COMPENDIUM OF
LDC-IL SENTENCE ALIGNED SPEECH CORPUS

EDITORS:
REJITHA K. S.
NARAYAN KUMAR CHOUDHARY

Linguistic Data Consortium for Indian Languages
Central Institute of Indian Languages
Mysuru

COMPENDIUM OF
LDC-IL 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.

Editors:
Rejitha K. S.
Narayan Kumar Choudhary

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: Compendium of LDC-IL Sentence Aligned Speech Corpus
Editors: Rejitha K. S., Narayan Kumar Choudhary

ISBN:978-81-19411-34-4
CIIL Publication No.: 1438

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: N. Rajesha

TABLE OF CONTENTS

FIGURES

TABLES

FOREWORD

Reflecting the unique charm and cultural heritage of human culture, Indian languages have an unparalleled diversity that shines as a glimpse of our social expression. Linguistic diversity in India is a sublime testament to the country's pluralism and reflects the complex interplay of our history and geography. In India farmers are the backbone of the country, once they could ask their questions in their native tongue, and a system would understand their query and suggest relevant information in their own language then the dream of Digital India would manifest. It cannot happen without bringing NLP research and application in regional languages of India.

In an era of rapid AI advancement, it is imperative that Indian languages are not left behind to nurture India's momentum. Therefore, Indian languages need to be used in technology as well. Only then, all the people of India, regardless of language, be able to take advantage of the potential of advanced technologies. Virtual assistant devices, Navigation maps and other smart devices make our daily lives possible today through the combination of artificial intelligence and machine learning. An efficient intelligent product providing relevant assistance to users' and such systems are possible through high-quality language data. To develop such AI and ML models needs humongous amount of quality annotated data. Annotated Speech data is generally used to produce training data for speech recognition and natural language processing systems. It directly affects algorithmic performance of AI and ML algorithms, and it is an essential resource to develop language-oriented systems. Good quality and consistent annotation support to develop and deploy ML models and it helps to give accurate output.

Speech annotation is the process of transcribing and annotating spoken language into words, sounds, and other vocal elements into written text. It is crucial for a wide range of applications, including Transcription Services, Chatbots, Voice recognition systems, Language Learning, Voice Assistants and Medical Transcription etc. LDC-IL prepares in-house data sets; hence it offers high quality labelling and data security.

LDC-IL has published 19 Raw text and 23 Raw Speech datasets. "Compendium of LDC-IL Sentence Aligned Speech Corpus" is a collection of 14 annotated speech data set consisting of Indian languages viz. Assamese, Bengali, Hindi, Kannada, Konkani, Maithili, Malayalam, Marathi, Nepali, Odia, Tamil, Urdu, Indian English - Bengali Variant, Indian English - Kannada Variant. Besides these data sets the compendium included Chhattisgarhi Raw Speech Corpus and Chhattisgarhi Raw Text Corpus. I hope that these data sets will be an asset in various fields related to language. This collection of speech corpus data enables linguistic and technical research, providing a strong base for the development of tools and applications that can overcome language barriers, promote linguistic preservation, and facilitate better communication. This venture is commendable for the advancement of Indian languages.

Prof. Shailendra Mohan
Director
Central Institute of Indian Languages
Mysuru

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 annotation. The orthographically normalised annotation is the prompt text in all of these cases.

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: *সোনকালেকাপোৰধোৱা, বৰষুণআহিবএতিয়া।*
xunkaːle kaːpur ɖʱuwaː, bɔrɔxun aːhibɔ etijaː(Actual given sentence)
তুমিসোনকালেকাপোৰধোৱা, বৰষুণআহিবএতিয়া।
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*
চিনী *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 pakg^{h}ɔɹ > পাকঘৰ pak g^{h}ɔɹ*

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.

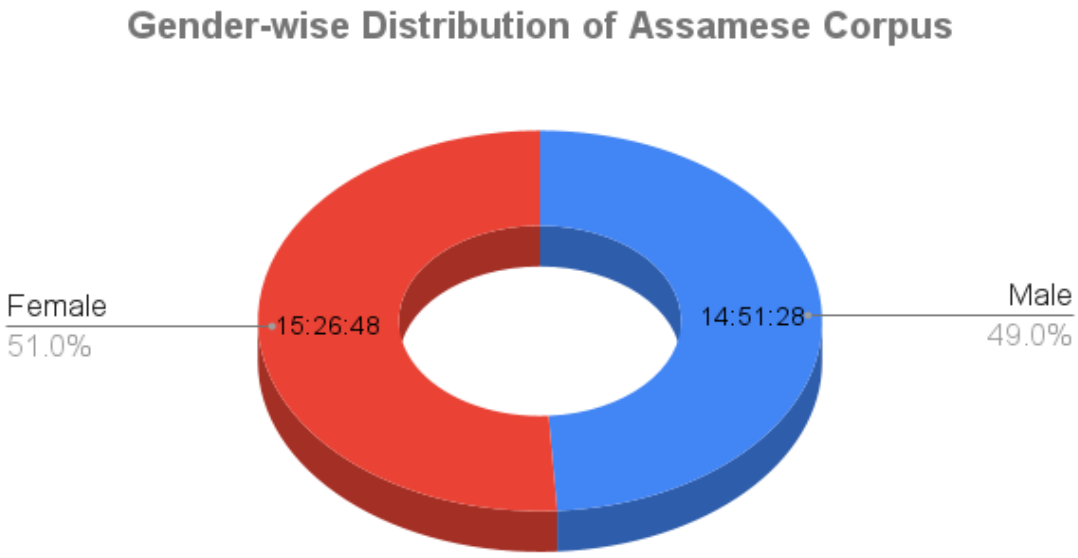

Figure 1: Gender-wise Distribution of Assamese Corpus

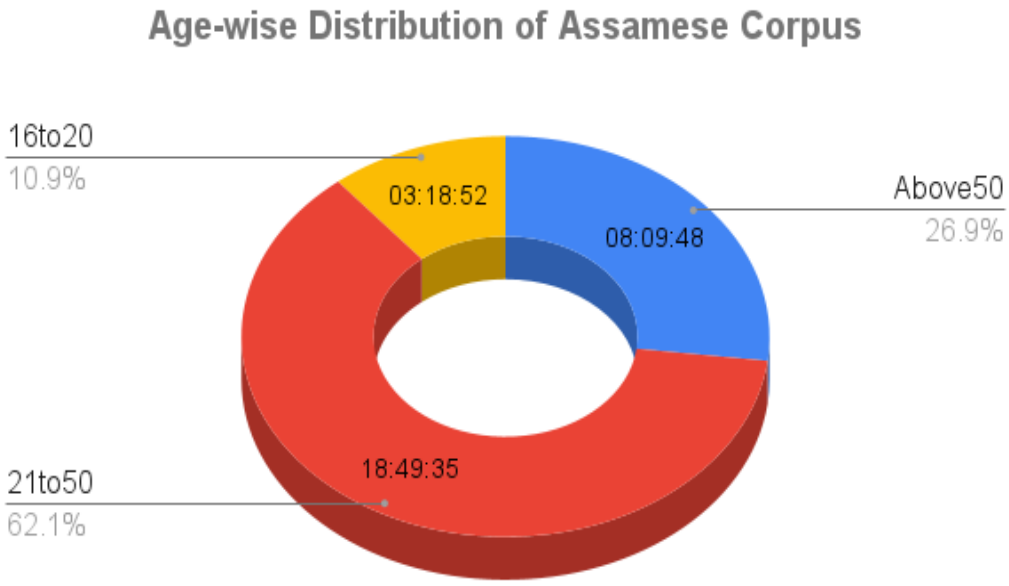

Figure 2: Age-wise Distribution of Assamese Corpus

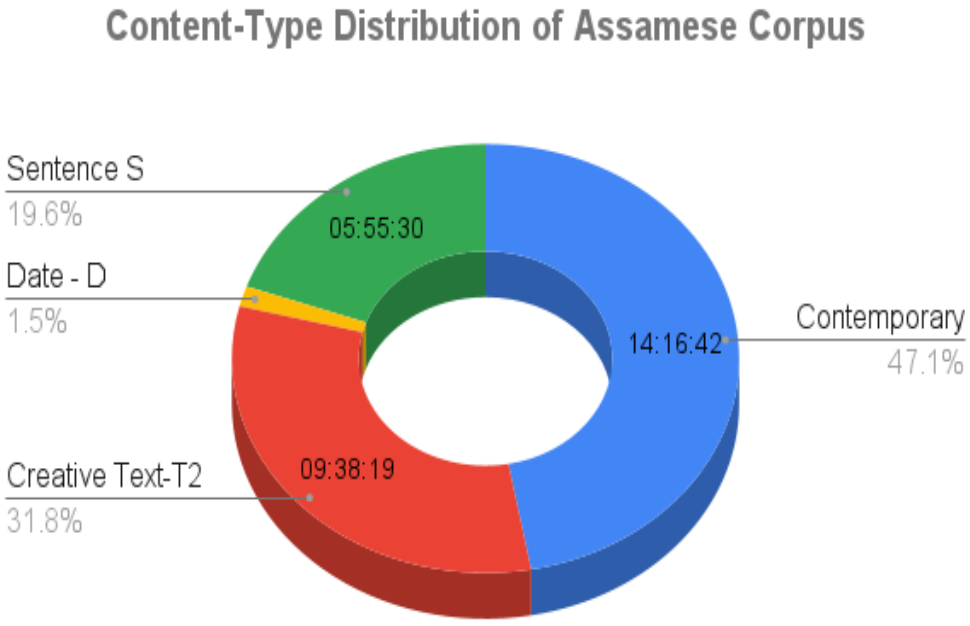

Figure 3: Content Type-wise Distribution of Assamese Corpus

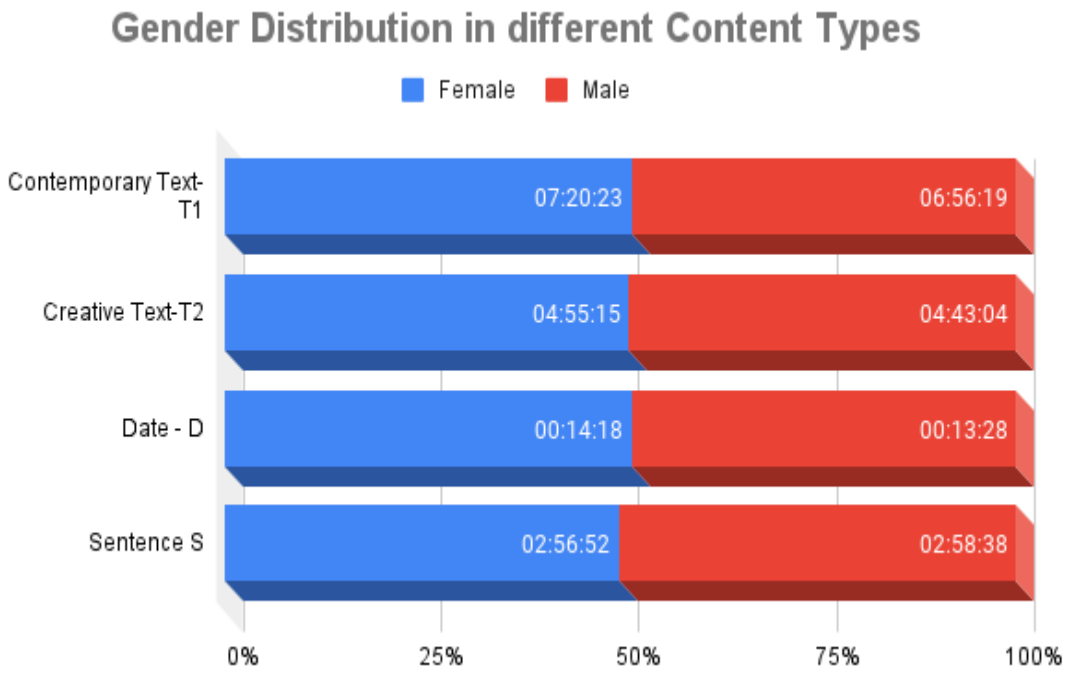

Figure 4: Gender Distribution in different Content Types of Assamese Corpus

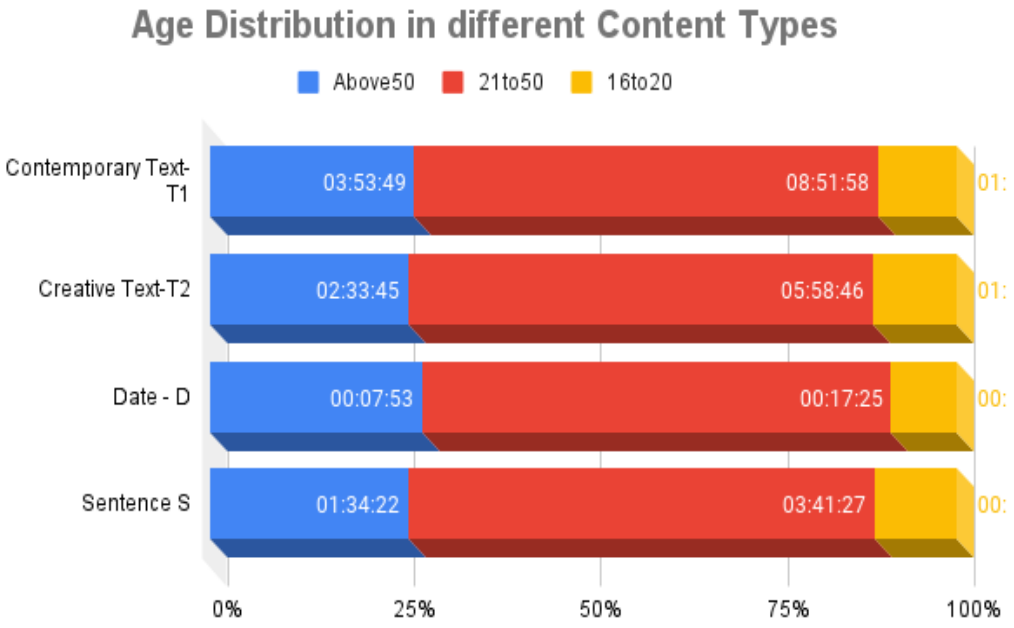

Figure 5: Gender Age Distribution in different Content Types of Assamese Corpus

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.

3 BENGALI SPEECH ANNOTATION

Sonali Sutradhar, Poulami Das, Narayan Kumar Choudhary

3.1 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.

3.2 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, /ɔrʤun ɾɔnotuŋgo/ "অর্জুন রনতুঙ্গো" for Arjuna Ranatunga etc.

3.2.1 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 atʃʰe/ "ভারোত কুনুক্‌ক্রোমে 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 pʰiɾia dekʰibaɾ puɾbe ekɔtʰa ʃikaɾ koɾia lɔoa utʃit/ "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 taɾataɾi poũtʃʰe gælo/ "সে খুব তাড়াতাড়ি পৌঁছে গেল" has been pronounced as /ʃe kʰub taɾataɾi poũtʃʰeælo/ "সে খুব তাড়াতাড়ি পৌঁছেঅ্যালো".
/ʃe ɾodʒ ɾate baɾi pʰeɾe/ "সে রোজ রাত করে বাড়ি ফেরে" has been pronounced as /ʃe ɾodʒ ɾate baɾi 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 udʒaɾ ɦoe gælo/ "গ্রামের পর গ্রাম উজাড় হয়ে গেল" has been pronounced as /geɾameɾ pɔɾ geɾam udʒaɾ 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 kɾiket̪ maʈʰ dækʰeni/ "অ্যাতো কেলেঙ্কারি এবং কুতশিত নোংরা পোক্খোপাতদুশ্টো আম্পায়ারিং ক্রিকেট মাড দ্যাখেনি" has been pronounced as /æto keleŋkaɾi eboŋ kutʃit noŋɾa pokkʰopat duʃʈo ampajaɾiŋ kɾiket̪ 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.

3.3 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.

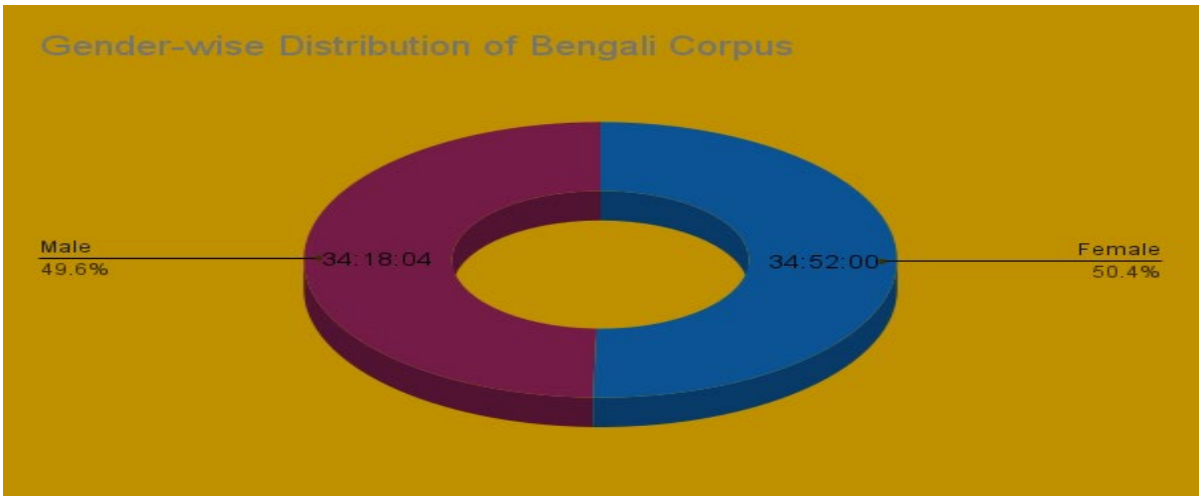

Figure 6: Gender-wise Distribution of Bengali Corpus

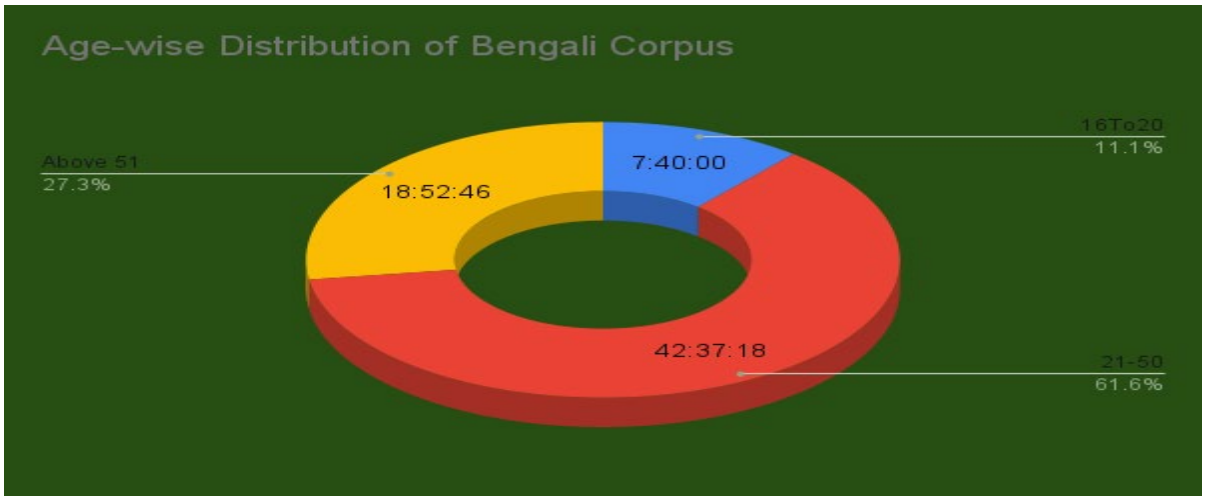

Figure 7: Age-wise Distribution of Bengali Corpus

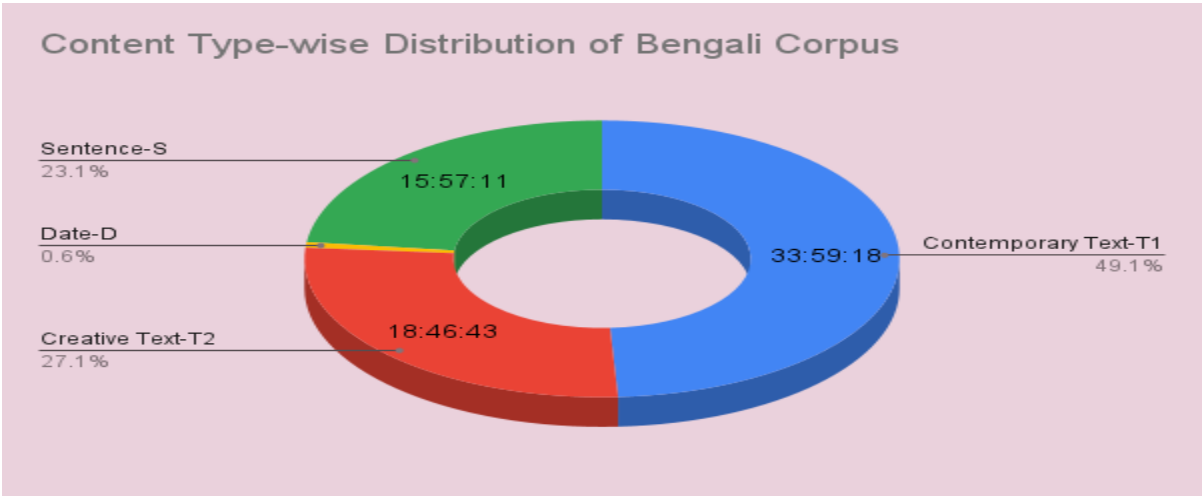

Figure 8: Content Type-wise Distribution of Bengali Corpus

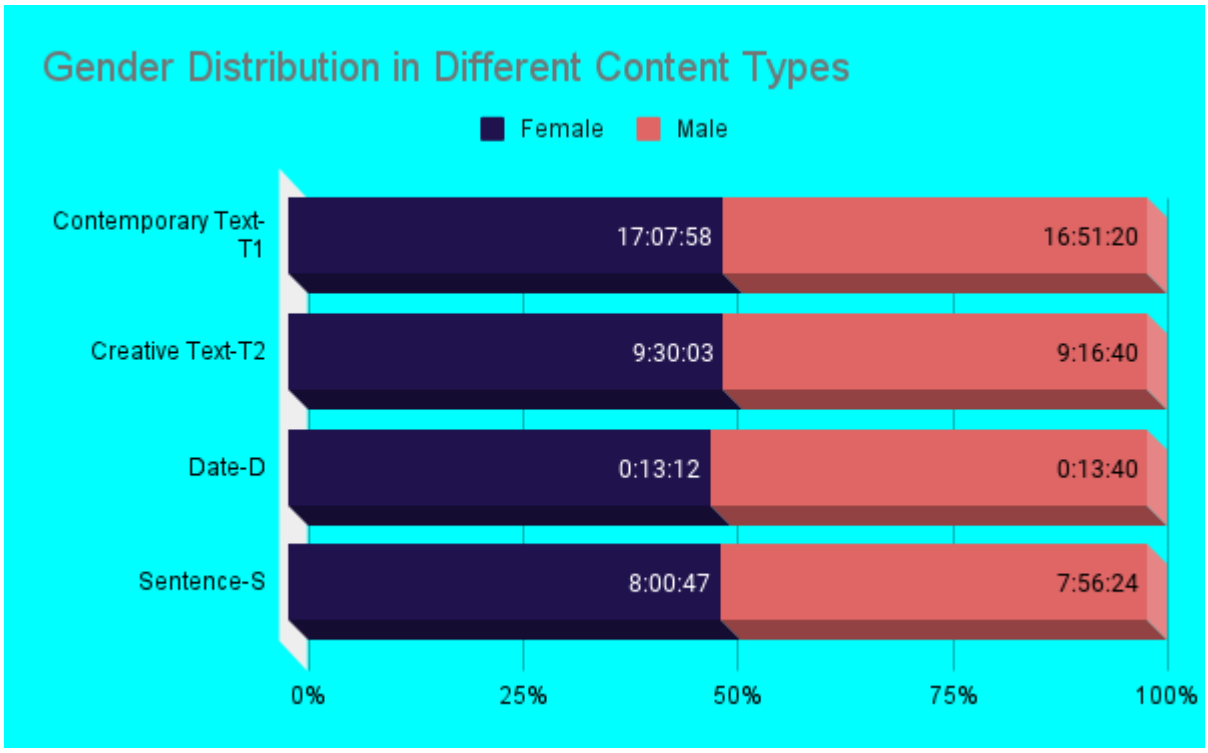

Figure 9: Gender Distribution in different Content Types of Bengali Corpus

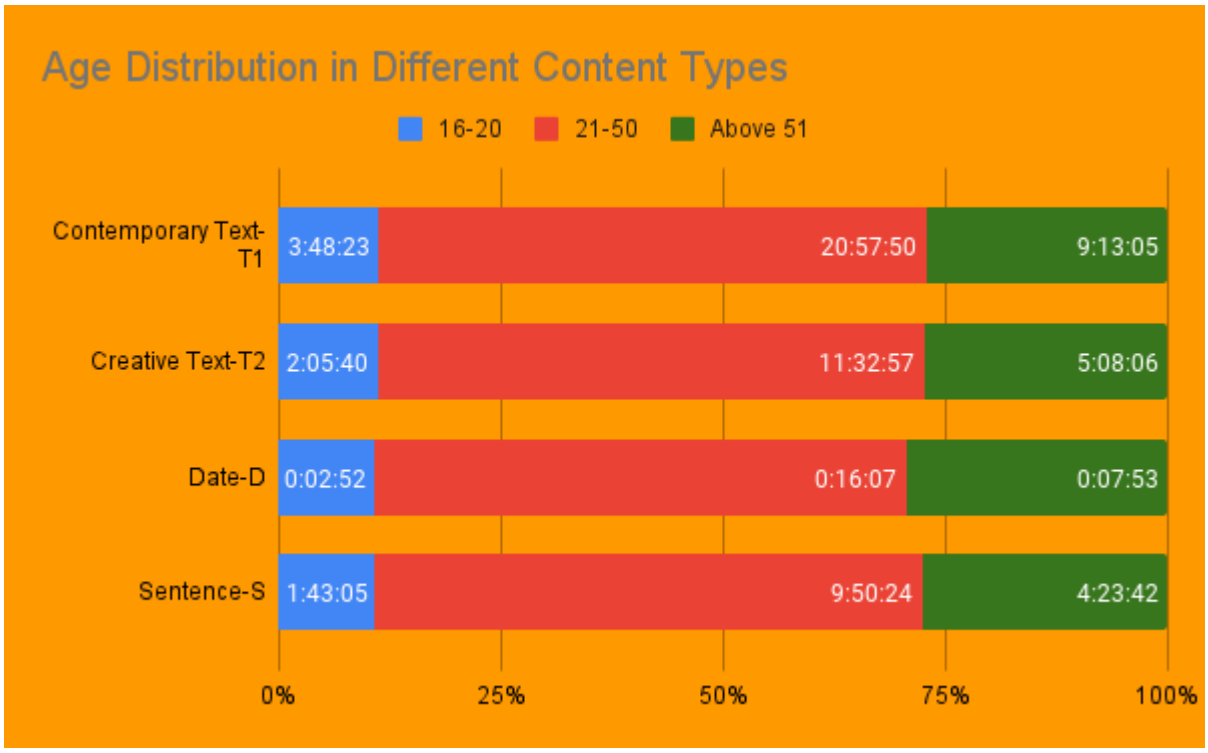

Figure 10: Age Distribution in different Content Types of Bengali Corpus

3.3.1 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 3: Representation of Bengali Sentence Aligned Speech Data Duration

3.4 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 4: Distribution of Speakers of Bengali Sentence Aligned Speech Data

3.5 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.

4 HINDI SPEECH ANNOTATION

Satyaendra Kumar Awasthi, Ankita Tiwari, Narayan Kumar Choudhary

4.1 OVERVIEW OF SENTENCE ALIGNED SPEECH CORPUS

Hindi Sentence Aligned Speech Corpus is created by annotating the speech data collected by LDC-IL. A detailed explanation of the Hindi Raw Speech Corpus will be available in the Hindi Speech Data Documentation (Ramamoorthy, L. et. Al, 2019). LDC-IL Hindi Sentence Aligned Speech files contain an audio file and two textual layers in Devanagari 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
'Hindi_Female_16To20_Contemporary_Text-T1_SP-0151_T1-0151-001.wav'

LDC-IL Sentence Aligned Speech corpus for Hindi 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's question. The corpus consists of an audio file for each recording and corresponding textual layer consisting of the phonetically normalized annotation. The orthographically normalized annotation is the prompt text in all of these cases.

4.2 OBSERVATIONS

LDC-IL sentence-level speech annotation strictly follows what the speaker pronounces to produce the phonetically normalized 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, some speakers pronounce /ɟəjɲ, kɑɾɟəkrəm, bikɑs, instead of /jəjɲ, kɑɾjəkrəm, ʋikɑs.
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 standardized 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.

4.2.1 PHONETIC ALTERNATION IN HINDI 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 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.

b. False start

False start is a common phenomenon in most of the speakers and 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.: b-ʋiɖ̪jɑ; ɾoɟ-sʋəɾoɟəgɑɾ; niɾɖ̪ʰɑ-niʃciʈ̪

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, if the audio is transcribed as [kʰəbəɾ]* ne pəɾeʃɑn kəɾ ɖ̪ijɑ ʈ̪ʰɑ, it shows that [kʰəbəɾ]* is not properly audible. In some words, some of the syllables or phone might not be audible to the listener or are skipped by the speaker.
For example, in [kʰəbəɾ]* some parts are not audible.

d. 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.: ɟimmeʋɑɾiː>ɟimmeɖ̪əʋɑɾiː; ɾəkʰɑ>ɾəkkʰɑ

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.: unniːs sɔ >unniːsɔ; səŋgəʈʰənõ>səŋgəʈʰõ; məʈ̪əɖ̪ɑʈ̪ɑõ>məɖ̪ɖ̪ɑõ

e. 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.: cəkr >cəkk; cʰəɖ̪mõ >cʰəɖ̪ɖ̪õ

Dissimilation is dropping out a syllable or a letter by the influence of adjacent speech segments.
E.g.: ɖ̪uːɾ>ɖ̪uːɽ ; kəɾijɑ>kəɳijɑ

f. Colloquial usage

Some of the speakers have pronounced colloquial forms instead of the standardized form written in the prompt sheet.
E.g.: ɒpəɾeʃən>ɑpəɾesən; ɖɒkʈəɾ>ɖɑkʈəɾ; ʋikɑs>bikɑs

g. Lengthening and Shortening

Short and long vowels are interchanged in the recordings at several places.
E.g.: niɖ̪ʱi>niɖ̪ʱi:; unni:s>unnis; pəcci:s>pəccis

h. Substandard alternation

It has been observed that some speakers have consistently replaced the aspirated sounds with their unaspirated counterparts.
E.g.: ɖ̪oʂ>ɖ̪os; ʃɑsən>sɑsən; ʃikʂiʈ̪ >ʃiccʰiʈ̪; prəɳɑli:>prənɑli:

i. 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.: kəfən>kəpʰən; məɲzil>mənɟil; məccʰəɾ>məccʰəʈ

j. Final vowel modification

In continuous speech the final vowel gets modified at times in some of the speakers:
E.g.: ɾɑʂʈ̺rəpəʈ̪i>ɾɑʂʈ̺rəpəʈ̪i:

k. Phonetic variation

While pronouncing a word that starts with ‘j’ the Hindi some native speaker changes in ‘ɟ’.

l. 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.

E.g.: kəmlənəjən> kəməl nəjən

4.3 SUMMARY OF THE CORPUS

The total duration of Hindi Sentence Aligned Speech Corpus is 72:34:52 (hh:mm:ss) comprising 42,275 audio segments from 473 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 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.

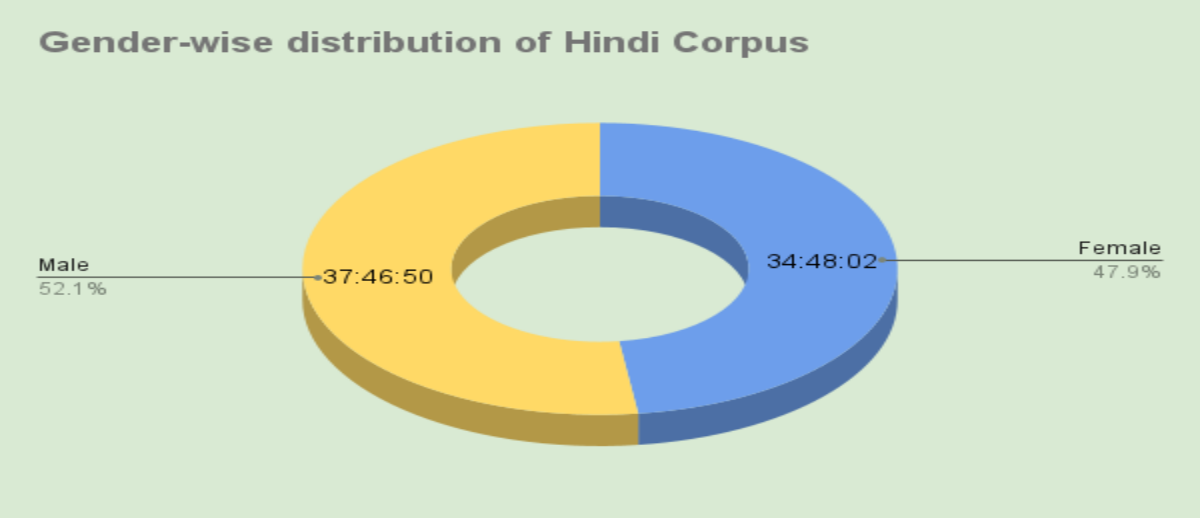

Figure 11: Gender-wise Distribution of Hindi Corpus

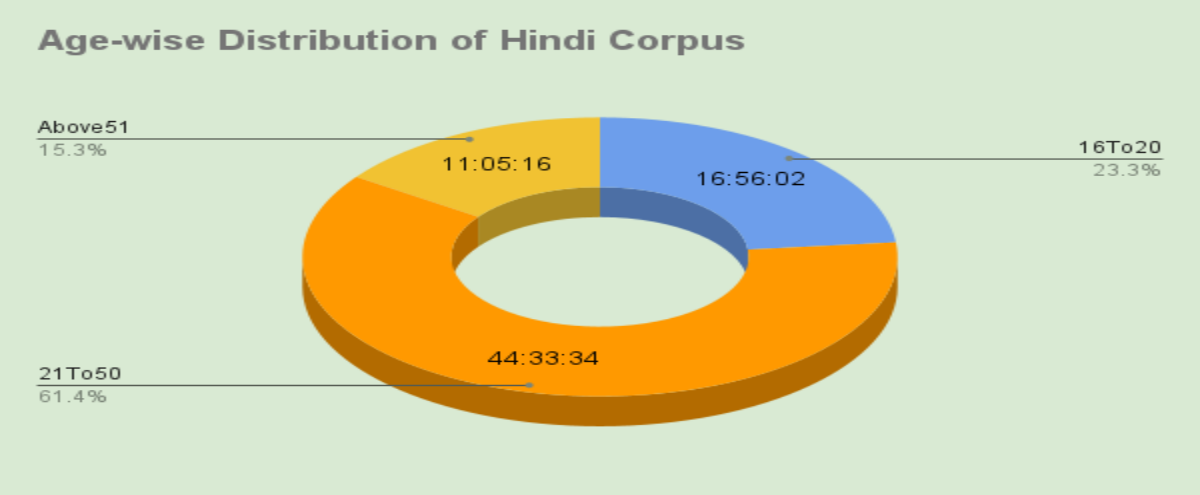

Figure 12: Age-wise Distribution of Hindi Corpus

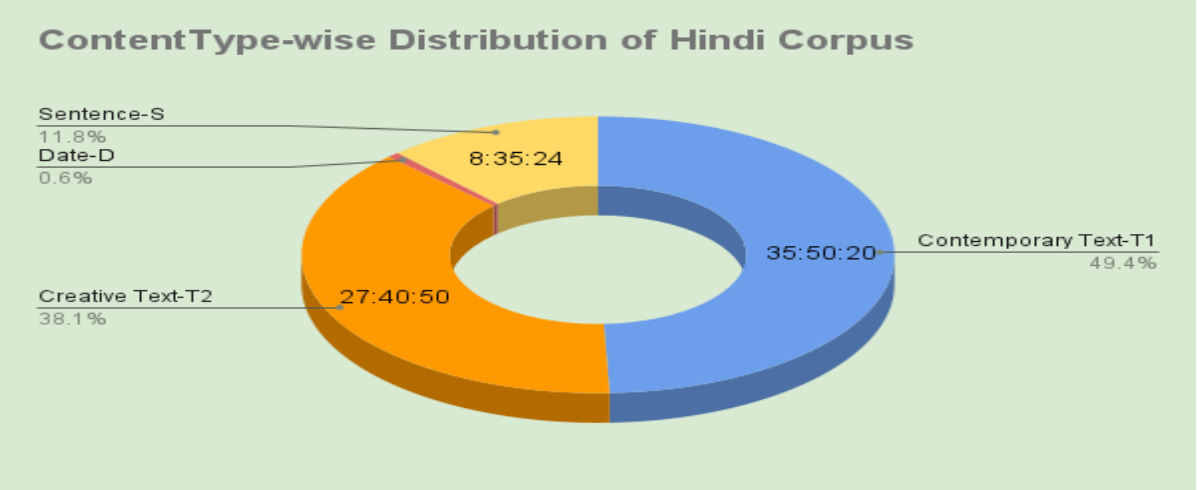

Figure 13: Content Type-wise Distribution of Hindi Corpus

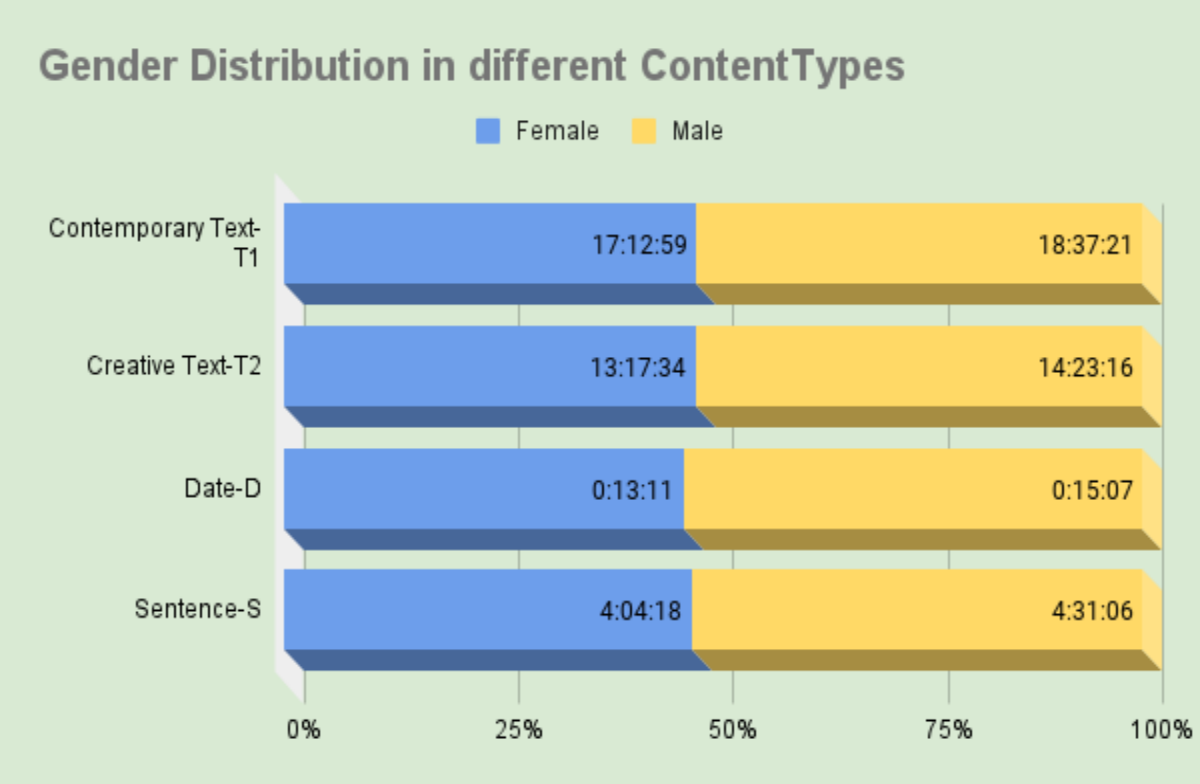

Figure 14: Gender Distribution in different Content Types of Hindi Corpus

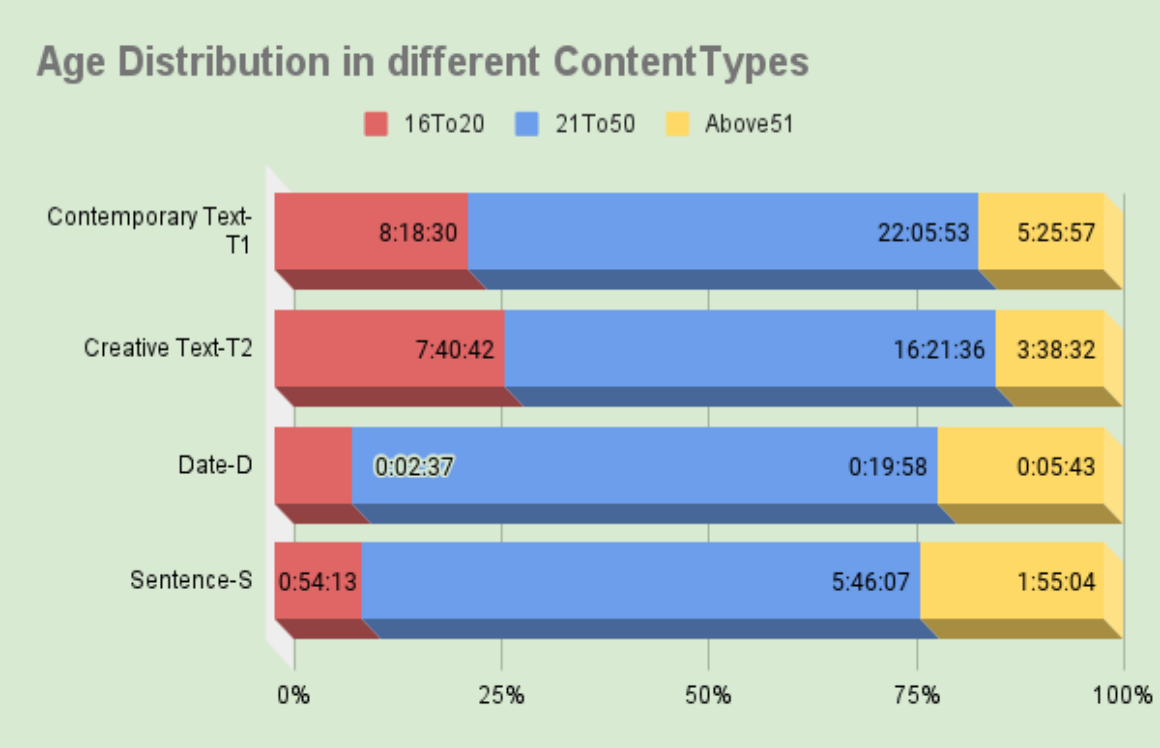

Figure 15: Age Distribution in different Content Types of Hindi Corpus

4.3.1 DURATION OF HINDI SENTENCE ALIGNED SPEECH DATA

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

Content Type	Gender	Age Group	Duration (hh:mm:ss.ms)		
Contemporary Text-T1	Female	16To20	03:48:55.819096	17:12:57.961371	35:50:18.745852
		21To50	11:46:53.599312		
		Above51	01:37:08.542962		
	Male	16To20	04:29:33.986646	18:37:20.784481	
		21To50	10:18:58.939584		
		Above51	03:48:47.858252		
Creative Text-T2	Female	16To20	03:46:50.761969	13:17:33.205978	27:40:49.401356
		21To50	08:33:05.558721		
		Above51	00:57:36.885288		
	Male	16To20	03:53:50.505135	14:23:16.195378	
		21To50	07:48:30.193023		
		Above51	02:40:55.497220		
Date-D	Female	16To20	00:00:50.717031	00:13:11.301597	00:28:18.264492
		21To50	00:10:42.297906		
		Above51	00:01:38.286660		
	Male	16To20	00:01:45.722761	00:15:06.962895	
		21To50	00:09:15.899074		
		Above51	00:04:05.341060		
Sentence-S	Female	16To20	00:19:01.900072	04:04:17.442270	08:35:22.877981
		21To50	03:12:38.942260		
		Above51	00:32:36.599939		
	Male	16To20	00:35:10.961558	04:31:05.435711	
		21To50	02:33:27.610685		
		Above51	01:22:26.863467		

Table 5: Representation of Hindi Sentence Aligned Speech Data Duration

4.4 SUMMARY OF SPEAKERS

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

Age Group	Female	Male	Total
16To20	31	43	74
21To50	168	141	309
Above51	26	64	90
Total	225	248	473

Table 6: Distribution of Speakers of Hindi Sentence Aligned Speech Data

4.5 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, Jitendra Kumar Singh, Richa, Anjali Sinha, Dheeraj Kumar Mishra, Arimardan Kumar Tripathi & Satyaendra Kumar Awasthi. 2019. Hindi Raw Speech Corpus. Central Institute of Indian Languages, Mysore.
5. Ramamoorthy, L., Narayan Choudhary, Jitendra Kumar Singh, Richa, Anjali Sinha, Dheeraj Kumar Mishra, Arimardan Kumar Tripathi, Aditi Debsharma, Satyaendra Kumar Awasthi & Madhupriya Pathak. 2019. A Gold Standard Hindi Raw Text Corpus. Central Institute of Indian Languages, Mysore.

5 KANNADA SPEECH ANNOTATION

Rajesha N., Vijayalaxmi F. Patil, Chetan Baji, Narayan Kumar Choudhary

5.1 OVERVIEW OF SENTENCE ALIGNED SPEECH CORPUS

Kannada Sentence Aligned Speech Corpus is created by annotating the speech data collected by LDC-IL (Ramamurthy, L. et. Al, 2019). A detailed explanation of the Kannada Raw Speech Corpus will be available in the Kannada Speech Data Documentation. (N. Rajesha. et. Al, 2019). LDC-IL Kannada Sentence Aligned Speech files contain an audio file and two textual layers in Kannada 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
'Kannada _Female_16To20_Contemporary Text-T1_SP-0031_T1-0084-001.wav'

The speech is annotated on the basis of specific language syllable structure. The words are labeled manually to the corresponding wave. LDC-IL Sentence Aligned Speech corpus contains four content types such as contemporary text, creative text, sentences and date format. The contemporary text and creative text are recordings of news and essays/novels. Each speaker has uttered typically 25 sentences randomly selected from phonetically balanced sentences list of LDC-IL speech data set. Date format content type contains date format uttered by the speaker.

5.2 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 speech data, the dialect variation drastically influences the pronunciation. Therefore, speakers from different regions speak the same word in different ways. In North Karnataka (Hyderabad Karnataka and Mumbai Karnataka regions) speakers tend to pronounce the years in the connotation of hundreds whereas the other regions prefer in thousands. For example '*hattombatnuːra embatrombattara*' The Canara and Old Mysore Regions it will be pronounced as '*saːvirada ombʱainuːra embʱattombattu*. Another noticeable thing is for every 9s of the decimal system the "*repha*" is inserted after *ta-kara* in the region of Hyderabad Karnataka and Mumbai Karnataka.

49 = nalaʋattombattu (Old Mysore and Canara) = nalaʋatrombattu (North Karnataka)
59 = aiʋattombattu (Old Mysore and Canara) = aiʋatrombattu (Old Mysore and Canara) etc.

The reading speed differs from reader to reader. Fast reading informants pose difficulty in annotation. Since news items contain sports news, it includes the informant reading all types of numbers. Speakers sometimes wrongly uttered large digit numbers 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 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.

5.2.1 PHONETIC ALTERNATION IN KANNADA 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 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 start is a common phenomenon in most of the speakers and 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.: *avar-avarade:* *pre:-pari:kʂisuttade* *sa:mag-sa:magriga[alli* *he:ridda-he:ridare*

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. If the text is annotated as '*a:dre suʈiŋ[ginalli]** *ɕa:ʈ he:[uva:ga sigare:p psedabeḍa antiddru*' shows *[ginalli]** is not properly audible but native speakers could easily understand the word because of language proficiency. In the long words the middle of the syllable or phone might not be audible to the listener or skip by the speaker. i.e. In, '*sammi[ɕra]** *sarka:ra[da]** *mu:lad^{h}armavannu maretidde:ve*' the middle pair is not audible.

d. 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.: *ɕastraʧikitsejinda:gijo:* > *ɕastraɕriʧikitsejinda:gijo:*
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.: *innitara* > *innita* *viʂɳuvard^{h}an* > *viʂɳuvard^{h}a* *naḍejuvudilla* > *naḍejudilla*

e. 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.: *kaŋgaɭa > kaɳgaɭa* (phoneme 'ŋ' moved to 'ɳ' because of its following consonant.)
nanage > naŋge (phoneme 'n' moved to 'ŋ' because of its following consonant.)

Dissimilation is dropping out a syllable or a letter by the influence of adjacent speech segment.

E.g.: *vriṯṯijaḷḷi > riṯṯijaḷḷa* *saːdʰaːraɳavaːgi > saːdarɳavagi*
iverḍu > iveraḍu *ʧalanavalanagaɭu > ʧalnavalnanagaɭu*

f. Colloquial usage

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

E.g.: *aravatta**naːlku**>aravat**naːku**; **eː**ɭu >**joː**ɭu;* ***naḍesa**beːkendu >**neḍsa**beːkendu*

g. Lengthening and Shortening

Short and long vowels are interchanged in the recordings at several places.

E.g.: *pragatipara > pragaːtipara* *ʧunaːvaɳeju > ʧuːnaːvaɳeju*
pramukʰavaːgi > pramukʰavagi

h. Substandard alternation

It has been observed that some speakers have consistently replaced the aspirated sounds with their unaspirated counterparts.

E.g.: *utsaːha > ustaːha;* *vidʒajoːtsava > vidʒajoːstava;* *prajatna > preːtana*

i. Interchange of Voiced fricative with Vowels

It is observed that some informants interchange the Voiced fricatives (h) with vowels, this is more observed in Kannada of old mysore region speech.

Eg. Vowel in place of Voiced fricative
hakku> akku; haːlu > aːlu; hinnele > innele; heʧʧaːgi > eʧʧaːgi; hoːraːʈakke > oːraːʈakke;

Eg.Voiced fricative in place of Vowel
aːguttiruva > haːguttiruva; aːrtʰika > haːrtika; ilaːkʰe > hilaːkʰe;

j. Interchange of Voiced and voiceless

Kannada has voiced and voiceless consonants; some speakers have pronounced voiced consonants as voiceless or vice versa in some instances.

Eg. Voiceless in place of Voiced Consonant: *laiŋgika > laiŋkika*
Eg. Voiced in place of Voiceless Consonant: *divaːnakʰaːnegaɭigeː > divaːnagʰaːnegaɭige*

k. Interchange of Aspirated to unaspirated

Speakers tend to pronounce aspirated letters in unaspirated, and vice versa across all dialects. Aspirated to unaspirated is more commonly observed in the Old Mysore region speech.
Eg.Aspirated in place of unaspirated Consonant:

bʰajo:tpa:daka > bajotpa:daka;	*ʧʰa:je > ʧa:je;*	*bʰaja > baja;*
samartʰani:ja> samartani:ja;	*ullaŋgʰane > ullaŋgane*	*gʰo:ʂisidare >go:ʂisidare*

l. Interchange of Voiceless fricatives

Kannada has three voiceless fricatives namely, ಶ [ɕ] which is voiceless alveolo-palatal fricative, voiceless retroflex fricative ಷ [ʂ] and voiceless dental fricative ಸ [s]. It is observed that some informants interchange the Voiceless fricatives.

prave:ɕisida:galu: > prave:sisida:galu:	*ɕikʂejinda > sikʂejinda*	*prasa:da > praɕa:da*
naɖesutta:ra: > naɖeɕuttara	*praɕnege > prasnege*	*ɕa:nti>sa:nti:*
ma:nasikava:gi > ma:naɕikava:gi	*ɕla:gʰisi > sla:gisi:*	*niʂpakʂa > nispakʂa*
anulakʂisi> anulakʂiɕi		

5.3 SUMMARY OF THE CORPUS

Below section is providing the tabular details of the different content types of the Kannada Sentence Aligned Speech Corpus. These figures may be helpful in tuning the corpus for various purposes of training, testing and evaluating various algorithms as well as provide useful insights into the dataset. The total duration of Kannada Sentence Aligned Speech Corpus is 107:48:50 (hh:mm:ss) comprising 600 speakers.

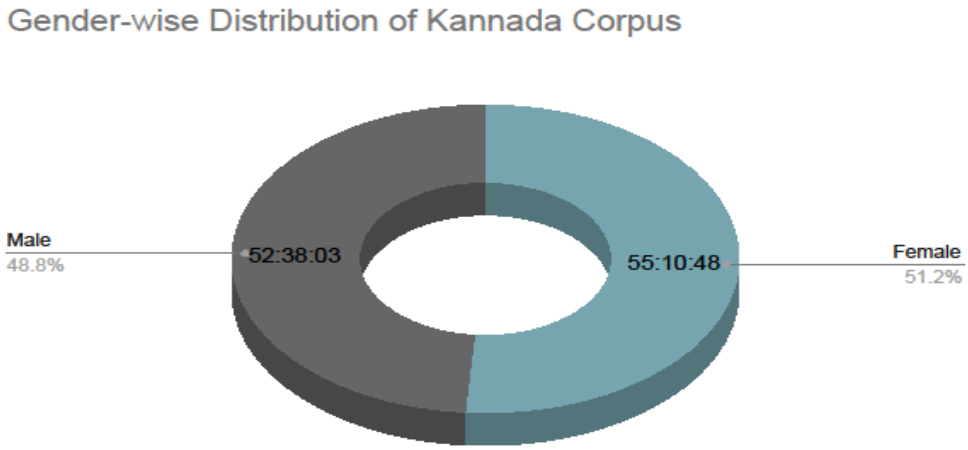

Figure 16: Gender-wise Distribution of Kannada Corpus

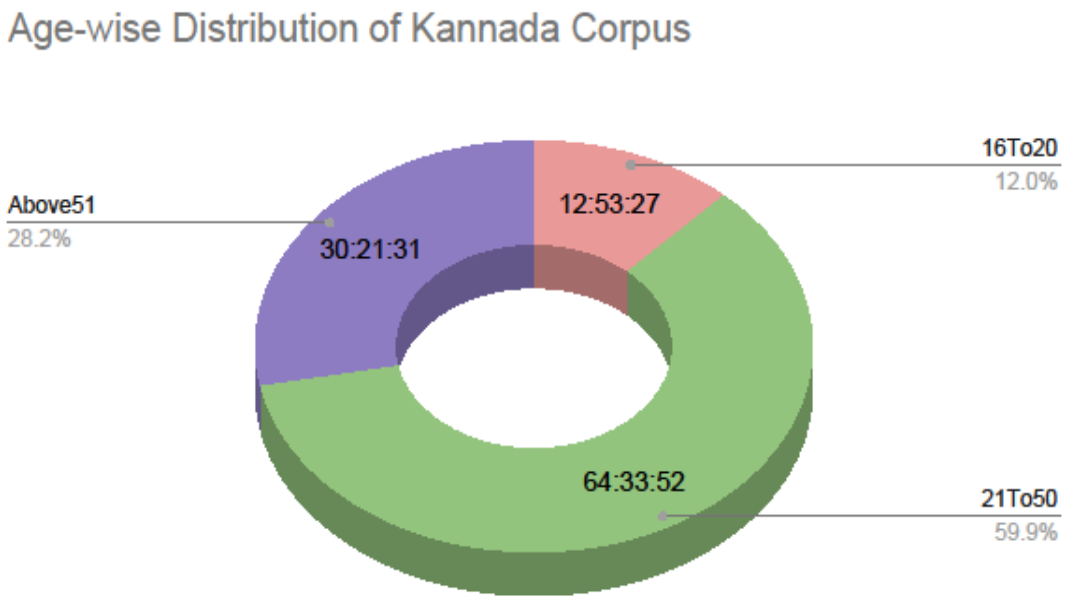

Figure 17: Age-wise Distribution of Kannada Corpus

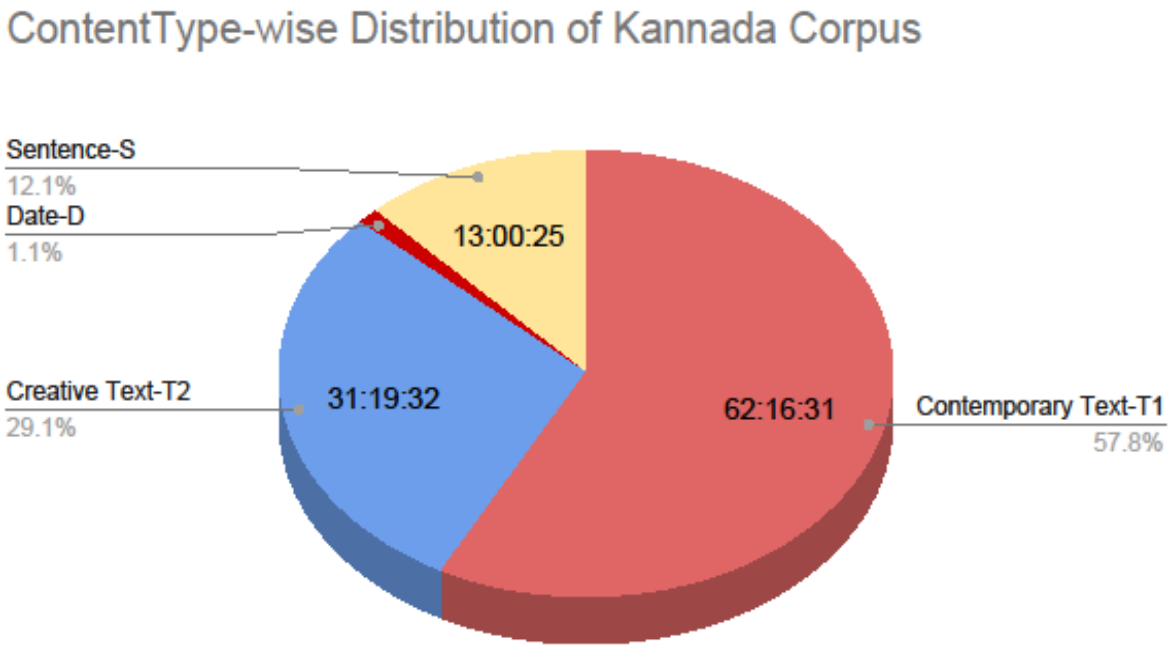

Figure 18: Content Type-wise Distribution of Kannada Corpus

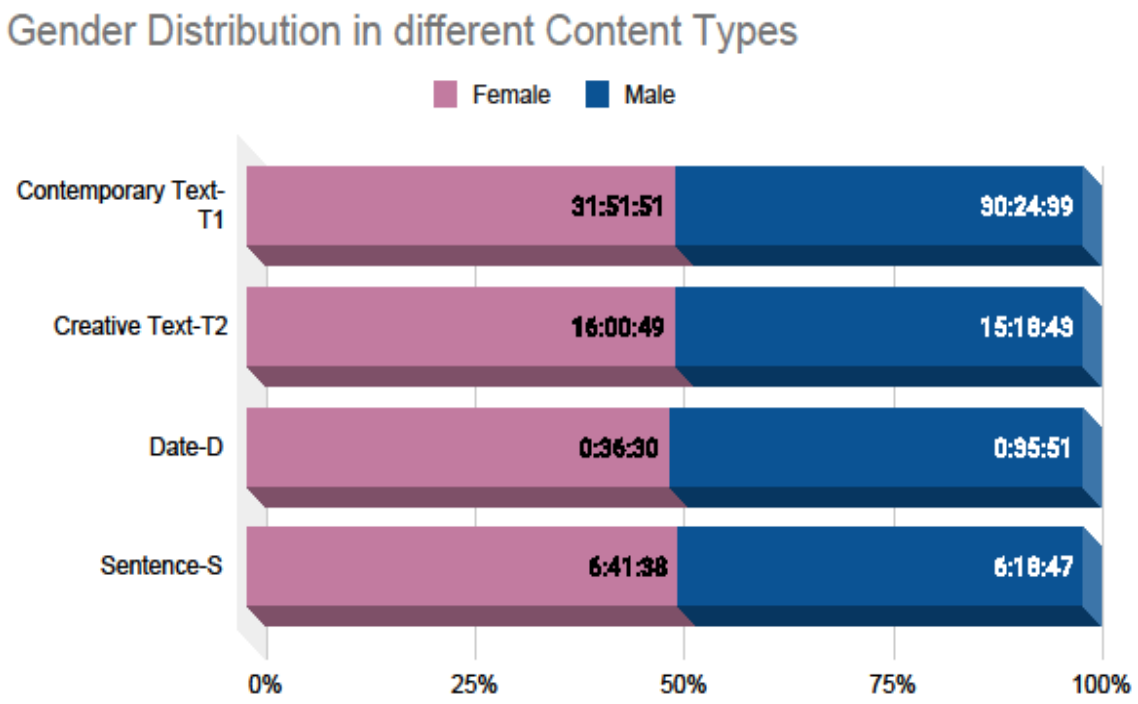

Figure 19: Gender Distribution in different Content Types of Kannada Corpus

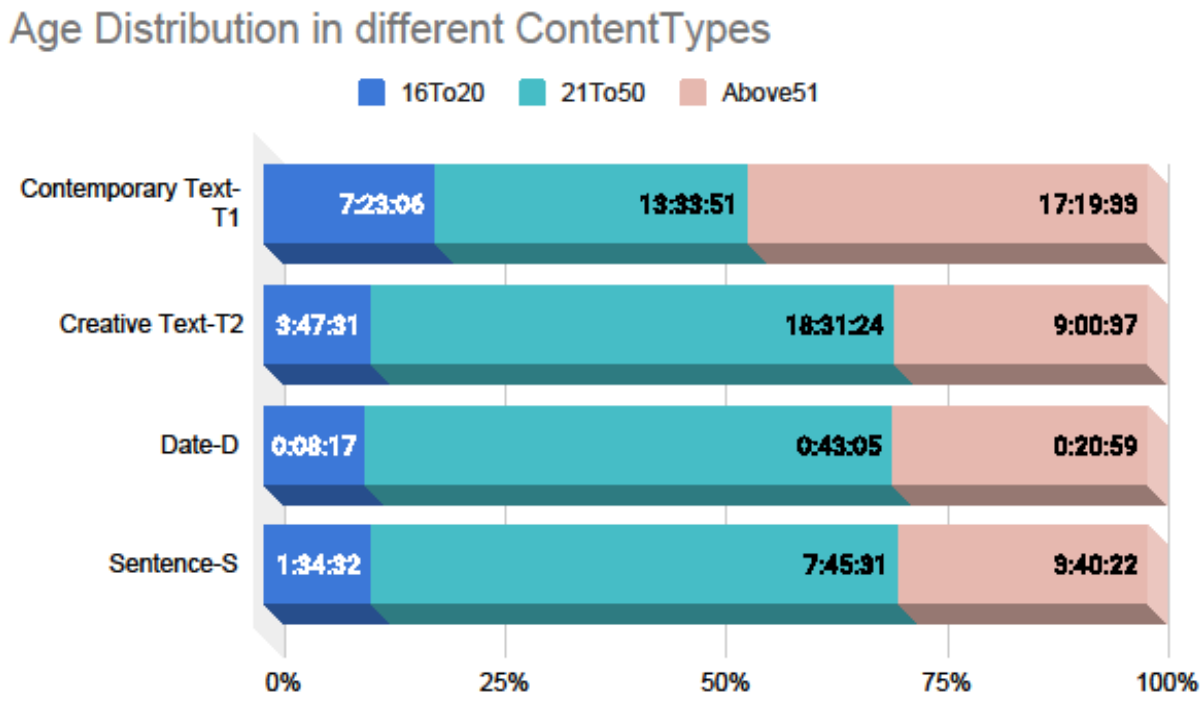

Figure 20: Age Distribution in different Content Types of Kannada Corpus

5.3.1 DURATION OF KANNADA SENTENCE ALIGNED SPEECH DATA

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

Content Type	Gender	Age Group	Duration (hh:mm:ss.ms)		
Contemporary Text-T1	Female	16To20	03:53:54.466314	31:51:51.440245	62:16:30.816962
		21To50	19:12:14.460695		
		Above51	08:45:42.513236		
	Male	16To20	03:29:11.927437	30:24:39.376717	
		21To50	18:21:37.332607		
		Above51	08:33:50.116673		
Creative Text-T2	Female	16To20	02:01:34.089509	16:00:48.283607	31:19:32.354477
		21To50	09:23:24.588579		
		Above51	04:35:49.605519		
	Male	16To20	01:45:57.261359	15:18:44.070870	
		21To50	09:07:59.454921		
		Above51	04:24:47.354590		
Date-D	Female	16To20	00:04:29.265599	00:36:30.198127	01:12:21.837967
		21To50	00:21:08.526276		
		Above51	00:10:52.406252		
	Male	16To20	00:03:48.484909	00:35:51.639839	
		21To50	00:21:56.158461		
		Above51	00:10:06.996469		
Sentence-S	Female	16To20	00:50:02.576569	06:41:37.684314	13:00:25.447202
		21To50	03:59:11.092277		
		Above51	01:52:24.015468		
	Male	16To20	00:44:28.954632	06:18:47.762889	
		21To50	03:46:20.421473		
		Above51	01:47:58.386784		

Table 7: Representation of Kannada Sentence Aligned Speech Data Duration

5.4 SUMMARY OF SPEAKERS

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

Age Group	Female	Male	Total
16To20	36	36	72
21To50	180	180	360
Above51	84	84	168
Total	300	300	600

Table 8: Distribution of Speakers of Kannada Sentence Aligned Speech Data

5.5 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, Vijayalaxmi F. Patil, Chetan Suryakant Baji, Malini N. Abhyankar, Rajesha N. & Manasa G. 2019. Kannada Raw Speech Corpus. Central Institute of Indian Languages, Mysore. ISBN: 978-81-7343-228-6
5. N., Rajesha, Vijayalaxmi F. Patil, Manasa G., Chetan Baji, Narayan Choudhary & L. Ramamoorthy. 2019. "Documentation of LDC-IL Kannada Raw Speech Corpus" in Linguistic Resources for AI/NLP in Indian Languages. Central Institute of Indian Languages, Mysore. pp. 205-214.

6 KONKANI SPEECH ANNOTATION

Saurabh Varik, Narayan Kumar Choudhary

6.1 OVERVIEW OF SENTENCE ALIGNED SPEECH CORPUS

Konkani Sentence Aligned Speech Corpus is created by annotating the speech data collected by LDC-IL. A detailed explanation of the Konkani Raw Speech Corpus will be available in the Konkani Speech Data Documentation (Ramamurthy, L. et. Al, 2019). LDC-IL Konkani Sentence Aligned Speech files contains an audio file and two textual layers in Konkani 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
'Konkani_Female_16To20_Contemporary Text-T1_SP-0031_T1-0031-001.wav'

LDC-IL Sentence Aligned Speech corpus for Konkani 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 layer consisting of the phonetically normalised annotation. The orthographically normalised annotation is the prompt text in all of these cases.

6.2 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, uɳ̊ɛ̃ kɐr "reduced" sometimes sounds like (uɳ̊ɛ̃ ŋkɐr), Kokum (Garcinia indica) - bʰɪrɪɳɖ̚ə/bʰɪrəɳɖ̚ə/bʰɪɳɳ̚ə.

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.

6.2.1 PHONETIC ALTERNATION IN KONKANI 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 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 start is a common phenomenon in most of the speakers and 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.: gʰə-gʰərə, gə-gʰərə

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, if the audio is transcribed as mə̃n̪ɾɑ:[ləjɑ:]*t͡ʃe:ɾə (मंत्राचेर*[लया]), it shows that [ləjɑ: (लया)] is not properly audible.In some longer words, the middle of the syllable or phone might not be audible to the listener or are skipped by the speaker. For example, in prəd̪ʰɑ:n̪ə*t̪ri:[प्रधान*त्री]the middle part is not audible.

d. 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.: s̪ə̃t͡ʃɑ:ɾə (संचार) > s̪ə̃t͡ʃət͡ʃɑ:ɾə (संचचार); s̪un̪ɑ:pəɾə̃:t̪ə (सुनापरांत) > s̪un̪ɑ:pəɾə̃:n̪ət̪ə (सुनापरांनत)

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.: kəɾu:n̪ə [करून] > kəɾən̪ə[करन]

e. 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.: d̪ɪlle:>me:l̪l̪e:दिल्ले>मेळ्ळे

Dissimilation is dropping out a syllable or a letter by the influence of adjacent speech segments.
E.g.: pʰərəmɑ:n̪ə > pərəmɑ:n̪ə फरमान>परमान

f. Colloquial usage

Some of the speakers have pronounced colloquial forms instead of the standardised form written in the prompt sheet.
E.g.: ɑːʈʰʈʰiːʂə > əɖəʈiːʂə आठ्ठीस>अडतीस, ɪkərɑː > əkərɑː इकरा > अकरा

g. Lengthening and Shortening

Short and long vowels are interchanged in the recordings at several places.
E.g.:,ɑːɳiː > ɑːɳɪ आनी > आनि, kəʋɪʈɑː > kəʋiːʈɑː कविता > कवीता

h. Substandard alternation

It has been observed that some speakers have consistently | replaced the aspirated sounds with their unaspirated counterparts. Moreover, in Kozhikode region people adapted easily articulated sounds.
E.g.: pɑːkʰoː > pɑːkoːपाखो > पाको, ʈɑːɾiːkʰə > ʈɑːɾiːkə तारीख > तारीक

i. Final vowel Elision

In the Kanara-Saraswat dialect, all solitary words end with a vowel, but in connected speech all word-final vowels in multisyllabic words are omitted when another word follows without a pause. hə̃va tākkā āppaytə̃ (hə̃va "me", tākka "he", āppaytə̃ "telephone") is pronounced hə̃v tāk āppaytə̃. Such vowel omissions in connected speech are also found in the Sashti dialect of Christianity. In both dialects, if the omitted vowel is a prevowel, the preceding consonant is palatalized.
Vowel Rounding in Christian Dialects: In the dialects of Christianity (Valdez Christian and Saxti Christian), the vowel a is fused with the o, resulting in fewer vowel phonemes.

j. 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.
E.g.: bɑːjələməɳiːʂə > bɑːjələ məɳiːʂə, gʰərəɖɑːɾə > gʰərə ɖɑːɾə

6.3 SUMMARY OF THE CORPUS

The total duration of Konkani Sentence Aligned Speech Corpus is 83:19:42 (hh:mm:ss) comprising 34,091 audio segments from 487 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.

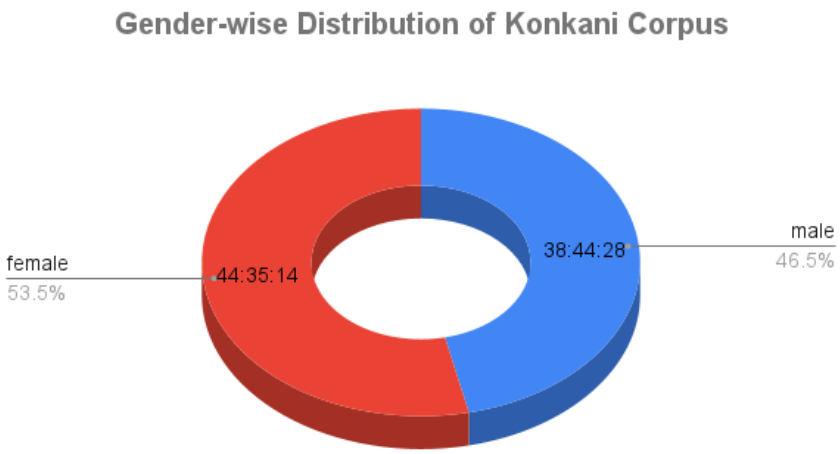

Figure 21: Gender-wise Distribution of Konkani Corpus

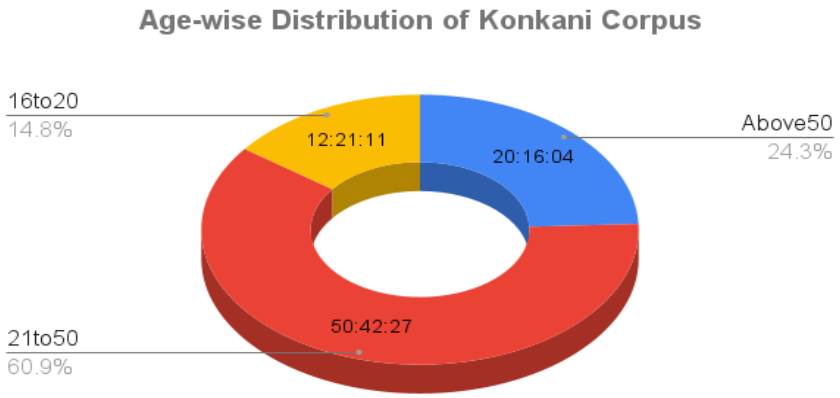

Figure 22: Age-wise Distribution of Konkani Corpus

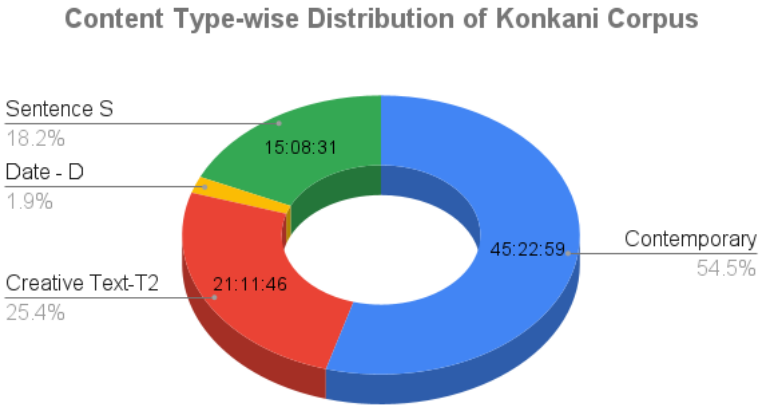

Figure 23: Content Type-wise Distribution of Konkani Corpus

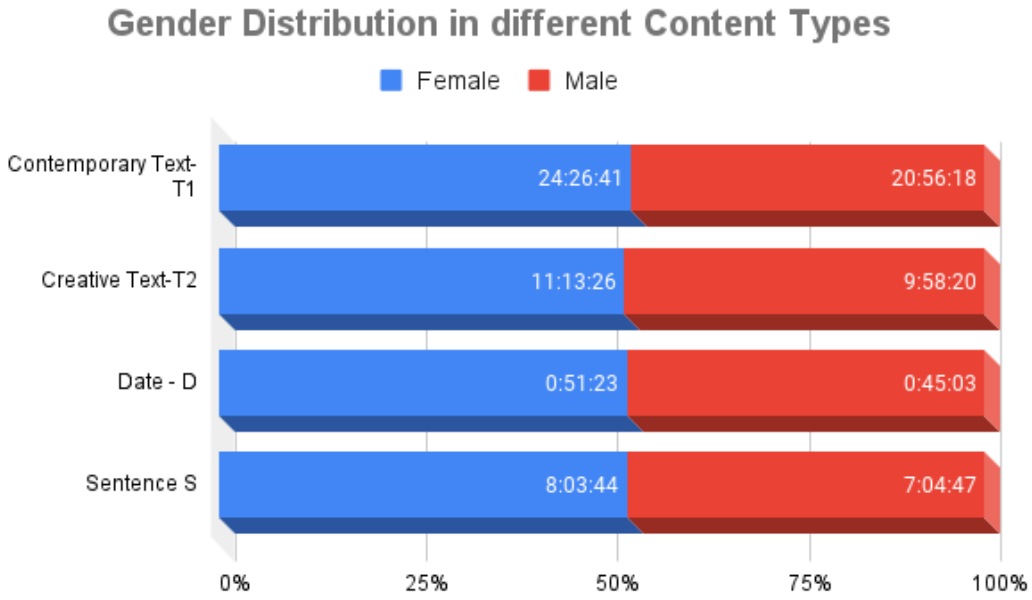

Figure 24: Gender Distribution in different Content Types of Konkani Corpus

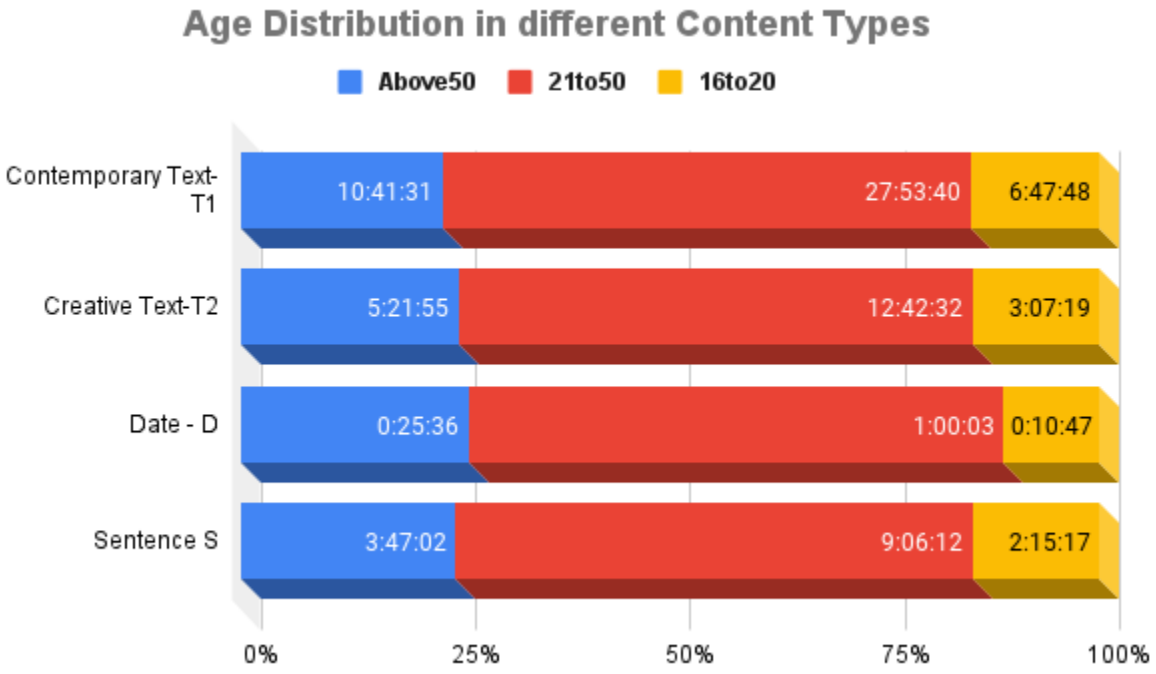

Figure 25: Age Distribution in different Content Types of Konkani Corpus

6.3.1 DURATION OF KONKANI SENTENCE ALIGNED SPEECH DATA

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

Content Type	Gender	Age Group	Duration (hh:mm:ss.ms)		
Contemporary Text-T1	Female	16To20	03:56:35.127344	24:26:41.071740	45:22:58.516021
		21To50	15:07:25.134619		
		Above51	05:22:40.809776		
	Male	16To20	02:51:12.656557	20:56:17.444281	
		21To50	12:46:15.072803		
		Above51	05:18:49.714921		
Creative Text-T2	Female	16To20	01:49:03.351690	11:13:26.284276	21:11:46.954730
		21To50	06:39:52.328808		
		Above51	02:44:30.603779		
	Male	16To20	01:18:15.916249	09:58:20.670454	
		21To50	06:02:40.489887		
		Above51	02:37:24.264317		
Date-D	Female	16To20	00:06:51.936300	00:51:22.595258	01:36:24.983446
		21To50	00:31:17.980378		
		Above51	00:13:12.678580		
	Male	16To20	00:03:54.638059	00:45:02.388189	
		21To50	00:28:45.135341		
		Above51	00:12:22.614789		
Sentence-S	Female	16To20	01:20:13.004817	08:03:44.234407	15:08:31.069589
		21To50	04:48:10.121925		
		Above51	01:55:21.107665		
	Male	16To20	00:55:03.584759	07:04:46.835182	
		21To50	04:18:01.760069		
		Above51	01:51:41.490353		

Table 9: Representation of Konkani Sentence Aligned Speech Data Duration

6.4 SUMMARY OF SPEAKERS

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

Age Group	Female	Male	Total
16To20	40	26	66
21To50	157	140	297
Above51	62	62	124
Total	259	228	487

Table 10: Distribution of Speakers of Konkani Sentence Aligned Speech Data

6.5 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, Saurabh Varik, Rashmi Shet Tanawade & Yashwant D Gawas. 2019. A Gold Standard Konkani Text Corpus. Central Institute of Indian Languages, Mysore.
5. Ramamoorthy, L., Narayan Choudhary, Saurabh Varik & Rashmi Shet Tanawade. 2019. Konkani Raw Speech Corpus. Central Institute of Indian Languages, Mysore.

7 MAITHILI SPEECH ANNOTATION

Shantanu Kumar, Dinesh Mishra, Narayan Kumar Choudhary

7.1 OVERVIEW OF SENTENCE ALIGNED SPEECH CORPUS

Maithili Sentence Aligned Speech Corpus is created by annotating the speech data collected by LDC-IL. A detailed explanation of the Maithili Speech Corpus is available in the Maithili Speech Data Documentation (Ramamurthy, L. et. Al, 2019). LDC-IL Maithili Sentence Aligned Speech files contain an audio file and two textual layers in Devanagari 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 annotation. The orthographically normalised annotation is the prompt text in all of these cases.

7.2 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.

7.2.1 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.: ब-विद्या निर्धा - निर्धारित

b-vɪɖ̪jaː nɪrɖ̪ʰaː - nɪrɖ̪ʰaːrɪt̪

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.: जिम्मेवारी > जिम्मवारी राखल > रक्खल

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.: उन्नीससौ>उन्नीसौ संगठनों > संगठों

ʊnniːs səʊ > ʊnniːsəʊ səŋgəʈʰnoː̃ > səŋgəʈʰoː̃

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.: चक्र>चक्क छद्मों>छद्दों

tʃəkr> tʃəkk tʃʰəɖ̪moː̃ > tʃʰəɖ̪oː̃

Dissimilation is dropping out a syllable or a letter by the influence of adjacent speech segments.

E.g.: करिया>कणिया

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.: ऑपरेशन>आपरेसन डॉक्टर>डाक्टर विकास>बिकास

ɒpəreːʃən > aːpəreːsən ɖɒkʈər > ɖɑkʈər vɪkaːs > bɪkaːs

The original form has been kept in the transcription.

f. Lengthening and Shortening

Short and long vowels are interchanged in the recordings at several places.

E.g.: उन्नीस>उन्निस पच्चीस>पच्चिस

unniːs > unnɪs pətʃʈʃiːs > pətʃʈʃɪs

g. Substandard alternation

It has been observed that some speakers have consistently replaced the aspirated sounds with their unaspirated counterparts.

E.g.: दोष>दोस शासन>सासन शिक्षित>शिच्छत;

d̪oʂ > d̪os ʃaːsən>saːsən ʃɪkʂɪt̪ >ʃɪtʃʈʃʰɪt̪

h. 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.: गाँधी>गांधी

gaː̃d̪ʱi :> gaːnd̪ʱiː

i. 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. ‘गेलहा’ /geːləhaː/is the word pronounced by the speaker whereas the correct form of the word should be ‘गेलाह’ /geːlaːh/. So, while transcribing, the correct form has been kept.

7.3 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.

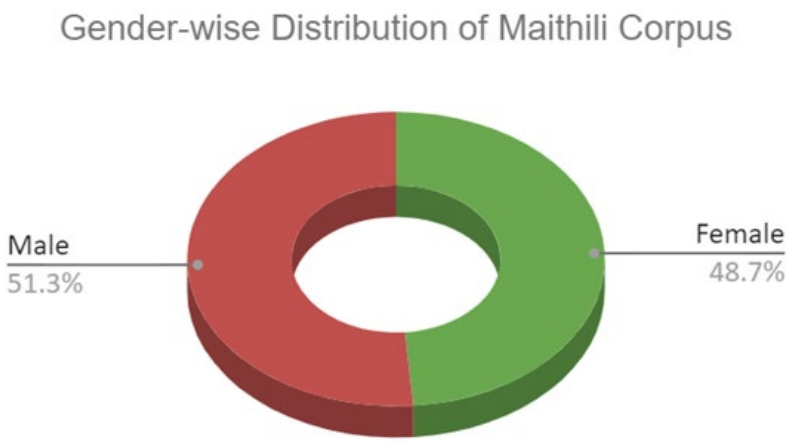

Figure 26: Gender-wise Distribution of Maithili Corpus

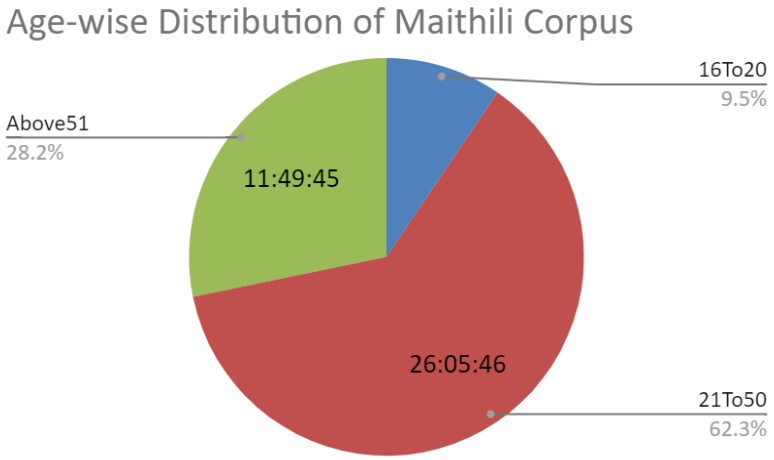

Figure 27: Age-wise Distribution of Maithili Corpus

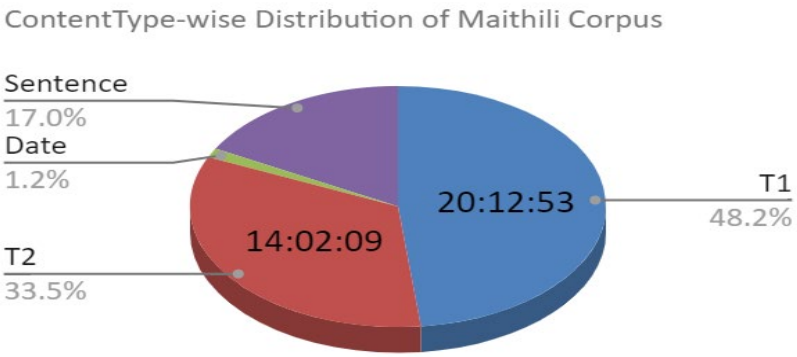

Figure 28: Content Type-wise Distribution of Maithili Corpus

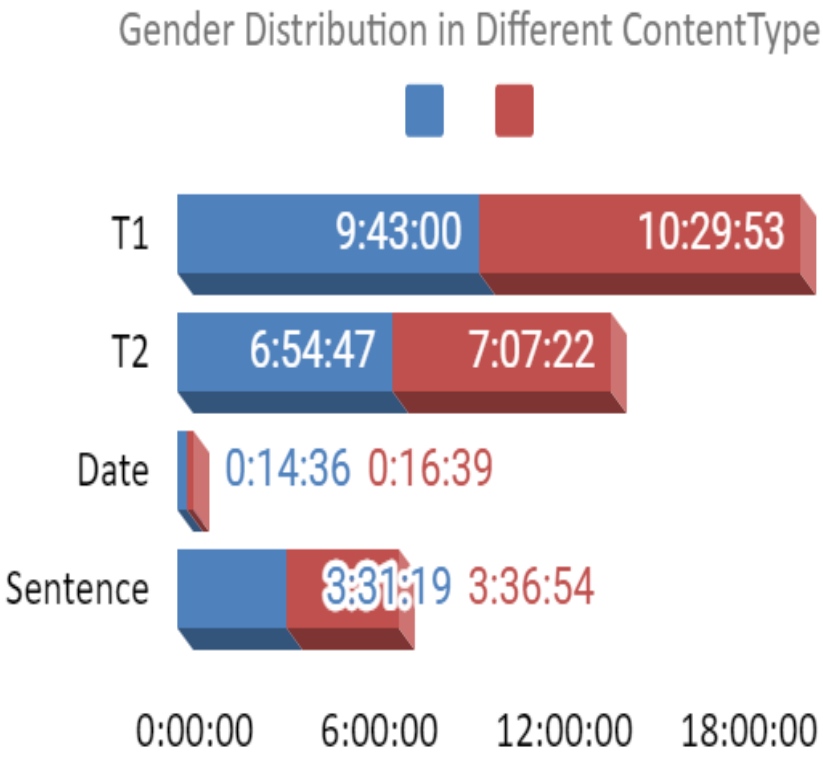

Figure 29: Gender Distribution in different Content Types of Maithili Corpus

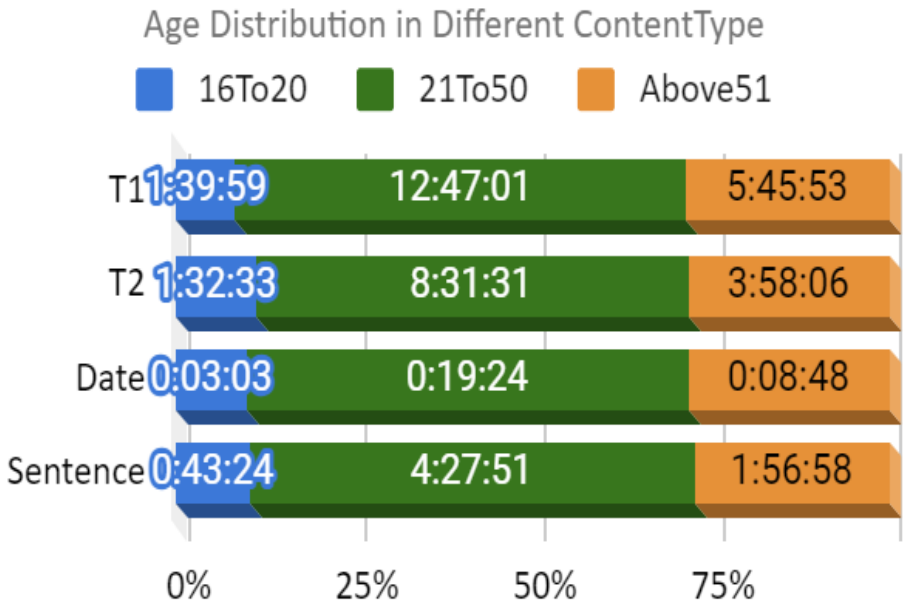

Figure 30: Age Distribution in different Content Types of Maithili Corpus

7.3.1 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 11: Representation of Maithili Sentence Aligned Speech Data Duration

7.4 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 12 : Distribution of Speakers of Maithili Sentence Aligned Speech Data

7.5 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 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.

8 MALAYALAM SPEECH ANNOTATION

Rejitha K. S., Sajila S., Saritha S.L., Narayan Kumar Choudhary

8.1 OVERVIEW OF SENTENCE ALIGNED SPEECH CORPUS

Malayalam 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 Malayalam Raw Speech Corpus will be available in the Malayalam Speech Data Documentation (Rejitha K.S., et. Al, 2019). LDC-IL Malayalam Sentence Aligned Speech files contain an audio file and two textual layers in Malayalam 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
'Malayalam_Female_16To20_Contemporary Text-T1_SP-0031_T1-0031-001.wav'

LDC-IL Sentence Aligned Speech corpus for Malayalam 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.

8.2 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 Kozhikode region few speakers never pronounce /ɻa/ and instead of /ɻa/ they consistently pronounce it as /ja/.

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.

8.2.1 PHONETIC ALTERNATION IN MALAYALAM 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 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 start is a common phenomenon in most of the speakers and 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.: bʰa-bʰinnaʈakaɭ; ɕa-sannaɖɖʰamaːɳennum

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, if the audio is transcribed as abʰjuːɦamunʈaːji[runnu], it shows that [runnu] is not properly audible. In some longer words, the middle of the syllable or phone might not be audible to the listener or are skipped by the speaker. For example, in ʋjaːpiʧʧukoɳʈiri[kkuka]jaːɳə the middle part is not audible.

d. 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.: bʰaːrgaʋan >bʰaːrgakaʋan; kaːʈʈirunnu>kaːʈʈiʈʈirunnu

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.: abʰjarʈʈʰanajkkum>abʰjarnajkkum

e. 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.: anpaʈə>ampaʈə; ʧĩkkanbirijaːɳi>ʧĩkkambirijaːɳi

Dissimilation is dropping out a syllable or a letter by the influence of adjacent speech segments.

E.g.: aʧʧʰanammamaːr >aʧʧʰanammaːr

f. Colloquial usage

Some of the speakers have pronounced colloquial forms instead of the standardised form written in the prompt sheet. E.g.: pariṯ̪apiʧʧuku:ʈa: >pariḏ̪avɪʧʧuku:ʈa:

g. Lengthening and Shortening

Short and long vowels are interchanged in the recordings at several places.
E.g.: saɦaja:ṯ̪rikan >saɦa:ja:ṯ̪rikan sʊa:ṯ̪anṯ̪rjaḏ̪inam>sʊaṯ̪anṯ̪rjaḏ̪inam

h. **Substandard alternation**

It has been observed that some speakers have consistently replaced the aspirated sounds with their unaspirated counterparts. Moreover, in Kozhikode region people adapted easily articulated sounds.
E.g.: bʰaːrja>pʰaːrja>baːrja; kalarnnu>kalajnnu

i. 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.: uḏ̪gʰaːʈanam>ulgʰaːʈanam

j. Final vowel modification

In continuous speech the final vowel gets modified at times in some of the speakers:
E.g.: muːnnə>muːnne

k. **Common phonetic variation**

While pronouncing a word which starts with ‘ʤa, ra, la and ba’ the Malayalam native speaker invariably changes the inherent vowel ‘a’ to ‘e’.

j. 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.
E.g.: aːjiraṯ̪ṯiṯ̪ṯoɭ̪aːjiraṯ̪ṯimuppaṯ̪ṯijeʈʈə > aːjiraṯ̪ṯi ṯ̪oɭ̪aːjiraṯ̪ṯI muppaṯ̪ṯi eʈʈə

8.3 SUMMARY OF THE CORPUS

The total duration of Malayalam Sentence Aligned Speech Corpus is 123:29:55 (hh:mm:ss) comprising 89,269 audio segments from 451 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 are showing 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.

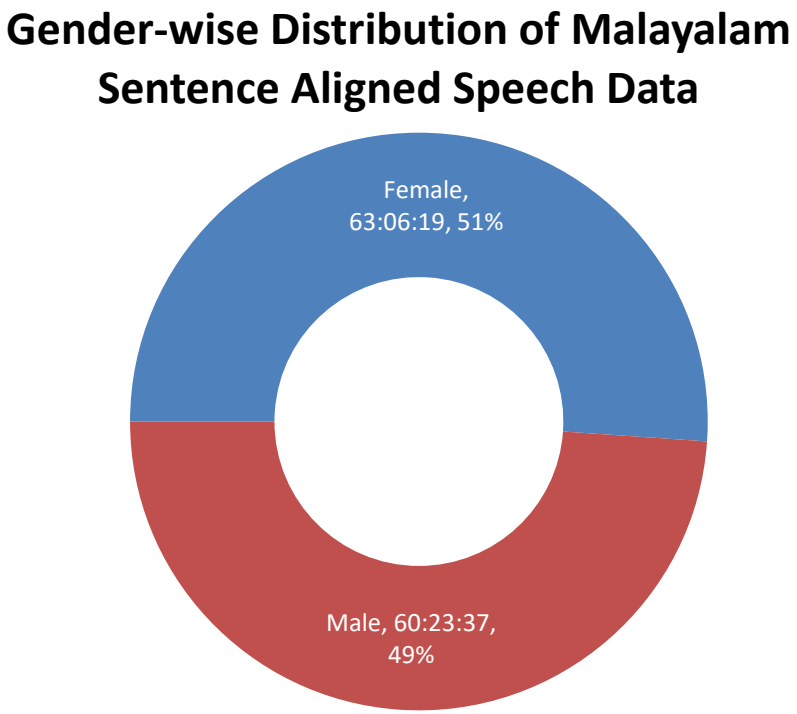

Figure 31: Gender-wise Distribution of Malayalam Corpus

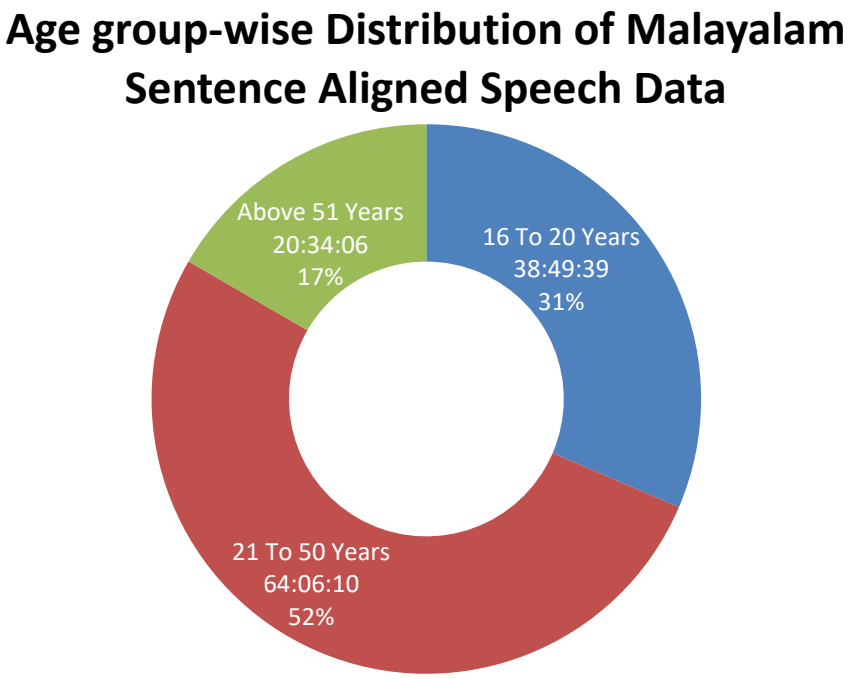

Figure 32: Age-wise Distribution of Malayalam Corpus

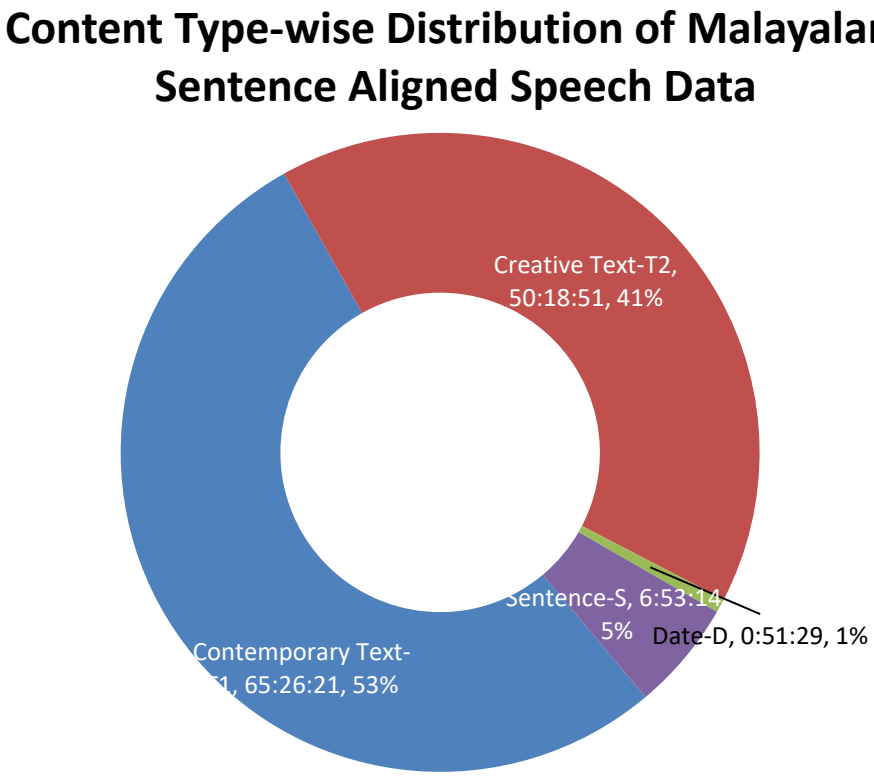

Figure 33: Content Type-wise Distribution of Malayalam Corpus

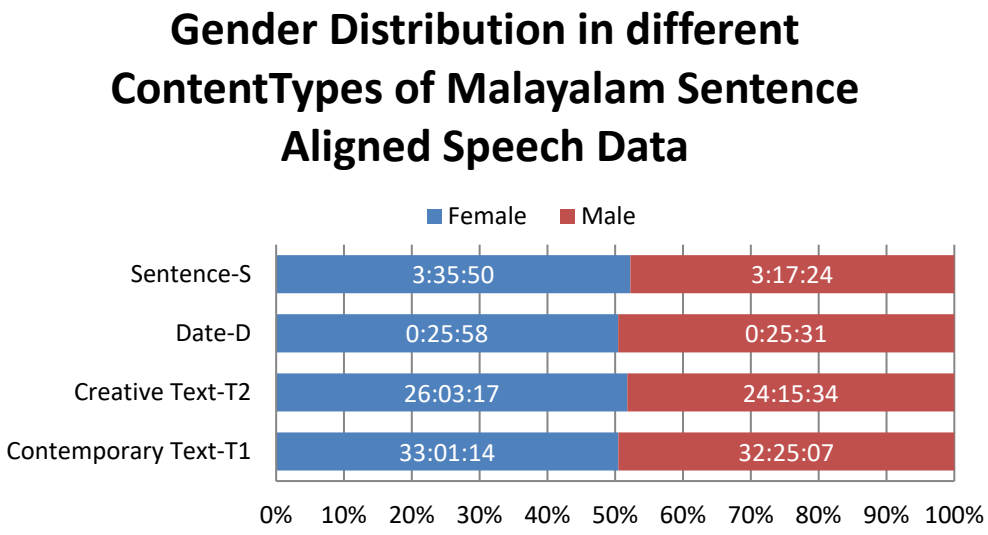

Figure 34: Gender Distribution in different Content Types of Malayalam Corpus

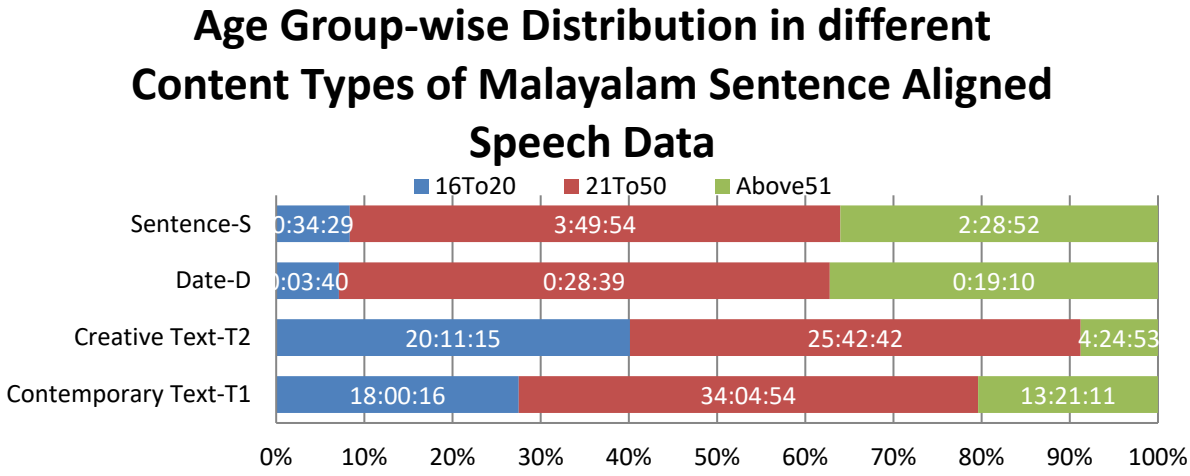

Figure 35: Age Distribution in different Content Types of Malayalam Corpus

8.3.1 DURATION OF MALAYALAM SENTENCE ALIGNED SPEECH DATA

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

Content Type	Gender	Age Group	Duration (hh:mm:ss.ms)		
Contemporary Text-T1	Female	16To20	09:15:44.501863	33:01:14.195152	65:26:21.306209
		21To50	17:07:31.341778		
		Above51	06:37:58.351511		
	Male	16To20	08:44:31.210327	32:25:07.111057	
		21To50	16:57:23.152534		
		Above51	06:43:12.748196		
Creative Text-T2	Female	16To20	10:35:19.614380	26:03:16.547225	50:18:50.676353
		21To50	13:16:23.653412		
		Above51	02:11:33.279433		
	Male	16To20	09:35:55.447264	24:15:34.129128	
		21To50	12:26:18.640281		
		Above51	02:13:20.041584		
Date-D	Female	16To20	00:01:57.028222	00:25:57.983541	00:51:28.869684
		21To50	00:14:27.662812		
		Above51	00:09:33.292507		
	Male	16To20	00:01:42.670001	00:25:30.886143	
		21To50	00:14:11.793371		
		Above51	00:09:36.422771		
Sentence-S	Female	16To20	00:19:02.770006	03:35:49.875736	06:53:14.307104
		21To50	01:58:50.264711		
		Above51	01:17:56.841019		
	Male	16To20	00:15:25.968186	03:17:24.431368	
		21To50	01:51:03.523515		
		Above51	01:10:54.939667		

Table 13: Representation of Malayalam Sentence Aligned Speech Data Duration

8.4 SUMMARY OF SPEAKERS

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

Age Group	Female	Male	Total
16To20	51	48	99
21To50	124	122	246
Above51	54	52	106
Total	229	222	451

Table 14: Distribution of Speakers of Malayalam Sentence Aligned Speech Data

8.5 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. K.S., Rejitha, Saritha S.L., Sajila S., Rajesha N., Manasa G., Narayan Choudhary & L. Ramamoorthy. 2019. “Documentation of LDC-IL Malayalam Raw Speech Corpus" in Linguistic Resources for AI/NLP in Indian Languages. Central Institute of Indian Languages, Mysore. pp. 233-243.
5. Ramamoorthy, L., Narayan Choudhary, Saritha S.L., Rejitha K.S., Sajila S. & Midhun P. G. 2019. Malayalam Raw Speech Corpus. Central Institute of Indian Languages, Mysore.

9 MARATHI SPEECH ANNOTATION

Bhageshree K Khandale, Narayan Kumar Choudhary

9.1 OVERVIEW OF SENTENCE ALIGNED SPEECH CORPUS

Marathi Sentence Aligned Speech Corpus is created by annotating the speech data collected by LDC-IL. A detailed explanation of the Marathi Raw Speech Corpus will be available in the Marathi Speech Data Documentation (Ramamurthy,L.et.Al,2019). LDC-IL Marathi Sentence Aligned Speech files contains an audio file and two textual layers in Marathi 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
'Marathi_Female_25To50_Contemporary Text-T1_SP-0301_T1-0301-001.wav'

LDC-IL Sentence Aligned Speech corpus for Marathi contains read speech from four content type's 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 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 annotation. The orthographically normalised annotation is the prompt text in all of these cases.

9.2 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 Vidharbh region few speakers never pronounce /ɳ/,/ɭ/and instead of /ɳ/,/ɭ/they consistently pronounce it as /n/,/l/.
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.

9.2.1 PHONETIC ALTERNATION IN MARATHI 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 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 start is a common phenomenon in most of the speakers and 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.:ʂə-ʂənmɑːnə(स-सन्मान)

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, if the audio is transcribed as kɑːɾjɑː [ləjɑː]*ʊərə(कार्या[लया]*वर), it shows that [ləjɑː(लया)] is not properly audible. In some longer words, the middle of the syllable or phone might not be audible to the listener or are skipped by the speaker. For example, inmukʰjə*ʈriː[मुख्य*त्री] the middle part is not audible.

d. 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.: ənənəsə > ənəsə (अननस>अनस); eːkəsə̃gʰə>eːkəkəsə̃gʰə (एकसंघ>एककसंघ)

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.: ʂʊəjə̃pɑːkə > ʂʊə̃pɑːkə(स्वयंपाक>स्वंपाक)

e. 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.: eːkəɖɑː> eːgəɖɑː (एकदा>एगदा)

Dissimilation is dropping out a syllable or a letter by the influence of adjacent speech segments.
E.g.: pʰənəsə > pənəsə(फणस>पनस);ɖɔkʈərə>ɖɑːkʈərə(डॉक्टर>डाक्टर)

f. Colloquial usage

Some of the speakers have pronounced colloquial forms instead of the standardised form written in the prompt sheet. E.g.: pə̃ɖʰərɑ:>pə̃ɖərɑ:(पंधरा>पंदरा);mərɑ:ʈʰi:>mərɑ:ʈi:(मराठी>मराटी)

g. Lengthening and Shortening

Short and long vowels are interchanged in the recordings at several places.
E.g.: pɑ:ɳi:>pɑ:ɳɪ(पाणी>पाणि);ʋəki:lə> ʋəkɪlə(वकील>वकिल); ɑ:i:>ɑ:ɪ(आई>आइ)

h. Substandard alternation

It has been observed that some speakers have consistently replaced the aspirated sounds with their unaspirated counterparts. Moreover, in Vidarbh region people adapted easily articulated sounds. E.g.: pə̃kʰɑ:> pə̃kɑ:(पंखा>पंका);ʈɑ:ɾi:kʰə>ʈɑ:ɾi:kə (तारीख>तारीक)

i. Final vowel modification

In continuous speech the final vowel gets modified at times in some of the speakers:
E.g.: mubə̃i:>mubə̃jə; (मुबंई>मुबंय)

j. 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.
E.g.: bɑ:i:mɑ:ɳu:ʂə > bɑ:i:mɑ:ɳu:ʂə (बाईमाणूस>बाईमाणूस); ʂʊəjə̃pɑ:kəgʰərə> ʂʊəjə̃pɑ:kə (स्वयंपाकघर>स्वयंपाक)

9.3 SUMMARY OF THE CORPUS

The total duration of Marathi Sentence Aligned Speech Corpus is 89:17:25 (hh:mm:ss) comprising 58544 audio segments from 307 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.

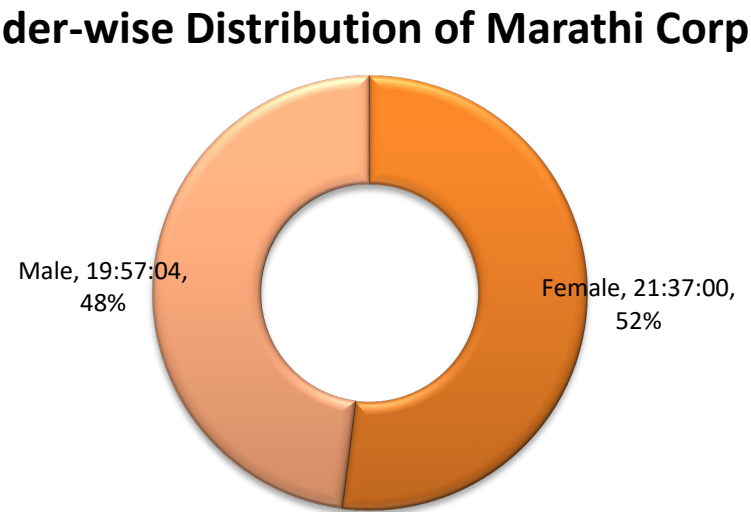

Figure 36: Gender-wise Distribution of Marathi Corpus

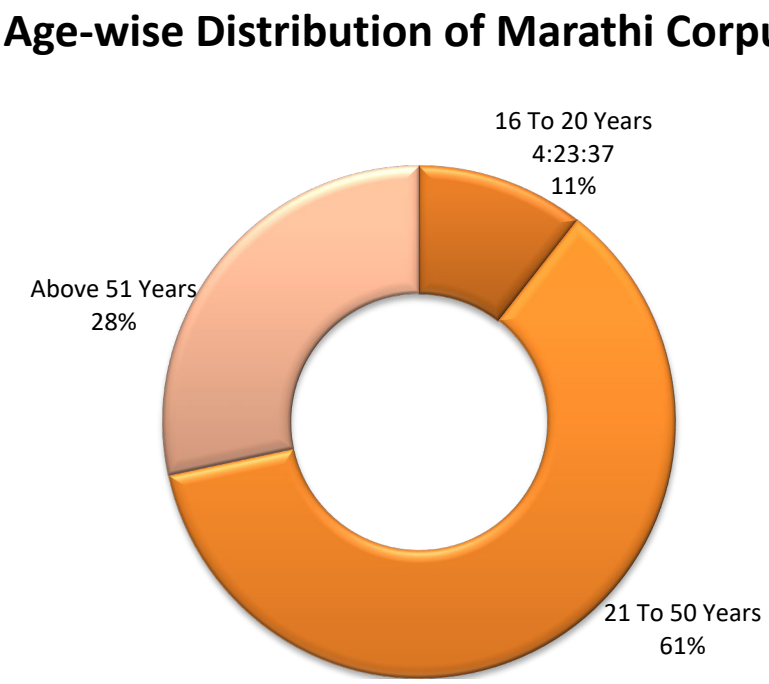

Figure 37: Age-wise Distribution of Marathi Corpus

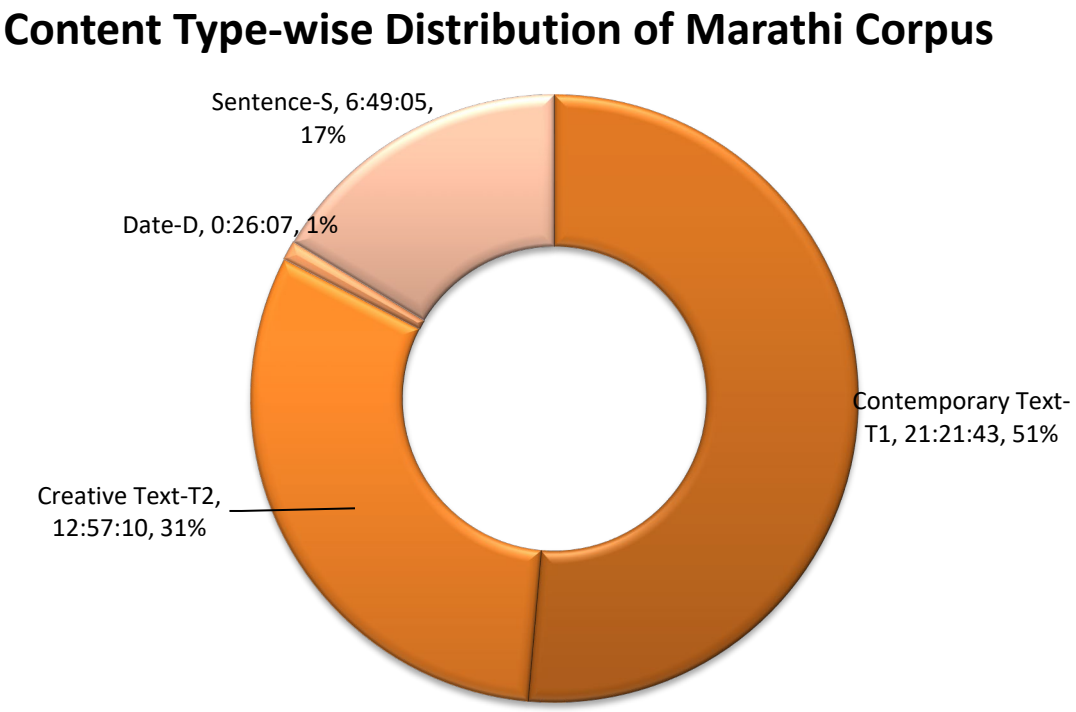

Figure 38: Content Type-wise Distribution of Marathi Corpus

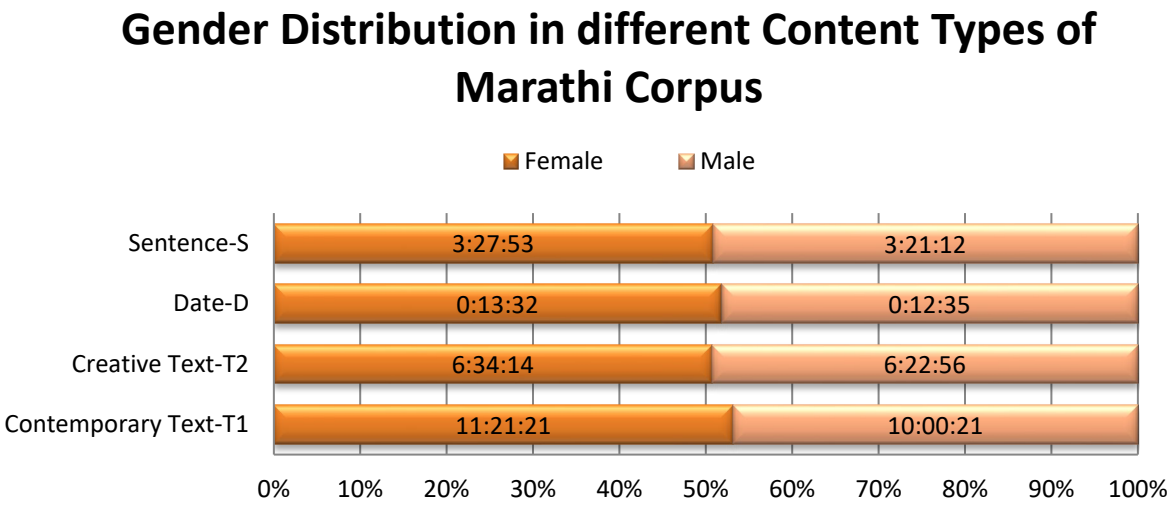

Figure 39: Gender Distribution in different Content Types of Marathi Corpus

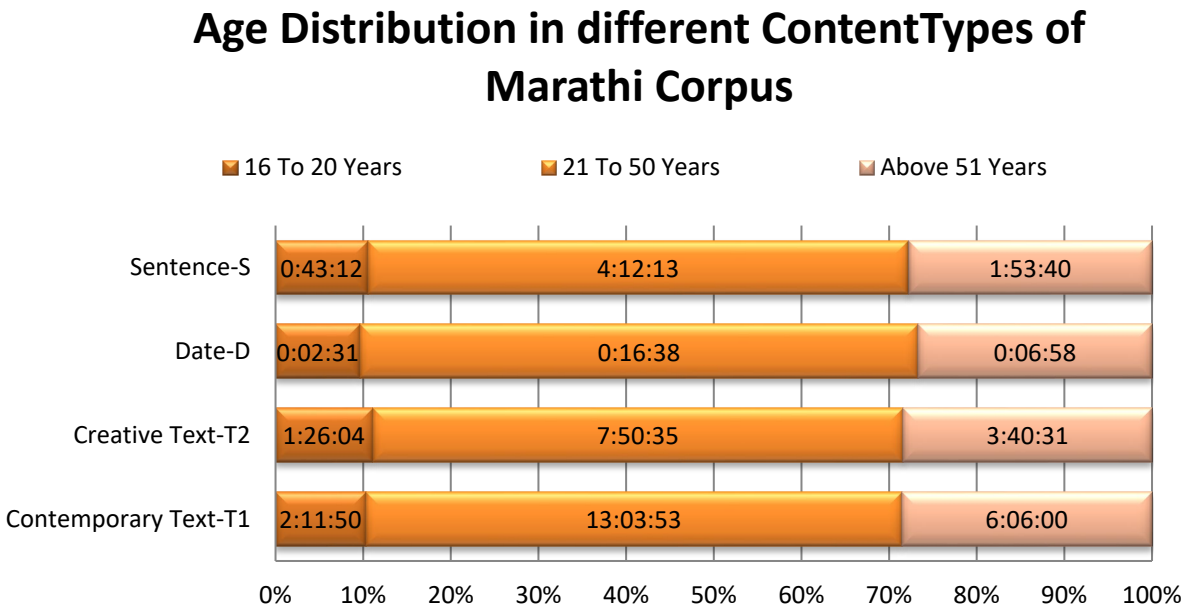

Figure 40: Age Distribution in different Content Types of Marathi Corpus

9.3.1 DURATION OF MARATHI SENTENCE ALIGNED SPEECH DATA

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

Content Type	Gender	Age Group	Duration (hh:mm:ss.ms)		
Contemporary Text-T1	Female	16To20	01:09:03.863110	11:21:21.136796	21:21:42.565240
		21To50	07:13:29.389145		
		Above51	02:58:47.884541		
	Male	16To20	01:02:45.921096	10:00:21.428443	
		21To50	05:50:23.236239		
		Above51	03:07:12.271109		
Creative Text-T2	Female	16To20	00:44:35.195643	06:34:14.106415	12:57:09.812678
		21To50	04:02:12.933200		
		Above51	01:47:25.977571		
	Male	16To20	00:41:28.864523	06:22:55.706264	
		21To50	03:48:22.083914		
		Above51	01:53:04.757826		
Date-D	Female	16To20	00:01:22.358870	00:13:31.959598	00:26:06.920888
		21To50	00:08:42.380087		
		Above51	00:03:27.220641		
	Male	16To20	00:01:08.940600	00:12:34.961290	
		21To50	00:07:55.413621		
		Above51	00:03:30.607068		
Sentence-S	Female	16To20	00:22:10.421093	03:27:53.199855	06:49:04.724197
		21To50	02:10:49.634350		
		Above51	00:54:53.144413		
	Male	16To20	00:21:01.884869	03:21:11.524342	
		21To50	02:01:23.258482		
		Above51	00:58:46.380991		

Table 15: Representation of Marathi Sentence Aligned Speech Data Duration

9.4 SUMMARY OF SPEAKERS

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

Age Group	Female	Male	Total
16To20	17	16	33
21To50	97	91	188
Above51	39	42	81
Total	153	149	302

Table 16: Distribution of Speakers of Marathi Sentence Aligned Speech Data

9.5 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, Gajanan R Apine & Apurva P Betkekar. 2019. Marathi Raw Speech Corpus.Central Institute of Indian Languages, Mysore. ISBN: 978-81-7343-251-4
5. Bhageshree Khandale, Rajesha N., Manasa G., Narayan Choudhary & L. Ramamoorthy. 2019. "Documentation of LDC-IL Marathi Raw Text Corpus" in Linguistic Resources for AI/NLP in Indian Languages. Central Institute of Indian Languages, Mysore. pp. 112-120.

10 NEPALI SPEECH ANNOTATION

Umesh Chamling Rai, Rupesh Rai, Narayan Kumar Choudhary

10.1 OVERVIEW OF SENTENCE ALIGNED SPEECH CORPUS

Nepali Sentence Aligned Speech Corpus is created by annotating the speech data collected by LDC-IL. A detailed explanation of the Nepali Raw Speech Corpus will be available in the Nepali Speech Data Documentation (Ramamurthy, L.et. Al, 2019). LDC-IL Nepali Sentence Aligned Speech files contains an audio file and two textual layers in Nepali 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
'Nepali_Female_16To20_Contemporary Text-T1_SP-0031_T1-0031-001.wav'

LDC-IL Sentence Aligned Speech corpus for Nepali 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 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 annotation. The orthographically normalised annotation is the prompt text in all of these cases.

10.2 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 Darjeeling region few speakers never pronounce /ŋ/ and instead of /ŋ/ they consistently pronounce it as /n/.
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.

10.2.1 PHONETIC ALTERNATION IN NEPALI 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 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 is 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 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.: *biralo-birala ji-jalisə; prə-səmmukʰə; gorəkʰə-gorəkʰapəʈrə*

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, if the audio is transcribed as *'ʃrī [cənrə]* əli pərəi ʈʰijo*', it shows that [*cənrə*] is not properly audible.In some longer words, the middle of the syllable or phone might not be audible to the listener or are skipped by the speaker. For example, in '*jəsə pə[ɖɖə]*ʈibaʈə gəribə kr̥ʂəkəhə[rə]*laī labʱə ɦuncʰə*' the middle part is not audible.

d. 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.: *viʃvəma ənjə kunəi grəhəma nīlo akaʃə rə səmuɖrə cʰəinə> viʃvəma ənjə kunəi grəhəma nīlo akaʃə rə səmunɖrə cʰəinə*
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.: *səɦəbʱagiʈa> səbʱagiʈa, vjəvəɦarə> vjəvarə, bevarə*

e. 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.: *gərĩɖa> gərinɖa; sĩcai> sincai*
Dissimilation is dropping out a syllable or a letter by the influence of adjacent speech segments.
E.g.: *prəʈjekə> prəʈekə, kəɦã> kã*

f. Colloquial usage

Some of the speakers have pronounced colloquial forms instead of the standardised form written in the prompt sheet.
E.g.: *məɦina> məina, jəʈa> eʈa, mə> mo, kənɦəijalaī ɦerẽ> kənəijalaī ɦerẽ*

g. Lengthening and Shortening

Short and long vowels are interchanged in the recordings at several places.
E.g.: Lengthening: *məɦanə > maɦanə, səɦasə> saɦasə,*
Shortening: *aɖʱjaʈmikə> əɖʱjaʈmikə; praɖurbʱavə> prəɖurbʱəvə*

h. Substandard alternation

There are sound variations in the colloquial or some dialect of Nepali.
Throughout the whole speech some speakers uttered some letters consistently in different ways. It might be the habit of speakers. It is observed that some informants replaced the aspirated sounds as unaspirated. This feature has been found in all three Darjeeling, Assamese and Uttarakhand regions, where people adapted easily articulated sounds.
E.g.: *sə̃gə> səŋgə, pĩɖula> piɖũla, snaʈəkə> sʈʰanəkə, uʈsukə> usʈukə*

i. 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.: jugəjugə > ɟugəɟugə, vərʂə> bərsɑː

j. Final vowel modification

In continuous speech the final vowel gets modified at times in some of the speakers:
E.g.: ɑːmɑː>ɑːmə

k. Interchange of Voiced and voiceless

Nepali has voiced and voiceless consonants; some speakers have pronounced voiced consonants as voiceless or vice versa in some instances.
Eg. Voiceless in place of Voiced Consonant:*kiʈabə> kiʈapə, pəsɖa> pəsʈa, lagcʰə> lakcʰə*
Eg. Voiced in place of Voiceless Consonant:*euʈa> euɖa*

l. Interchange of Aspirated to unaspirated

Speakers tend to pronounce aspirated letters in unaspirated, and vice versa across all dialects. Aspirated to unaspirated is more commonly observed in the Darjeeling, Assamese and Uttarakhand region speech.
Eg. Aspirated in place of unaspirated Consonant:
sərəkarə> sərəkʰarə,pəʃcimə> pəʃcʰimə, sukkʰa> sukʰkʰa

Eg. Unaspirated in place Aspirated of Consonant:
soɖʱə> soɖə, məɖʱje> məɖje, buɟnə> buɟʱnə, pəɖʱnu> pəɖnu, sʈʰiʈi> sʈiʈi

m. Interchange of Voiceless fricatives

Nepali has three voiceless fricatives namely, voiceless alveolo-palatal fricative (ɕ), voiceless retroflex fricative [ʂ] and voiceless dental fricative [s]. It is observed that some informants interchange the Voiceless fricatives.
ʃikʂa> sikʂa, puʂpə> puspə, ʃoɖʱə> soɖʱə, sənʈoʂə> sənʈosə, aɖeʃə> aɖesə

10.3 SUMMARY OF THE CORPUS

The total duration of Nepali Sentence Aligned Speech Corpus is 43:04:23 (hh:mm:ss) comprising 21,481audio segments from 346 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.

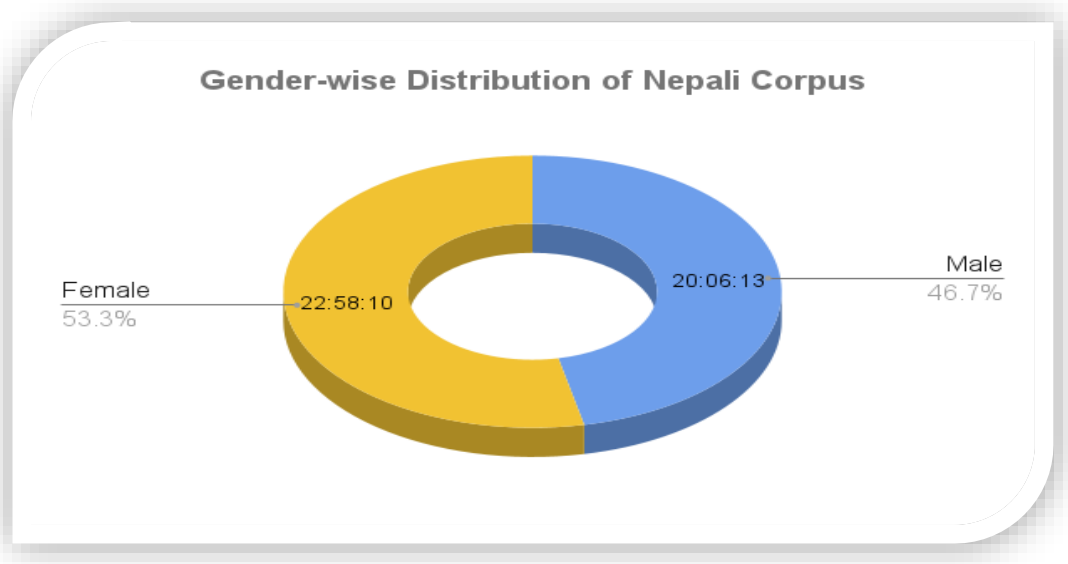

Figure 41: Gender-wise Distribution of Nepali Corpus

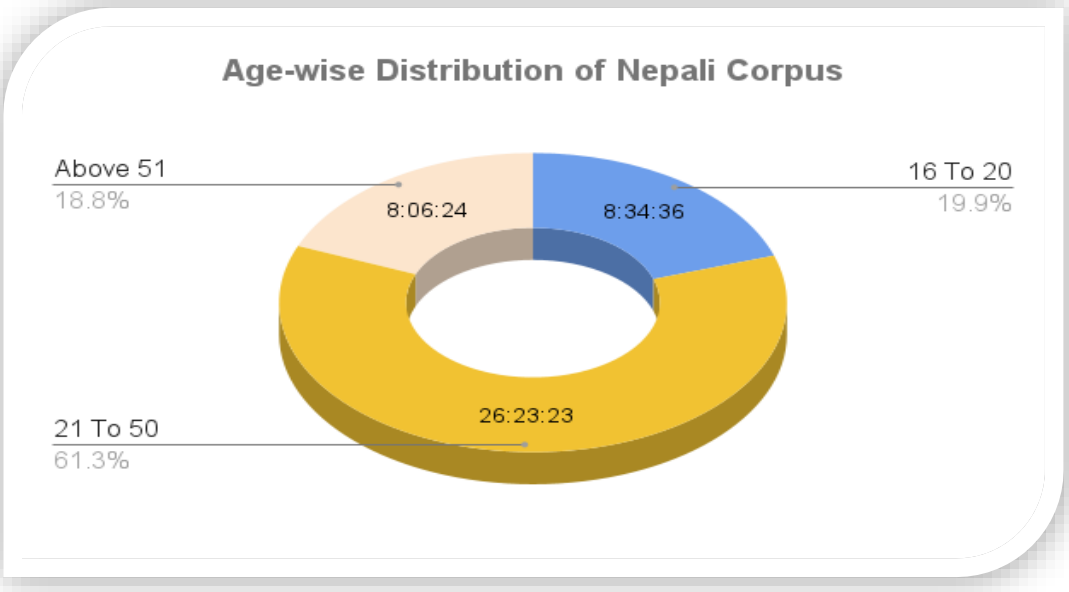

Figure 42: Age-wise Distribution of Nepali Corpus

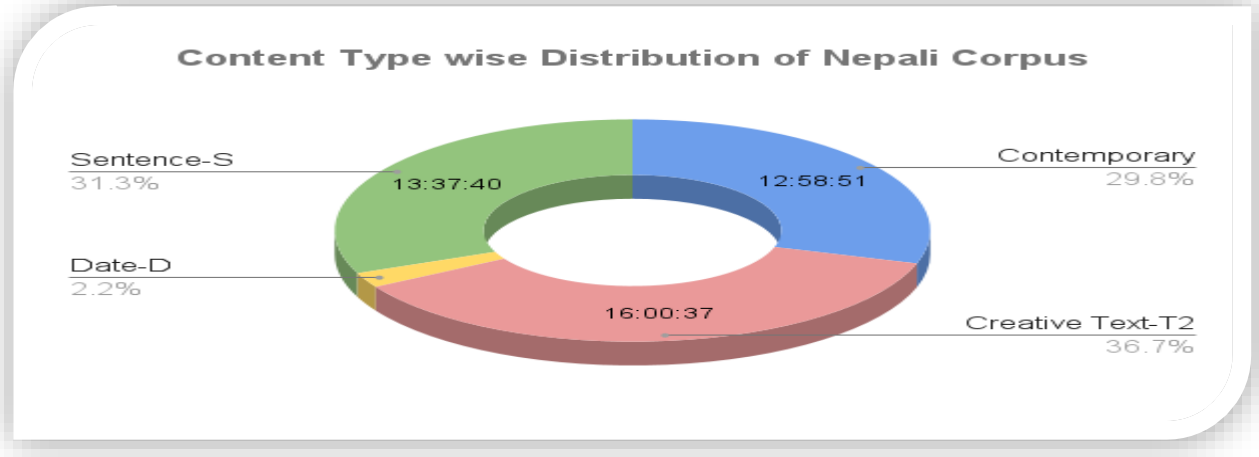

Figure 43: Content Type-wise Distribution of Nepali Corpus

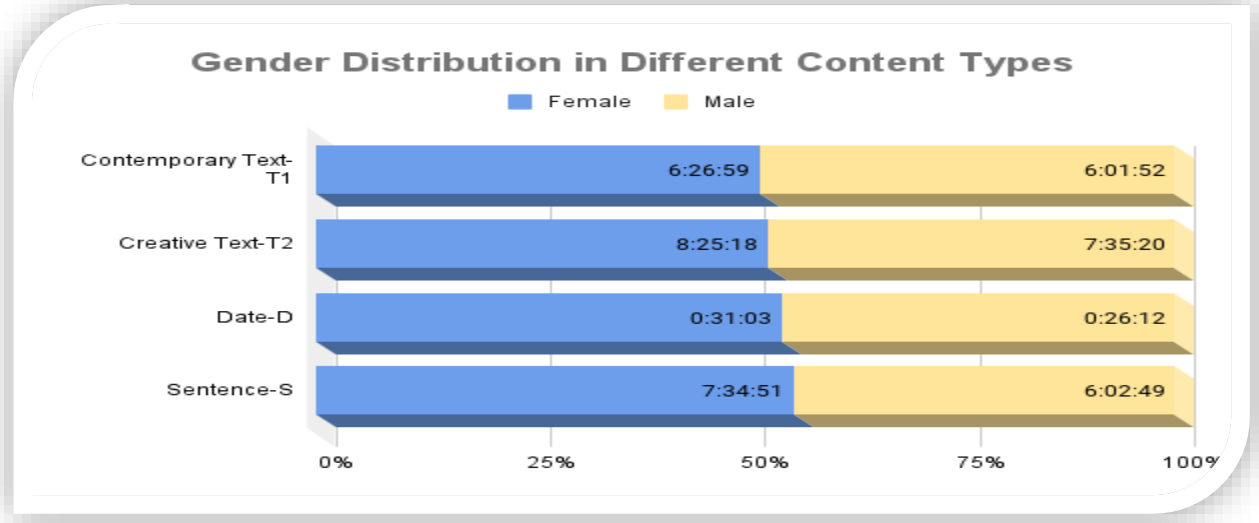

Figure 44: Gender Distribution in different Content Types of Nepali Corpus

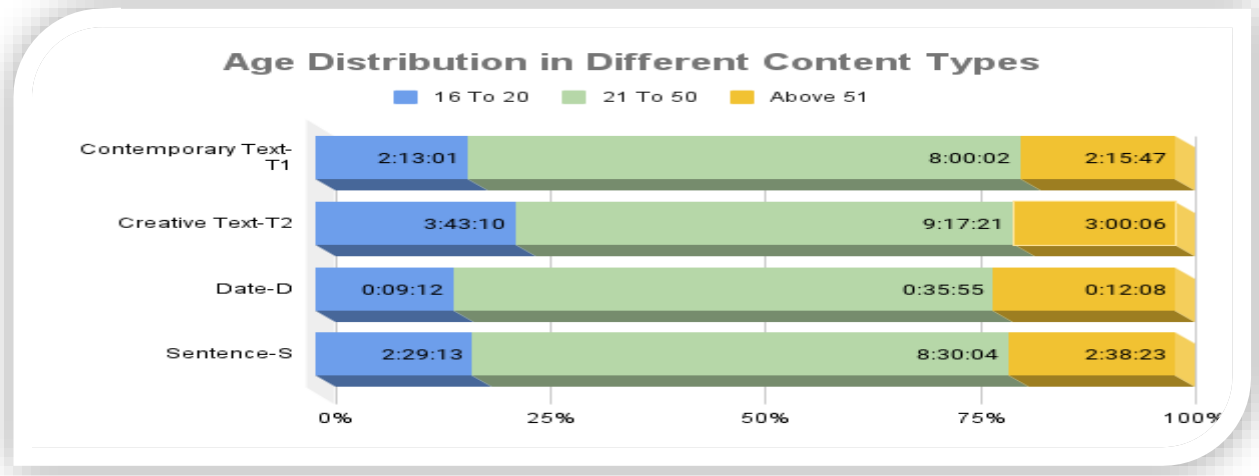

Figure 45: Age Distribution in different Content Types of Nepali Corpus

10.3.1 DURATIONOF NEPALI SENTENCE ALIGNED SPEECH DATA

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

Content Type	Gender	Age Group	Duration (hh:mm:ss.ms)		
Contemporary Text-T1	Female	16To20	01:17:48.178175	06:26:58.853151	12:28:50.755098
		21To50	04:15:16.461072		
		Above51	00:53:54.213905		
	Male	16To20	00:55:13.136145	06:01:51.901947	
		21To50	03:44:45.781300		
		Above51	01:21:52.984502		
Creative Text-T2	Female	16To20	02:00:41.739864	08:25:17.564278	16:00:37.471853
		21To50	05:07:39.926901		
		Above51	01:16:55.897513		
	Male	16To20	01:42:28.664898	07:35:19.907575	
		21To50	04:09:41.495883		
		Above51	01:43:09.746794		
Date-D	Female	16To20	00:05:18.093910	00:31:02.911317	00:57:14.723878
		21To50	00:20:26.069006		
		Above51	00:05:18.748401		
	Male	16To20	00:03:53.446701	00:26:11.812561	
		21To50	00:15:29.186632		
		Above51	00:06:49.179228		
Sentence-S	Female	16To20	01:30:12.422158	07:34:51.124167	13:37:40.277667
		21To50	04:57:50.612331		
		Above51	01:06:48.089678		
	Male	16To20	00:59:00.476626	06:02:49.153499	
		21To50	03:32:13.375968		
		Above51	01:31:35.300906		

Table 17: Representation of Nepali Sentence Aligned Speech Data Duration

10.4 SUMMARY OF SPEAKERS

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

Age Group	Female	Male	Total
16To20	35	25	60
21To50	124	94	218
Above51	28	40	68
Total	187	159	346

Table 18: Distribution of Speakers of Nepali Sentence Aligned Speech Data

10.5 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, Samar Sinha, Jeena Rai, Umesh Chamling Rai & Rupesh Rai. 2019. A Gold Standard Nepali Raw Text Corpus. Central Institute of Indian Languages, Mysore. ISBN: 978-81-7343-255-2.
5. Ramamoorthy, L., Narayan Choudhary, Samar Sinha, Jeena Rai, Umesh Chamling Rai & Rupesh Rai. 2019. Nepali Raw Speech Corpus. Central Institute of Indian Languages, Mysore. ISBN: 978-81-7343-255-2.

11 ODIA SPEECH ANNOTATION

Santosh Kumar Mohanty, Narayan Kumar Choudhary

11.1 OVERVIEW OF SENTENCE ALIGNED SPEECH CORPUS

Odia Sentence Aligned Speech Corpus is created by annotating the speech data collected by LDC-IL. A detailed explanation of the Odia Raw Speech Corpus will be available in the Odia Speech Data Documentation (Ramamoorthy, L. et. al, 2021). LDC-IL Odia Sentence Aligned Speech files contain an audio file and two textual layers in Odia 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
'Odia_Female_16To20_Contemporary Text-T1_SP-0031_T1-0031-001.wav'

LDC-IL Sentence Aligned Speech Corpus for Odia 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 only the response of the speaker. The corpus consists of an audio file for each recording and corresponding textual layer consisting of the phonetically normalised annotation. The orthographically normalised annotation is the prompt text in all of these cases.

11.2 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.
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.
In Odia it is, as found in this speech corpus, very rare that the spoken forms of long vowels- /i:/ and /u:/ and the corresponding secondary forms occur. The occurrence of the palatal /tʃʰ/ as well as fricative /ʃʰ/ is apparent in a rare manner. Interestingly, the two phonemes /dʒ/ and /z/ are pronounced as /dʒ/ only even though orthographically written differently in the words they appear.

11.2.1 PHONETIC ALTERNATION IN ODIA 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 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 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.: bʰaː-bʰaːrətɔ; kɔ-kɔlɔmɔ

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, if the audio is transcribed as [apubɔ]*, it shows that the word/apurbɔ/is not properly audible.

d. 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.: jaːuʧʰɔi jaːuaʧʰɔi
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.: raːʤaː >rɔʤaː

e. 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.: ʧɔkrɔ >ʧɔkɔ
Dissimilation is dropping out a syllable or a letter by the influence of adjacent speech segments.
E.g.: dʰɔnɔiaː > dʰɔɳɔiaː

f. Colloquial usage
Some of the speakers have pronounced colloquial forms instead of the standardised form written in the prompt sheet.
E.g.: gʰɔruku>gʰɔrɔku

g. Substandard alternation
It has been observed that some speakers have consistently replaced the aspirated sounds with their unaspirated counterparts.
E.g.: bɻɔbɔstʰaː >bɻɔbɔstaː; kɔnɔiʂtʰɔ >kɔnɔiʂtɔ

h. Free 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.: laːu> naːu; baːigɔɳɔ >baːiɜ̃gɔɳɔ

i. Final vowel modification
In continuous speech the final vowel gets modified at times in some of the speakers:
E.g.: jeuɜ̃ > jɔu

j. Compound word splitting
The compound 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.
E.g.: pɔdmɔloʧɔnɔ >pɔdmɔ loʧɔnɔ

11.3 SUMMARY OF THE CORPUS

The total duration of Odia Sentence Aligned Speech Corpus is 69:07:50 (hh:mm:ss) comprising 43,448 audio segments from 450 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 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.

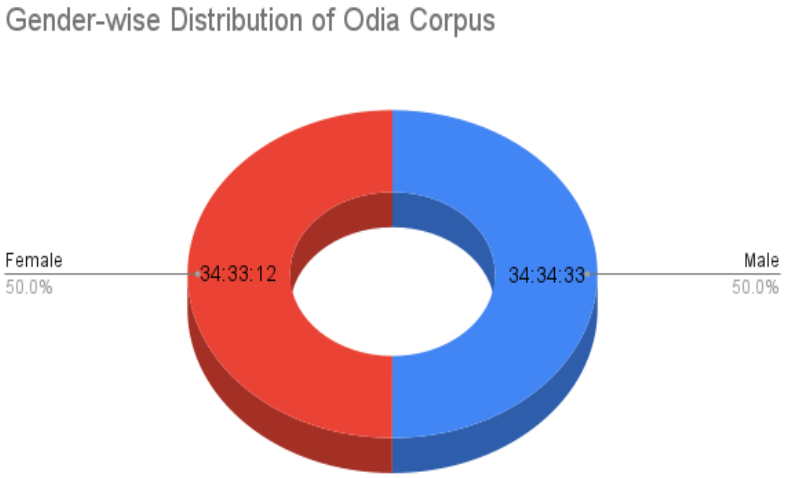

Figure 46: Gender-wise Distribution of Odia Corpus

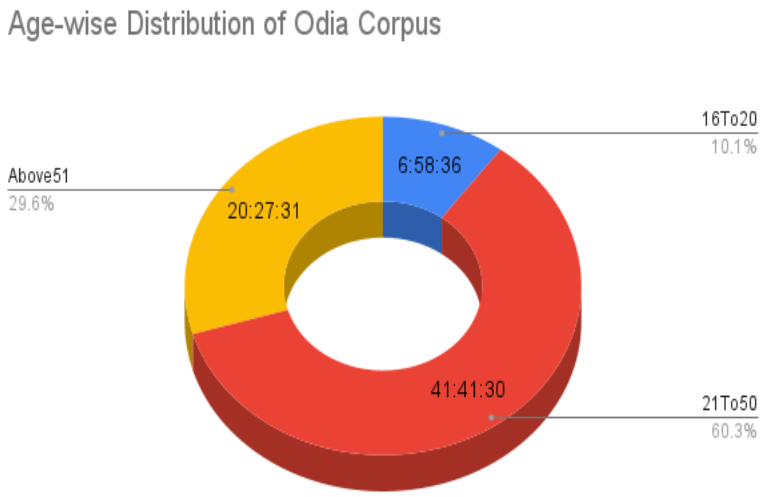

Figure 47: Age-wise Distribution of Odia Corpus

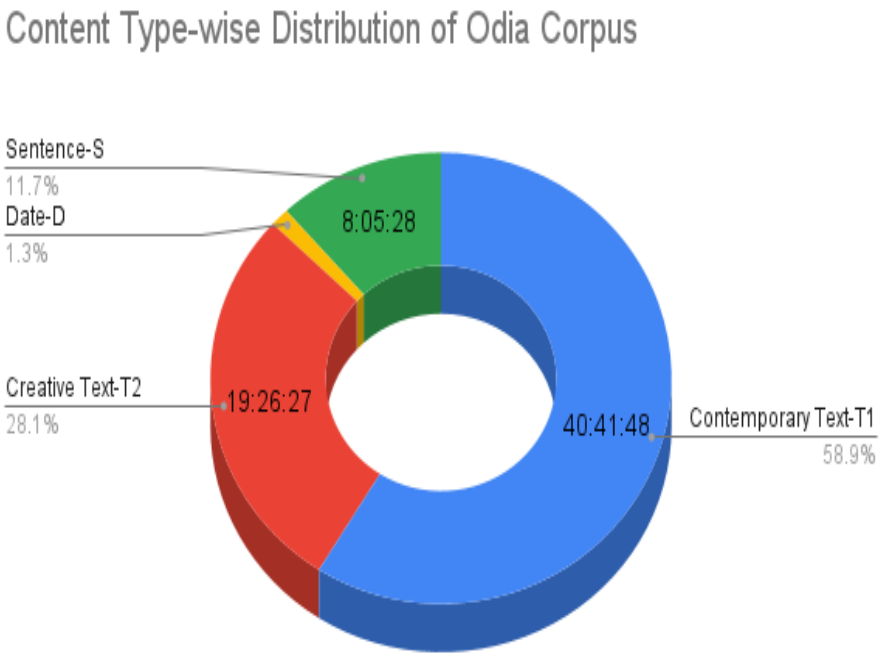

Figure 48: Content Type-wise Distribution of Odia Corpus

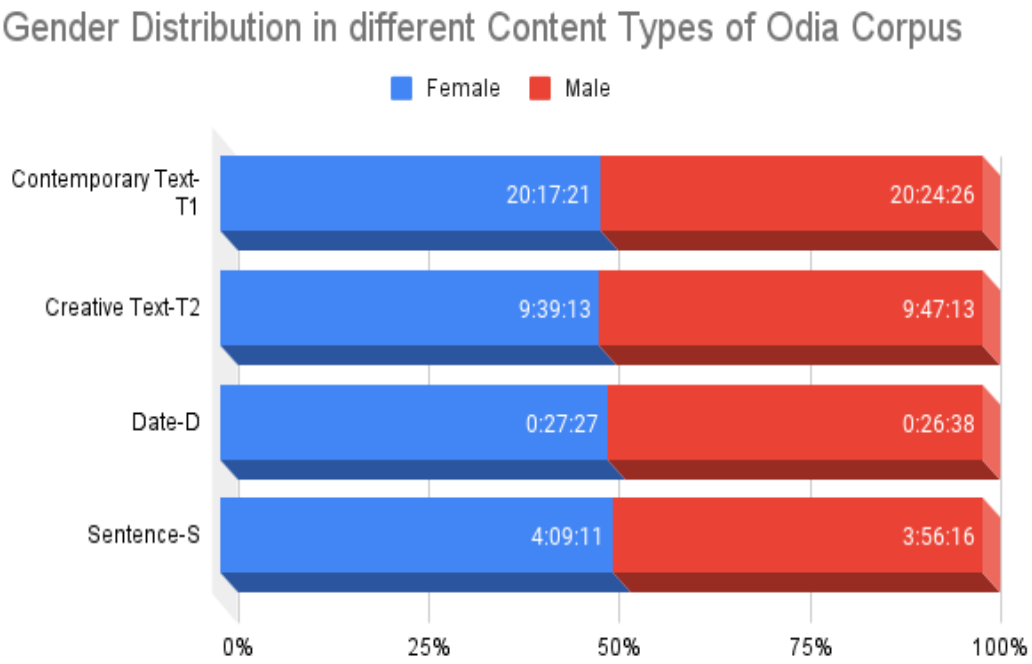

Figure 49: Gender Distribution in different Content Types of Odia Corpus

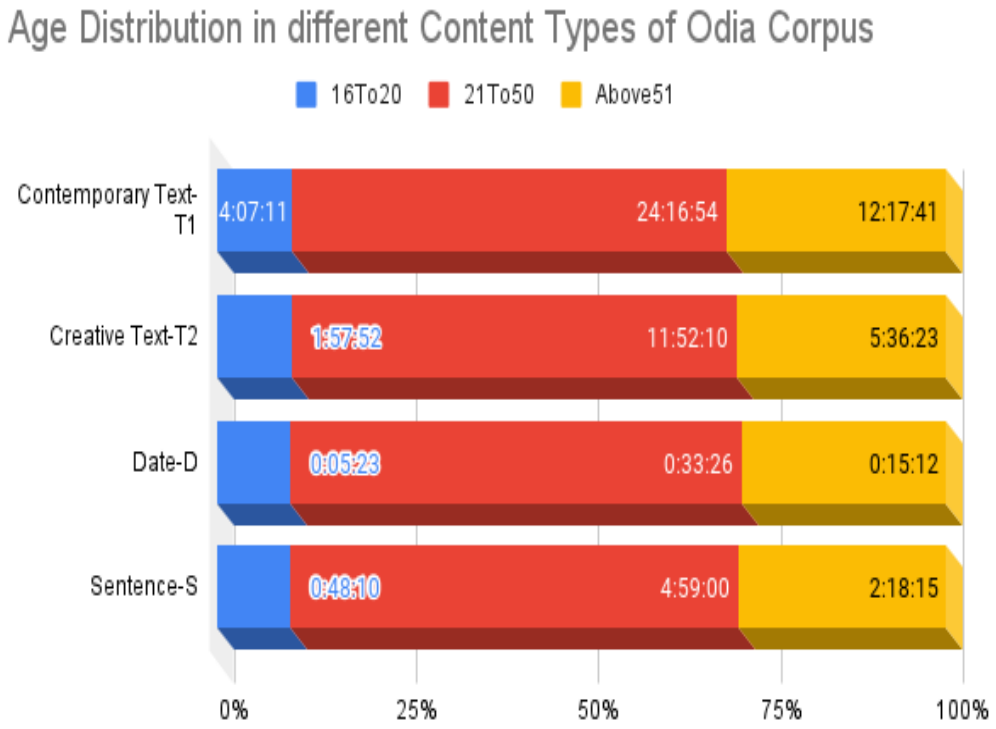

Figure 50: Age Distribution in different Content Types of Odia Corpus

11.3.1 DURATION OF ODIA SENTENCE ALIGNED SPEECH DATA

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

Content Type	Gender	Age Group	Duration (hh:mm:ss.ms)		
Contemporary Text-T1	Female	16To20	01:51:44.582985	20:17:21.207222	40:41:48.120178
		21To50	12:18:22.141428		
		Above51	06:07:14.482808		
	Male	16To20	02:15:27.255307	20:24:26.912956	
		21To50	11:58:32.491294		
		Above51	06:10:27.166355		
Creative Text-T2	Female	16To20	00:55:30.446158	09:39:13.922383	19:26:27.831432
		21To50	06:00:42.906791		
		Above51	02:43:00.569434		
	Male	16To20	01:02:22.366946	09:47:13.909049	
		21To50	05:51:28.320609		
		Above51	02:53:23.221495		
Date-D	Female	16To20	00:02:35.722483	00:27:27.344698	00:54:05.480509
		21To50	00:17:06.827494		
		Above51	00:07:44.794720		
	Male	16To20	00:02:48.673422	00:26:38.135811	
		21To50	00:16:20.468914		
		Above51	00:07:28.993475		
Sentence-S	Female	16To20	00:23:56.501821	04:09:11.392170	08:05:28.373433
		21To50	02:34:40.880005		
		Above51	01:10:34.010344		
	Male	16To20	00:24:14.816522	03:56:16.981262	
		21To50	02:24:20.711123		
		Above51	01:07:41.453618		

Table 19: Representation of Odia Sentence Aligned Speech Data Duration

11.4 SUMMARY OF SPEAKERS

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

Age Group	Female	Male	Total
16To20	23	25	48
21To50	142	138	280
Above51	61	61	122
Total	226	224	450

Table 20: Distribution of Speakers of Odia Sentence Aligned Speech Data

11.5 REFERENCES

1. Choudhary, N. and D. G. Rao. 2020."The LDC-IL Speech Corpora". In *Proceedings of the 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". In *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.
4. Ramamoorthy, L., Narayan Choudhary, Santosh Kumar Mohanty, Raja Kumar Naik, Pramod Kumar Rout & Kshirod Kumar Das. 2019. *A Gold Standard Odia Raw Text Corpus*. Central Institute of Indian Languages, Mysore.
5. Ramamoorthy, L., Narayan Choudhary, Raja Kumar Naik, Pramod Kumar Rout, Kshirod Kumar Das & Santosh Kumar Mohanty. 2021. *Odia Raw Speech Corpus*. Central Institute of Indian Languages, Mysore.

12 TAMIL SPEECH ANNOTATION

Amudha R., Kamaraj S., Narayan Kumar Choudhary

12.1 OVERVIEW OF SENTENCE ALIGNED SPEECH CORPUS

Tamil Sentence Aligned Speech Corpus is created by annotating the speech data collected by LDC-IL. A detailed explanation of the Tamil Raw Speech Corpus will be available in the Tamil Speech Data Documentation (Ramamurthy, L. et. Al, 2019). LDC-IL Tamil Sentence Aligned Speech files contain an audio file and two textual layers in Tamil 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
'Tamil_Female_16To20_Contemporary Text-T1_SP-0031_T1-0031-001.wav'

LDC-IL Sentence Aligned Speech corpus for Tamil 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 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 annotation. The orthographically normalised annotation is the prompt text in all of these cases.

12.2 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 Vazhaippazham region few speakers never pronounce /ɻa/ and instead of /ɻa/ they consistently pronounce it as /la/.
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.

12.2.1 PHONETIC ALTERNATION IN TAMIL 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 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 start is a common phenomenon in most of the speakers and some speakers it is frequent. Usually, it is the repetition of the first word or syllable of the word but sometimes speaker start with some other letter as well.
E.g.: vA-pOn'An', EE-vA ing'kE

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, if the audio is transcribed as vaNTi allatu mOTTAr kAr allatu mitivaNTi [kAr], it shows that [kAr] is not properly audible. In some longer words, the middle of the syllable or phone might not be audible to the listener or are skipped by the speaker. For example, in vaNTi allatu mOTTAr kA allatu mitivaNTithe middle part is not audible.

d. 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.: ceypavar >ceyvavar; pirAyANam >ppirAyANam
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.: a Avai eZutittArung'kaL>a Avai eZutittAng'kaL

e. 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.: eNpatu >empatu;vETam>vEsham;munnURu>muNNURu
Dissimilation is dropping out a syllable or a letter by the influence of adjacent speech segments.
E.g.: maramvElai >maravElai

f. Colloquial usage
Some of the speakers have pronounced colloquial forms instead of the standardised form written in the prompt sheet.
E.g.: paZanj'cORu>paZaiyacORu; talai > talay

g. Lengthening and Shortening
Short and long vowels are interchanged in the recordings at several places.
E.g.: aNi > ANi Etu > etu

h. 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.: kOvam >kOpam

i. Final vowel modification
In continuous speech the final vowel gets modified at times in some of the speakers:
E.g.: mu:npu > mu:npe

j. Common phonetic variation
While pronouncing a word which starts with 'ka, pa and i' the Tamil native speaker invariably changes the inherent vowel 'ai' to 'a'.

k. 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.
E.g.: uNavaiccAppiTTukkoNTirukkumpOtE>uNavaic cAppiTTuk koNTirukkum pOtE

12.3 SUMMARY OF THE CORPUS

The total duration of Tamil Sentence Aligned Speech Corpus is 139:11:41 (hh:mm:ss) comprising 60,287 audio segments from 452 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.

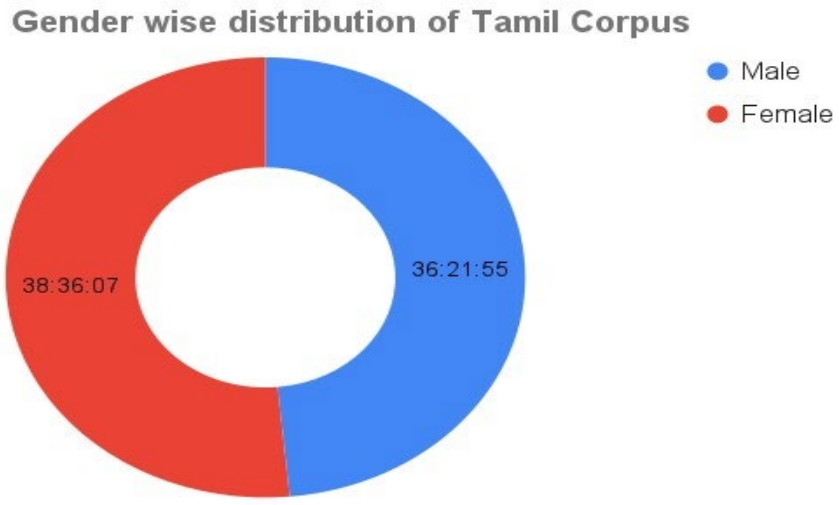

Figure 51: Gender-wise Distribution of Tamil Corpus

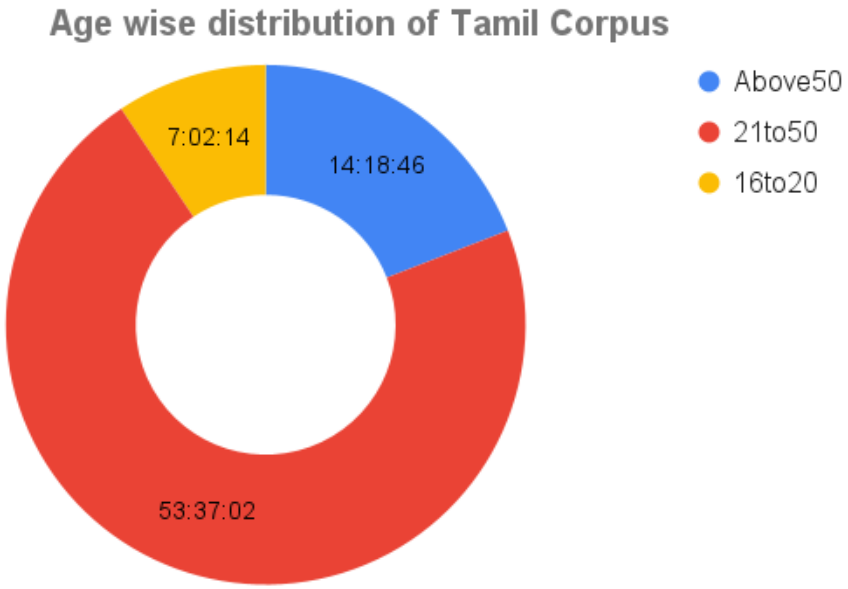

Figure 52: Age-wise Distribution of Tamil Corpus

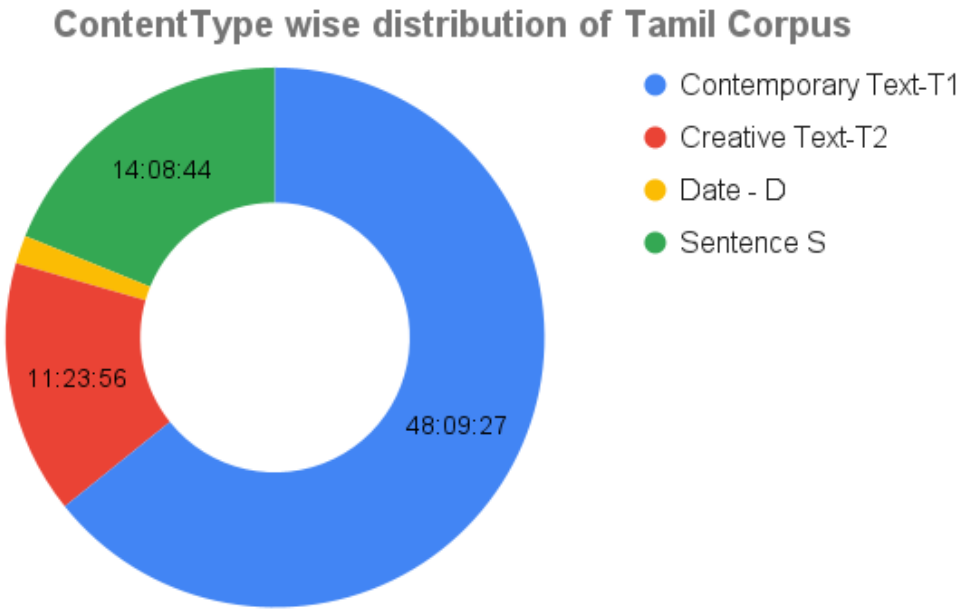

Figure 53: Content Type-wise Distribution of Tamil Corpus

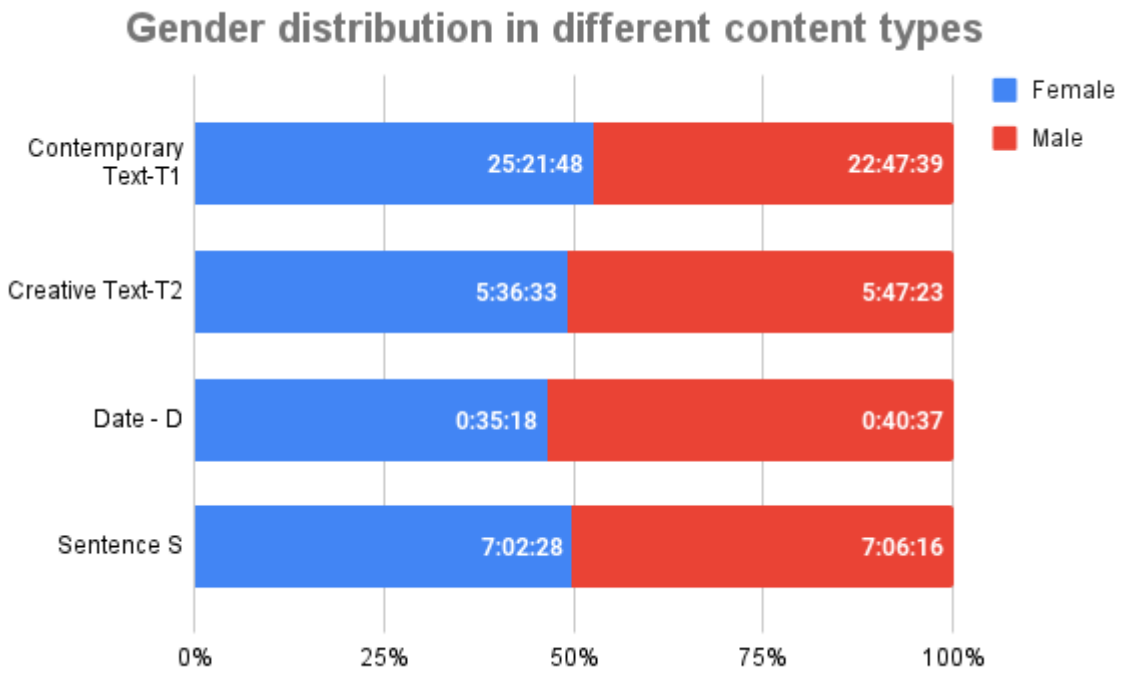

Figure 54: Gender Distribution in different Content Types of Tamil Corpus

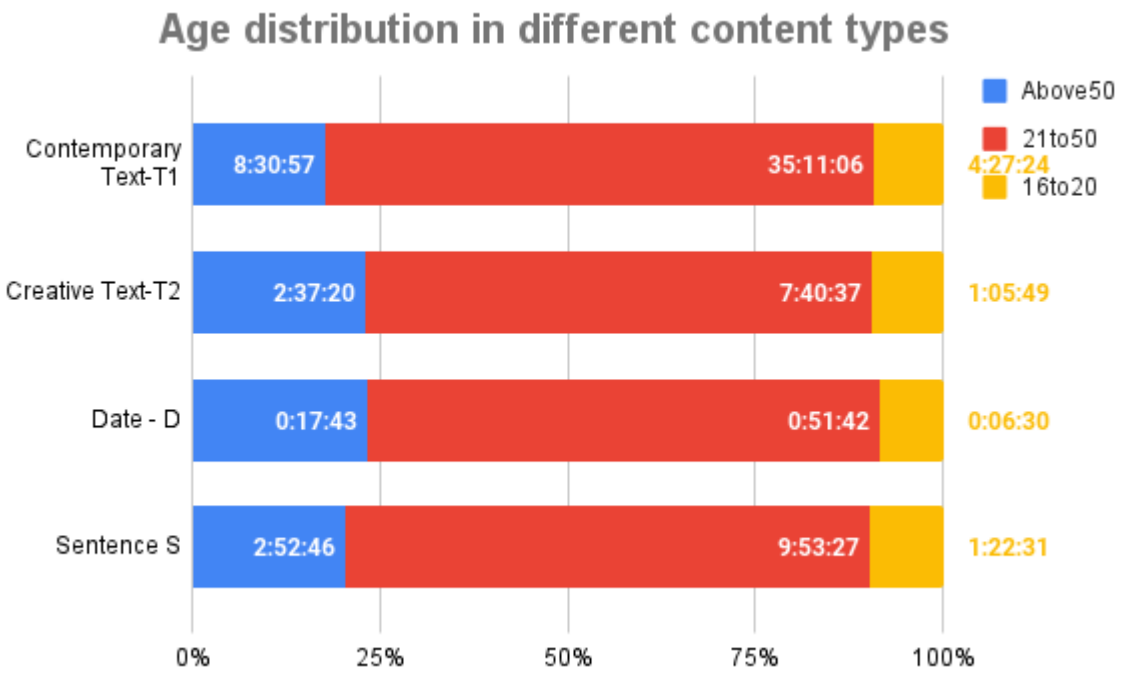

Figure 55: Age Distribution in different Content Types of Tamil Corpus

12.3.1 DURATION OF TAMIL SENTENCE ALIGNED SPEECH DATA

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

Content Type	Gender	Age Group	Duration (hh:mm:ss.ms)		
Contemporary Text-T1	Female	16To20	02:44:00.792117	25:21:47.531683	48:09:26.899929
		21To50	18:22:09.862490		
		Above51	04:15:36.877076		
	Male	16To20	01:43:22.859529	22:47:39.368247	
		21To50	16:48:56.346624		
		Above51	04:15:20.162093		
Creative Text-T2	Female	16To20	00:31:59.646245	05:36:32.697937	11:23:55.555935
		21To50	03:47:17.055709		
		Above51	01:17:15.995983		
	Male	16To20	00:33:49.140795	05:47:22.857998	
		21To50	03:53:29.825104		
		Above51	01:20:03.892099		
Date-D	Female	16To20	00:03:31.576500	00:35:17.851852	01:15:54.116604
		21To50	00:24:02.993432		
		Above51	00:07:43.281920		
	Male	16To20	00:02:57.680771	00:40:36.264753	
		21To50	00:27:38.762731		
		Above51	00:09:59.821251		
Sentence-S	Female	16To20	00:46:00.572510	07:02:27.255198	14:08:42.603582
		21To50	04:54:40.923627		
		Above51	01:21:45.759061		
	Male	16To20	00:36:29.559549	07:06:15.348384	
		21To50	04:58:46.066909		
		Above51	01:30:59.721927		

Table 21: Representation of Tamil Sentence Aligned Speech Data Duration

12.4 SUMMARY OF SPEAKERS

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

Age Group	Female	Male	Total
16To20	22	18	40
21To50	144	145	289
Above51	48	56	104
Total	214	219	433

Table 22: Distribution of Speakers of Tamil Sentence Aligned Speech Data

12.5 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, Thennarasu S, Prem Kumar L R, Amudha R, Prabagaran R, Srikanth D. 2021. Tamil Raw Speech Corpus. Central Institute of Indian Languages, Mysore.

5. Narayan Choudhary, 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.

13 URDU SPEECH ANNOTATION

Shahnawaz Alam, Mansoor Khan, Bi Bi Mariyam, Narayan Kumar Choudhary

13.1 OVERVIEW OF SENTENCE ALIGNED SPEECH CORPUS

Urdu Sentence Aligned Speech Corpus is created by annotating the speech data collected by LDC-IL. A detailed explanation of the Urdu Raw Speech Corpus will be available in the Urdu Speech Data Documentation (Ramamurthy, L. et. Al, 2019). LDC-IL Urdu Sentence Aligned Speech files contain an audio file and two textual layers in Urdu 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
'Urdu_Female_16To20_Contemporary Text-T1_SP-0031_T1-0031-001.wav'

LDC-IL Sentence Aligned Speech corpus for Urdu 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 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 annotation. The orthographically normalised annotation is the prompt text in all of these cases.

13.2 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 Bihar region few speakers never pronounce /ڑ/ /Ra/ and instead of /ڑ/ /Ra/ they consistently pronounce it as /ر/ /ra/, and in the region of South India few speakers never pronounce /ق/ /qa/ and instead of /ق/ /qa/ they consistently pronounce it as /خ/ /Kʰa/.
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.

13.2.1 PHONETIC ALTERNATION IN URDU 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 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.

b. False start

False start is a common phenomenon in most of the speakers and 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.: ب۔باہر(ba-baahar); چ۔چار(cha-chaar)

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, if the audio is transcribed as daulatman[d], it shows that [d] is not properly audible. In some compound words, the middle of the syllable or phone might not be audible to the listener or are skipped by the speaker. For example, in تراش [گ]سن (san[ga]trash) the middle part is not audible.

d. 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.: انتظام(intizaam) >انتظاظام (intizaa-zaam); گیارہ (gayaarah) >اگیارہ (igayaarah)
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.: دلدار(dildaar) >دلدا (dildaa); تالیاں (taaliyaaN) >تالیا (taaliyaa)

e. 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.: ملاقات(mulaaqaat) >ملاکات (mulaakaat); صفائی (safaaii) >صپائی (sapaaii)
Dissimilation is dropping out a syllable or a letter by the influence of adjacent speech segment.
E.g.: سرکار (sarkaar)>سرکا (sakaar)

f. Colloquial usage

Some of the speakers have pronounced colloquial forms instead of the standardised form written in the prompt sheet.
E.g.: نوے (navve) >نبے(nabbe)

g. Lengthening and Shortening

Short and long vowels are interchanged in the recordings at several places.
E.g.: خواہش (khwaahish) >خواہیش (khwahiish); چالیس (chaalees) >چالس (chaalis);

h. Substandard alternation

It has been observed that some speakers have consistently replaced the aspirated sounds with their unaspirated counterparts.
E.g.: بھائی (b^{h}aaii) >بائی (baaii); سبھی(sab^{h}ii)>سبی (sabii)

i. 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.: شام(sham) >سام (saam); خانہ(K^{h}aana) >کھانا (k^{h}aana); غم (Gam) >گم (gam)

j. Final vowel modification

In continuous speech the final vowel gets modified at times in some of the speakers:
E.g.:کہ (ke) >کی (ki) [with short vowel 'i']

13.3 SUMMARY OF THE CORPUS

Below section is providing the tabular details of the different content types of the Urdu Sentence Aligned Speech Corpus. These figures may be helpful in tuning the corpus for various purposes of training, testing and evaluating various algorithms as well as provide useful insights into the dataset. The total duration of Urdu Sentence Aligned Speech Corpus is 50:09:56 (hh:mm:ss) comprising 32,384 audio segments from 434 speakers.

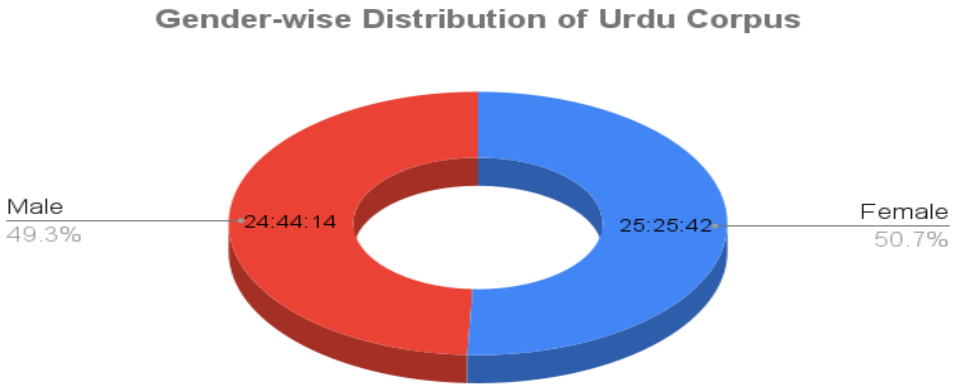

Figure 56: Gender-wise Distribution of Urdu Corpus

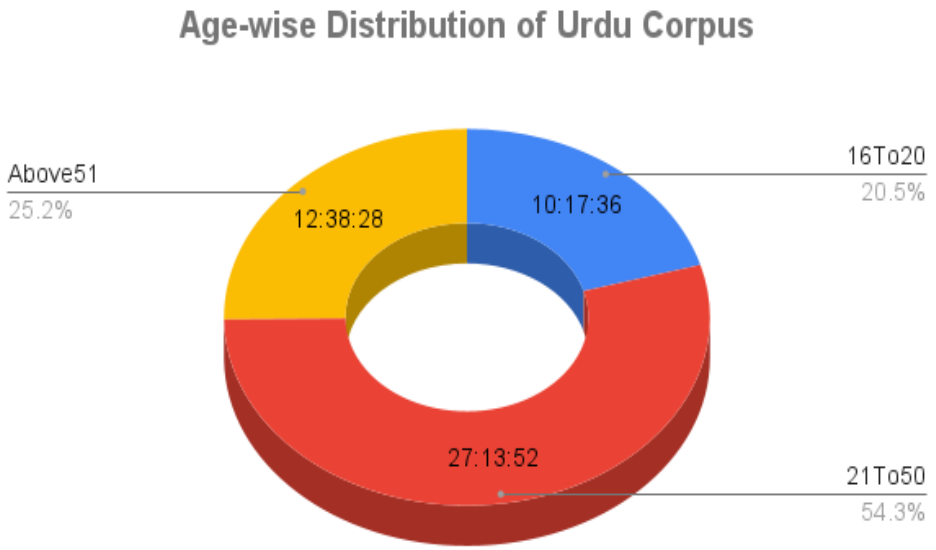

Figure 57: Age-wise Distribution of Urdu Corpus

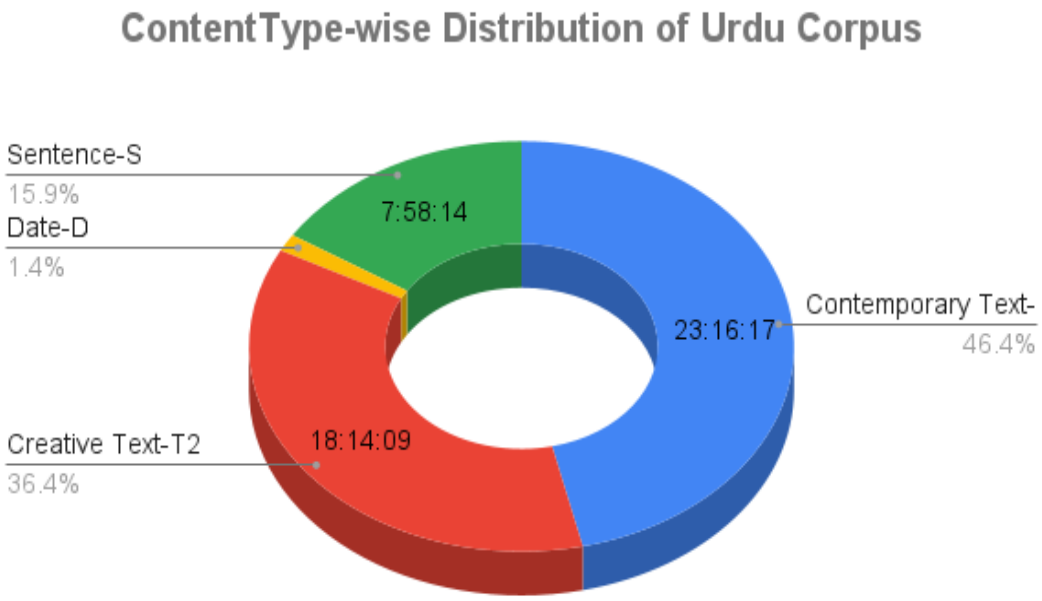

Figure 58: Content Type-wise Distribution of Urdu Corpus

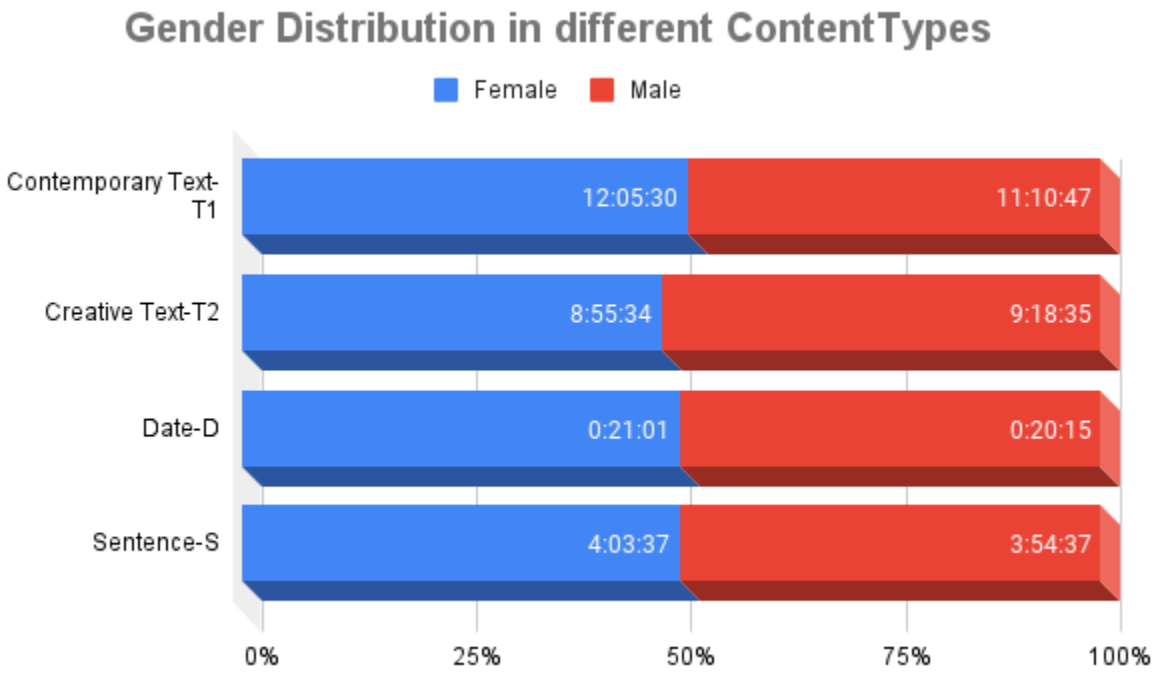

Figure 59: Gender Distribution in different Content Types of Urdu Corpus

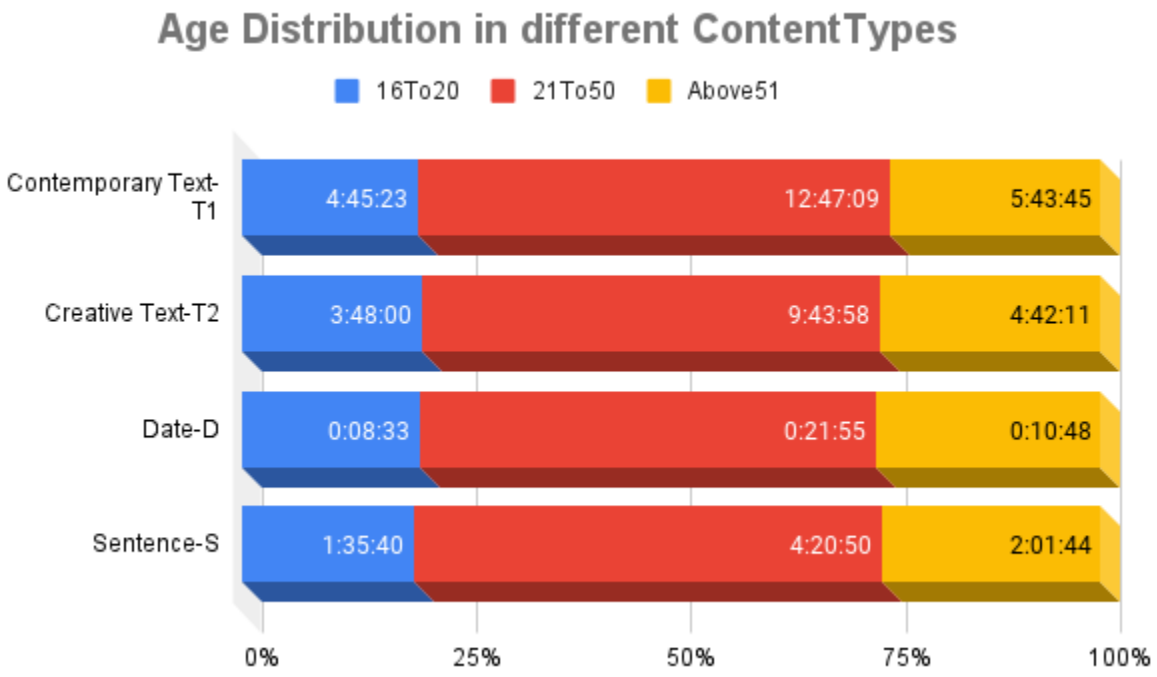

Figure 60: Age Distribution in different Content Types of Urdu Corpus

13.3.1 DURATION OF URDU SENTENCE ALIGNED SPEECH DATA

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

Content Type	Gender	Age Group	Duration (hh:mm:ss.ms)		
Contemporary Text-T1	Female	16To20	02:46:56.948528	12:05:29.225194	23:16:15.920631
		21To50	06:21:27.541888		
		Above51	02:57:04.734778		
	Male	16To20	01:58:25.731017	11:10:46.695438	
		21To50	06:25:41.343095		
		Above51	02:46:39.621326		
Creative Text-T2	Female	16To20	02:07:59.060739	08:55:34.507443	18:14:09.975684
		21To50	04:42:02.083260		
		Above51	02:05:33.363444		
	Male	16To20	01:40:00.847508	09:18:35.468241	
		21To50	05:01:56.195321		
		Above51	02:36:38.425412		
Date-D	Female	16To20	00:05:12.440700	00:21:01.608444	00:41:15.970948
		21To50	00:10:46.465144		
		Above51	00:05:02.702601		
	Male	16To20	00:03:20.604460	00:20:14.362504	
		21To50	00:11:08.602825		
		Above51	00:05:45.155219		
Sentence-S	Female	16To20	00:57:21.921357	04:03:36.311648	07:58:13.261270
		21To50	02:06:03.614898		
		Above51	01:00:10.775392		
	Male	16To20	00:38:17.698018	03:54:36.949622	
		21To50	02:14:46.483272		
		Above51	01:01:32.768331		

Table 23: Representation of Urdu Sentence Aligned Speech Data Duration

13.4 SUMMARY OF SPEAKERS

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

Age Group	Female	Male	Total
16To20	53	38	91
21To50	114	124	238
Above51	51	54	105
Total	218	216	434

Table 24: Distribution of Speakers of Urdu Sentence Aligned Speech Data

13.5 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, Mansoor Khan, Shahnawaz Alam & Bi Bi Maryam 2019. A Gold Standard Urdu Raw Text Corpus. Central Institute of Indian Languages, Mysore.
5. Ramamoorthy, L., Narayan Choudhary, Mansoor Khan, Shahnawaz Alam & Bi Bi maryam 2019. Urdu Raw Speech Corpus. Central Institute of Indian Languages, Mysore.

14 INDIAN ENGLISH - BENGALI VARIANT SPEECH ANNOTATION

Rejitha K. S., Rajesha N., Narayan Kumar Choudhary

14.1 OVERVIEW OF SENTENCE ALIGNED SPEECH CORPUS

Indian English – Bengali Variant Sentence Aligned Speech Corpus is created by annotating the speech data (Ramamoorthy L., et. Al, 2021) collected by LDC-IL. A detailed explanation of the Indian English – Bengali variant Speech Corpus is available in the Bengali Speech (Rejitha K.S.,. et. Al, 2021). LDC-IL Indian English – Bengali Variant Sentence Aligned Speech files contains an audio file and a textual layer in Roman 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 'EnglishBen_Male_16To20_Creative_Text-T2_SP-0002_T2-0003-001.wav'

The speech is annotated on the basis of English spelling convention. The words are labelled manually to the corresponding wave. LDC-IL Sentence Aligned Speech corpus contains four content types such as contemporary text, creative text, sentences and date format. The contemporary text and creative text are recordings of news and essays/novels. Each speaker has uttered typically 25 sentences randomly selected from phonetically balanced sentences list of LDC-IL speech data 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 orthographically normalised annotation.

14.2 OBSERVATIONS

LDC-IL sentence level speech annotation strictly follows the Standard English spellings. Deviations in the phonological features are evident in Indian English because English is an inter-language in India. Indians have their own mother tongue and those language features are drastically influencing their English pronunciation. Therefore, speakers from different regions speak the same word in different ways. Most of the Indian-English speakers do not know the Received Pronunciation properly. Stress, accent, intonation pattern of Received Pronunciation is difficult to follow for a common Indian speaker.
The reading speed differs from reader to reader. Fast reading informants pose difficulty in annotation. Since news items contain sports news, it includes the informant reading all types of numbers. Speakers sometimes wrongly utter large digit numbers 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 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.

14.2.1 PHONETIC ALTERNATION IN ENGLISH SPEECH DATA

Read speech has disfluency like unwanted pauses, elongated syllables, word fragments, self-corrections, and repeated 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:
English is a non-phonetic language and there is no one-to-one correspondence between the letters and the sounds. Since Indian languages are almost phonetic, people have difficulty in the production of English sounds. Stress, intonation and pitch are essential factors to produce English words and sentences which are difficult for a foreign speaker. These factors vary word by word. Such specific language oriented characteristics make trouble in language learning.

- **Deviation in Vowels**

Bengali speakers equate the English vowels /a:, ʌ, ɜ: and ə/ equate with /a/. i.e., perspective /pəspektɪv/ is pronounced as /parspektɪb/. People use /o/ extensively for /ə/, /ou/ and other diphthongs because of the Bangla language Influence. For e.g.: ‘concern’ /kənˈsɜːn/ is /konˈsɜrn/, low /loʊ/ is /lo/. Bangla speakers typically do not make any distinction between the long and short vowels of English, that means there is no difference between the vowels /i/and /i:/, /u/ and /u:/ etc.

- **Deviation in Consonants**

English fricative sounds are challenging for Bengali speakers. So they pronounce these fricatives /f, v, θ, ð, s, z / in a different way. Commonly they substitute /f/ with /ph/, aspirated plosive which is present in the Bangla language. For e.g.: If /ɪf/ as /ɪph/, put / /pʊt// as /puth/, value /vælju:/ as /balju/, zoom /zu:m/ as /ju:m/, question / /kwɛstjən/ as / /koʃtjan/ etc. They speak the final position /r/ which is normally dropped in the Received Pronunciation.

- **Other Deviations**

There is a tendency of changing letters into other letters which is similar or familiar to the speaker. Moreover speakers try to articulate easy phonemes instead of correct phonemes. These deviations are consistent throughout the particular speaker’s data.
For e.g.: Bangladesh /bangladeʃ/ as /ba:ngladeʃ/ Term /tə:m/ as /ta:m/

It is observed that some words are pronounced as some other word which is frequently in use or assume that the word is something and uttered the assumed word which is not in the prompt sheet.
For e.g.: Regional as religional

Mispronunciation is a common phenomenon of English speaking Indians.

For e.g.: demarcate /diːmɑːkeɪt/ as /dimikraɪt/ delimiting /dɪlɪmɪtɪŋ/ as /deɪlɪmɪtɪŋ/ Interchanging words is also observed in the speech data. This and that are read as the and vice versa. Similarly he and she are interchanged frequently by Indian speakers. Intermediate pause is another feature observed in some speakers' data. i.e., simultaneously is pronounced as 'simultaneous' after a pause 'ly'

14.3 SUMMARY OF THE CORPUS

The total duration of Indian English – Bengali variant Sentence Aligned Speech Corpus is 09:21:08 (hh:mm:ss) comprising 5,676 audio segments from 52 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 are showing 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.

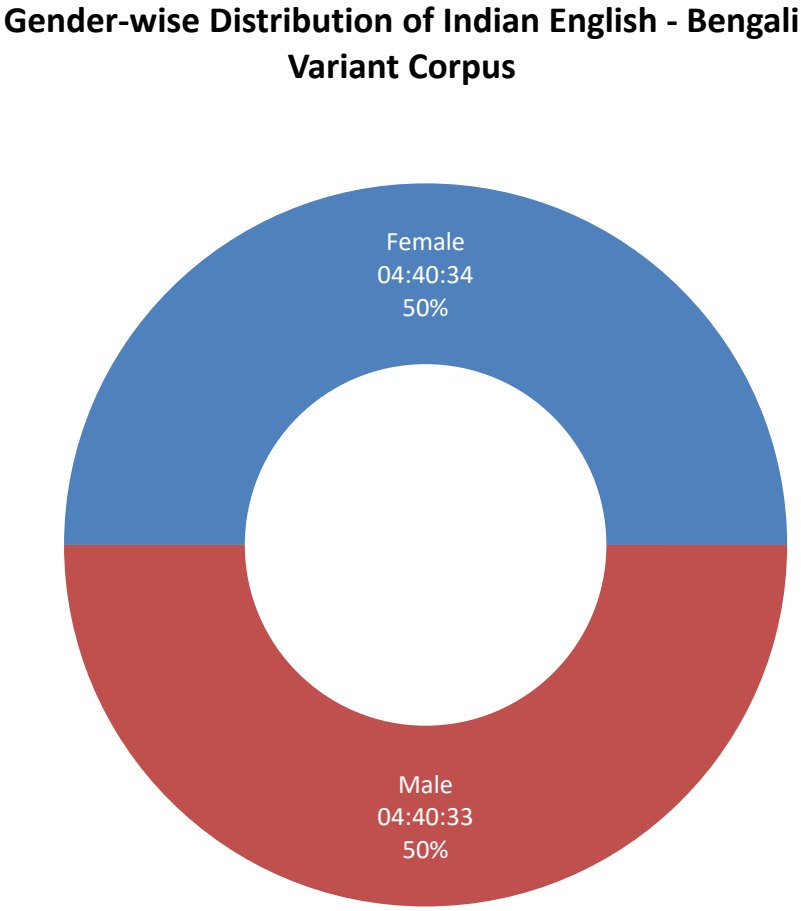

Figure 61: Gender-wise Distribution of Indian English - Bengali Variant Corpus

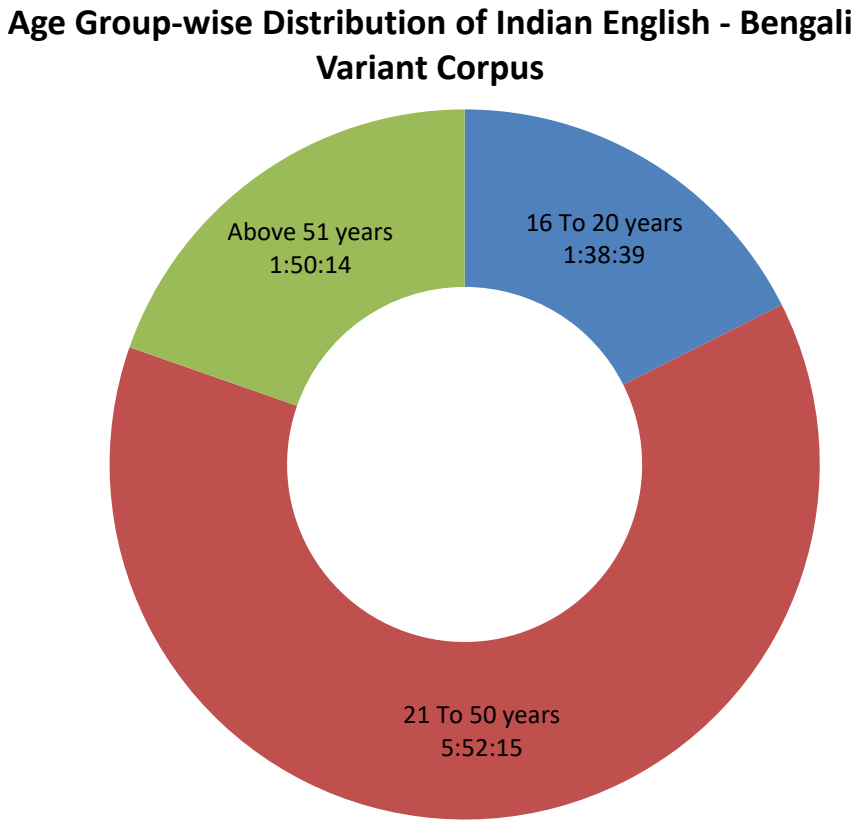

Figure 62: Age Group-wise Distribution of Indian English - Bengali Variant Corpus

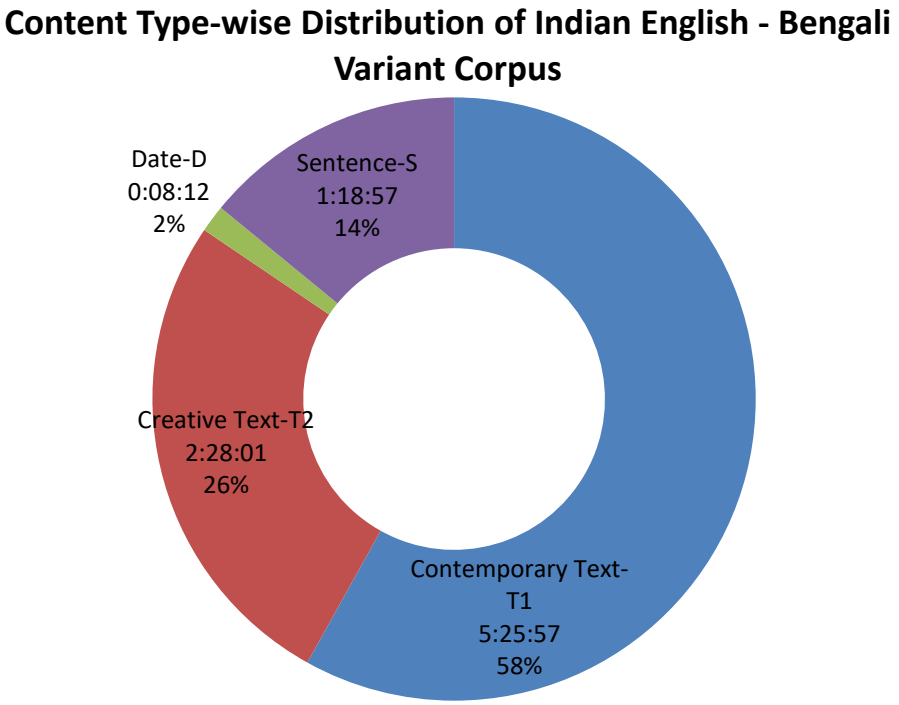

Figure 63: Content Type-wise Distribution of Indian English - Bengali Variant Corpus

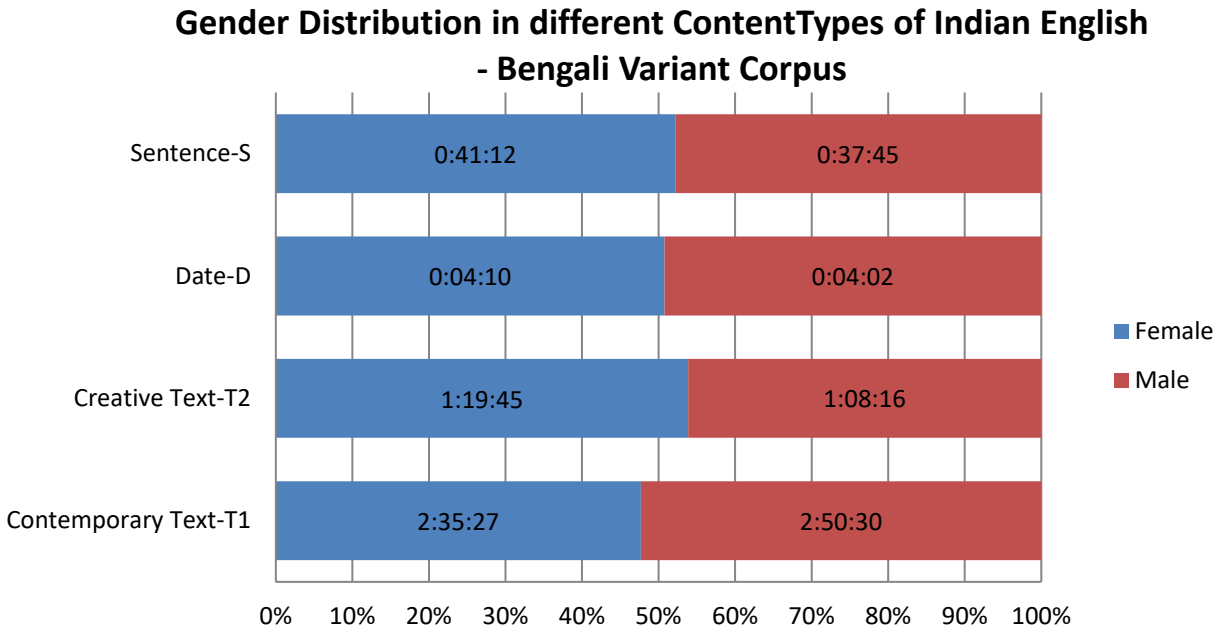

Figure 64: Gender Distribution in different Content Types of Indian English - Bengali Variant Corpus

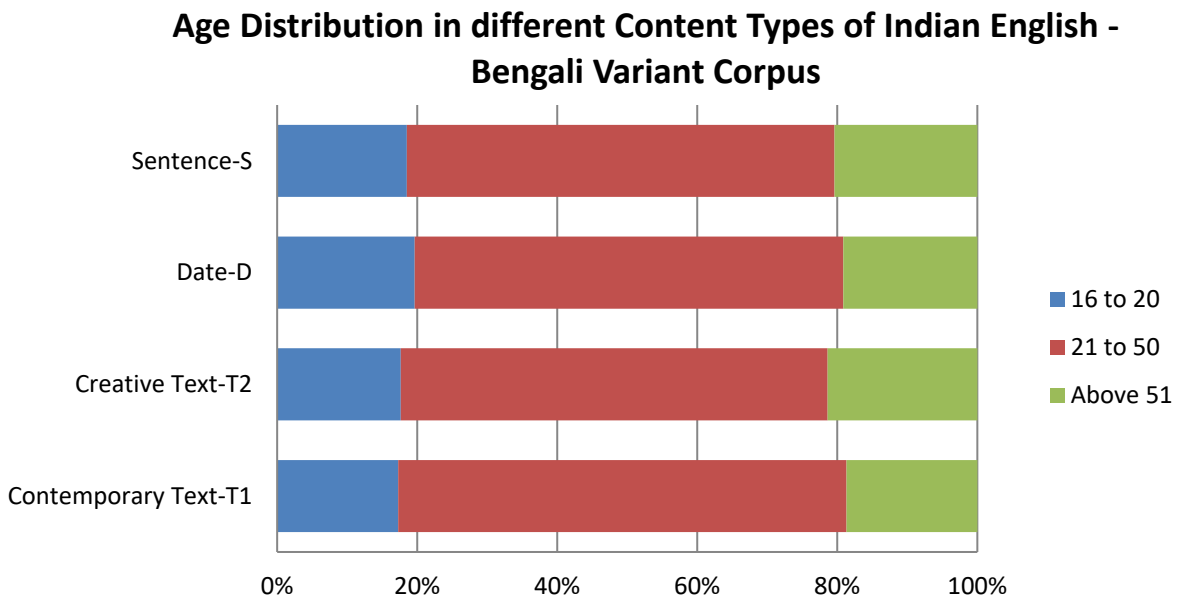

Figure 65: Age Distribution in different Content Types of Indian English - Bengali Variant Corpus

14.3.1 DURATION OF THE INDIAN ENGLISH – BENGALI VARIANT SENTENCE ALIGNED SPEECH DATA

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

Content Type	Gender	Age Group	Duration (hh:mm:ss.ms)		
Contemporary Text-T1	Female	16To20	00:26:04.885888	02:35:26.635196	05:25:56.824282
		21To50	01:40:20.070439		
		Above51	00:29:01.678869		
	Male	16To20	00:30:13.503441	02:50:30.189086	
		21To50	01:48:23.299945		
		Above51	00:31:53.385700		
Creative Text-T2	Female	16To20	00:13:32.162689	01:19:45.109971	02:28:01.137527
		21To50	00:49:17.548729		
		Above51	00:16:55.398553		
	Male	16To20	00:12:34.645914	01:08:16.027556	
		21To50	00:40:58.679443		
		Above51	00:14:42.702199		
Date-D	Female	16To20	00:00:43.320174	00:04:10.138457	00:08:12.428193
		21To50	00:02:39.073622		
		Above51	00:00:47.744662		
	Male	16To20	00:00:53.352199	00:04:02.289736	
		21To50	00:02:22.402920		
		Above51	00:00:46.534616		
Sentence-S	Female	16To20	00:07:54.189107	00:41:12.434842	01:18:57.165056
		21To50	00:25:00.628623		
		Above51	00:08:17.617112		
	Male	16To20	00:06:42.882659	00:37:44.730214	
		21To50	00:23:12.927829		
		Above51	00:07:48.919726		

Table 25: Representation of Indian English – Bengali Variant Sentence Aligned Speech Data Duration

14.4 SUMMARY OF SPEAKERS

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

Age Group	Female	Male	Total
16To20	5	5	10
21To50	16	16	32
Above51	5	5	10
Total	26	26	52

Table 26: Distribution of Speakers of Indian English – Bengali Variant Sentence Aligned Speech Data

14.5 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-ordinationand 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, Issue1.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 IndianLanguages,Mysore.pp.160-174.
4. Ramamoorthy L., Narayan Kumar Choudhary, Arundhati Sengupta, Rejitha KS, Rajesha N., Manasa,G.. 2021 Indian English Raw Speech Corpus - Bengali Variant Central Institute of Indian Languages,Mysore.978-81-948885-1-2.
5. Rejitha K.S., Rajesha N., Manasa G., Narayan Choudhary. 2021. "Indian English Raw Speech Corpus -Bengali Variant"in Compendium of Linguistic Resources in Indian Languages, Central Institute ofIndianLanguages,Mysore.pp.50-57.

15 INDIAN ENGLISH - KANNADA VARIANT SPEECH ANNOTATION

Rejitha K. S., Rajesha N., Narayan Kumar Choudhary

15.1 OVERVIEW OF SENTENCE ALIGNED SPEECH CORPUS

Indian English – Kannada variant Sentence Aligned Speech Corpus is created by annotating the speech data (Ramamoorthy L., et. Al, 2021) collected by LDC-IL. A detailed explanation of the Indian English – Kannada variant Speech Corpus is available in the Indian English -Kannada Speech Data Documentation (Rejitha K.S., et. Al, 2021). LDC-IL Indian English – Kannada Variant Sentence Aligned Speech files contains an audio file and a textual layer in Roman 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 'EnglishKan_Female_Above51_Creative_Text-T2_SP-0046_T2-0002-038.wav'

The speech is annotated on the basis of English spelling convention. The words are labelled manually to the corresponding wave. LDC-IL Sentence Aligned Speech corpus contains four content types such as contemporary text, creative text, sentences and date format. The contemporary text and creative text are recordings of news and essays/novels. Each speaker has uttered typically 25 sentences randomly selected from phonetically balanced sentences list of LDC-IL speech data 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 orthographically normalised annotation.

15.2 OBSERVATIONS

LDC-IL sentence level speech annotation strictly follows the Standard English spellings. Deviations in the phonological features are evident in Indian English because English is an inter-language in India. Indians have their own mother tongue and those language features are drastically influencing their English pronunciation. Therefore, speakers from different regions speak the same word in different ways. Most of the Indian-English speakers do not know the Received Pronunciation properly. Stress, accent, intonation pattern of Received Pronunciation is difficult to follow for a common Indian speaker. (Rejitha K.S., N. Rajesha, 2020).

The reading speed differs from reader to reader. Fast reading informants pose difficulty in annotation. Since news items contain sports news, it includes the informant reading all types of numbers. Speakers sometimes wrongly utter large digit numbers 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 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.

15.2.1 PHONETIC ALTERNATION IN ENGLISH SPEECH DATA

Read speech has disfluency like unwanted pauses, elongated syllables, word fragments, self-corrections, and repeated 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:
English is a non-phonetic language and there is no one-to-one correspondence between the letters and the sounds. Since Indian languages are almost phonetic, people have difficulty in the production of English sounds. Stress, intonation and pitch are essential factors to produce English words and sentences which are difficult for a foreign speaker. These factors vary word by word. Such specific language oriented characteristics make trouble in language learning.

- **Deviational Vowels**

Kannada speakers equate the English vowels /a, ʌ, ɜ: and ə/ equate with / ə or a /. South Karnataka people use /a/ whereas north Karnataka people use /ə/. /ɪ/ become /e/ and /ɔ/ become /ʌ/ in most of the observations. It is a common phenomenon among Kannada speakers that the English short vowels tend to be long vowels and diphthong becoming monophthong and vice versa. Pillar /pɪlə/ as /pailar/. The mother tongue influence introduces vowels at the end of words like school /sku:l/ become /sku:lu/. This is a common phenomena even in loan words. For e.g.: [bus] as [basu], [car] as [caru]

- **Deviation in Consonants**

Several Kannada speakers have difficulty to articulate specifically /s/ and /ʃ/, /f/ and /ph/, /w/ and /v/ etc. Some of the Kannada speakers pronounce words in different way such as so /so/ as /ʃo/ and /bats /bæts/ as /bætz/. Alveolar fricative /s/ is spread across post-alveolar, retroflex and palatal positions. Kannada speakers are not properly distinguishing /s/, /z/, /ʒ/ and /ʤ/. Non-standard English speakers in Karnataka pronounce/morning /ˈmɔ:nɪŋ/ as/ˈmo:rnɪŋ/ and learn /lɜ:n/ as /lɜ:rn/. Most of the Kannada Speakers are rhotic. They utter 'r' in middle and final positions even when it is silent.
Words ending with l, m, n are pronounced as əl, əm, ən. It is noticed in southern Karnataka, the ‘h’ insertion or deletion in Kannada words' pronunciation in large masses. Gemination is common in Dravidian languages so it is a habit to utter the double consonants. For e.g. summer /sʌmə(r)/ as /səmmər/ and cindrella /sɪndərelə/ as /sɪndərellə/ or /sɪndərella/

- **Other Observation**

When there are unfamiliar words in the prompt sheet the speakers try to utter according to the spelling. For eg: ‘aisle’ /aɪl/ is pronounced in many ways such as /aɪl/, /asəɪl/, /əɪsel/, /asel/; dining room /daɪnɪŋ ru:m/ as /dɪnɪŋ ru:m/. The speaker usually interchanges the articles ‘a, an, the’, pronouns ‘he and she’, article ‘the’ with the pronoun ‘that’. Moreover they miss or alter the

prepositions. The speakers assume the forthcoming words and utter words which might not be in the prompt sheet. It is extensively observed that the plural marker ‘s’ is omitted or added where it is not needed. Indian English speakers' Suprasegmental features like intonation, stress and tone are different from Received Pronunciation.

15.3 SUMMARY OF THE CORPUS

The total duration of Indian English – Kannada variant Sentence Aligned Speech Corpus is 11:17:40 (hh:mm:ss) comprising 6,166 audio segments from 53 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 are showing 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.

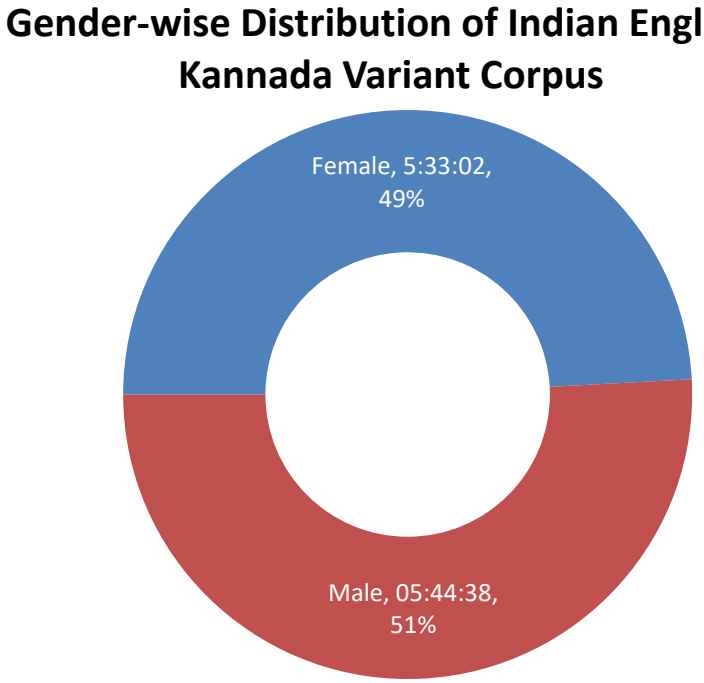

Figure 66: Gender-wise Distribution of Indian English - Kannada Variant Corpus

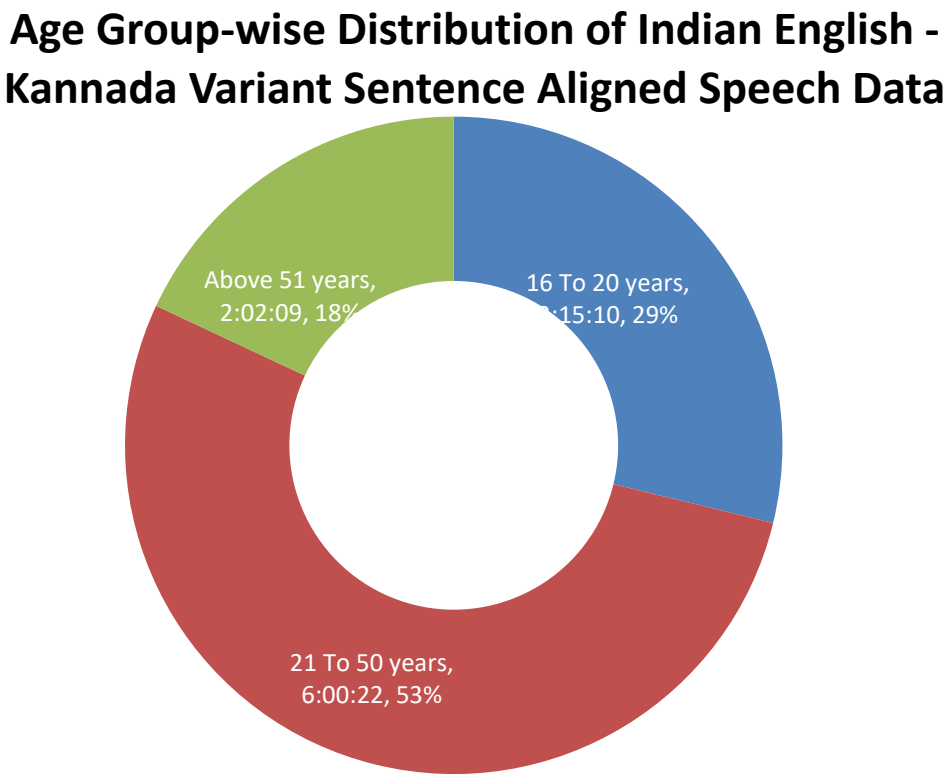

Figure 67: Age-wise Distribution of Indian English - Kannada Variant Corpus

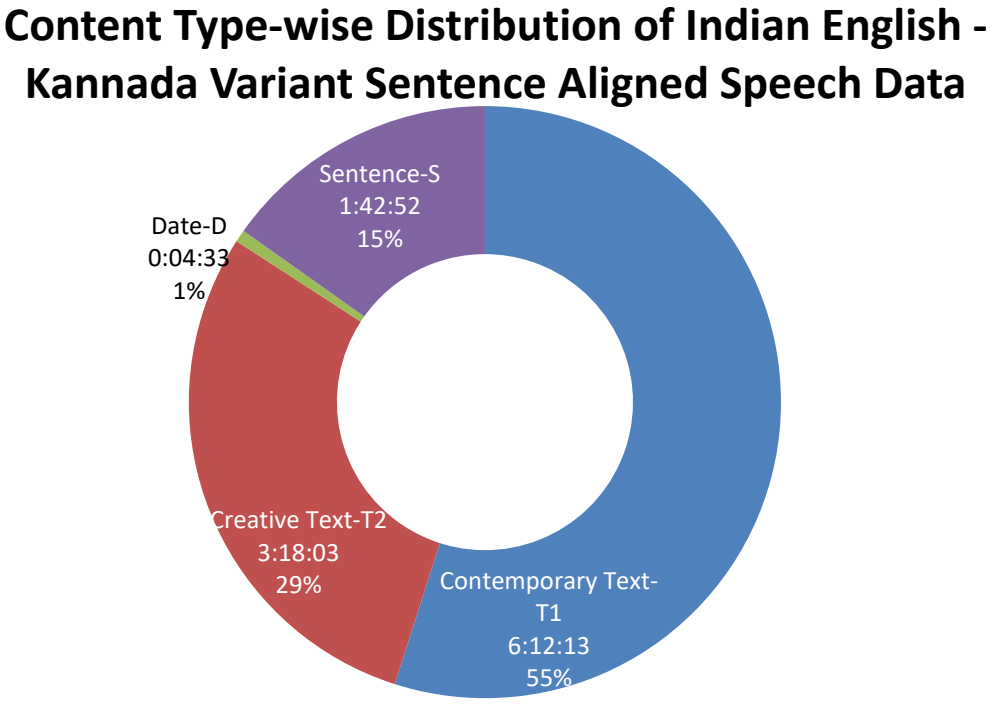

Figure 68: Age Group-wise Distribution of Indian English - Kannada Variant Corpus

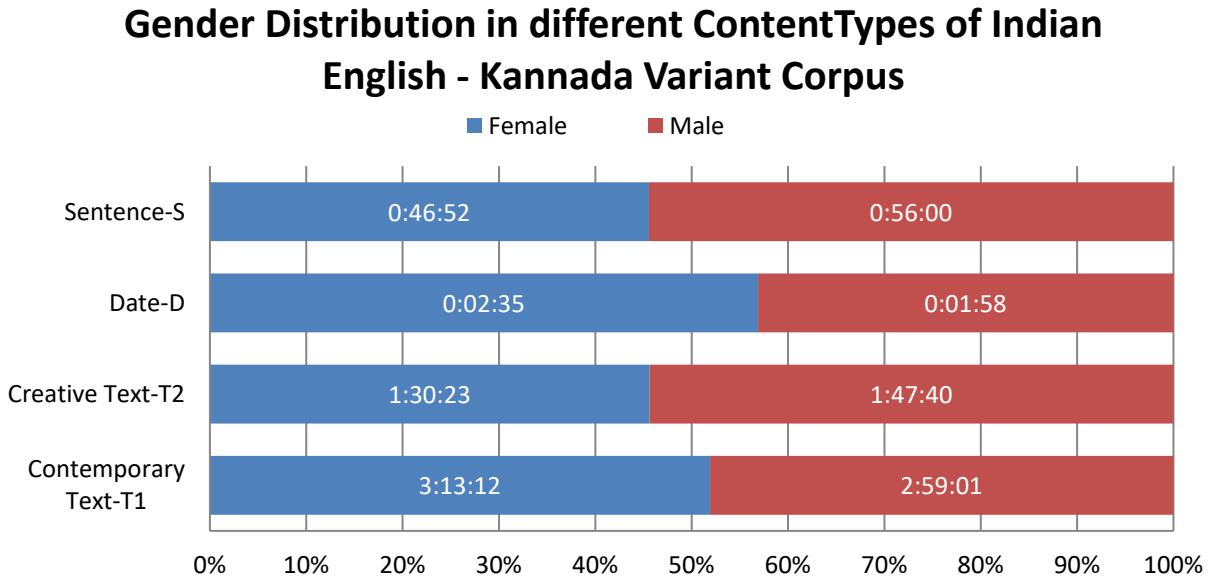

Figure 69: Gender Distribution in different Content Types of Indian English - Kannada Variant Corpus

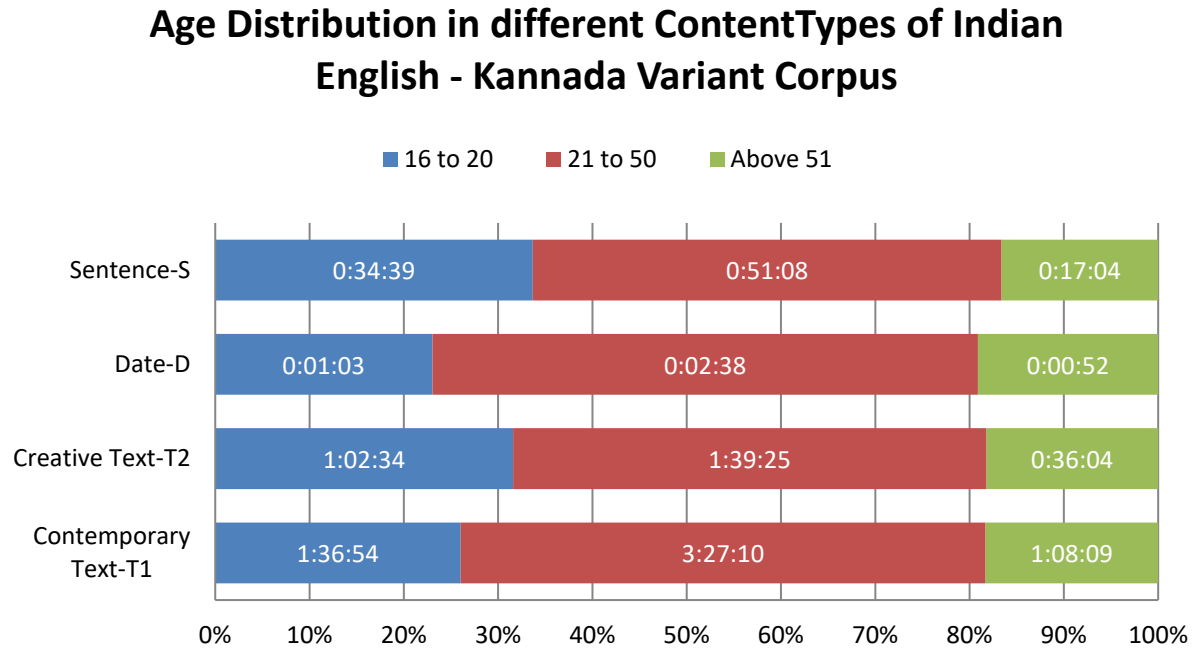

Figure 70: Age Distribution in different Content Types of Indian English - Kannada Variant Corpus

15.3.1 DURATION OF THE INDIAN ENGLISH - KANNADA VARIANT SENTENCE ALIGNED SPEECH DATA

The table below shows the duration of each of the content types and their distribution across a few factors in Indian English – Kannada Variant Sentence Aligned Speech Data.

Content Type	Gender	Age Group	Duration(hh:mm:ss.ms)		
Contemporary Text-T1	Female	16To20	01:02:28.741077	03:13:11.782010	06:12:13.027237
		21To50	01:40:24.956010		
		Above51	00:30:18.084923		
	Male	16To20	00:34:25.205666	02:59:01.245227	
		21To50	01:46:45.262068		
		Above51	00:37:50.777493		
CreativeText-T2	Female	16To20	00:23:16.523228	01:30:23.163705	03:18:02.845505
		21To50	00:50:25.904379		
		Above51	00:16:40.736099		
	Male	16To20	00:39:17.552623	01:47:39.681799	
		21To50	00:48:59.361319		
		Above51	00:19:22.767856		
Date-D	Female	16To20	00:00:36.554232	00:02:35.074273	00:04:32.600275
		21To50	00:01:27.910021		
		Above51	00:00:30.610020		
	Male	16To20	00:00:26.264155	00:01:57.526002	
		21To50	00:01:09.751868		
		Above51	00:00:21.509979		
Sentence-S	Female	16To20	00:11:32.067080	00:46:52.231114	01:42:51.771055
		21To50	00:26:25.565244		
		Above51	00:08:54.598789		
	Male	16To20	00:23:06.898949	00:55:59.539942	
		21To50	00:24:42.836948		
		Above51	00:08:09.804044		

Table 27: Representation of Indian English – Kannada Variant Sentence Aligned Speech Data Duration

15.4 SUMMARY OF SPEAKERS

The table below shows the total number of speakers and their distribution in the Indian English – Kannada Variant Sentence Aligned Speech Data.

Age Group	Female	Male	Total
16To20	7	6	13
21To50	14	16	30
Above51	5	5	10
Total	26	27	53

Table 28: Distribution of Speakers of Indian English – Kannada Variant Sentence Aligned Speech Data

15.5 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. LanguageResources&Evaluation.Springer,Vol.55,Issue1.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. K.S. Rejitha & N.Rajesha, “A Corpus Analysis of English Pronunciation Deviations among Kannada Speakers” in Working Papers on Linguistics and Literature, Department Of Linguistics, Bharathiyar University, Coimbatore. Volume: XIV 2020 .pp.367-373. ISSN 2349-8420.
5. Rejitha K.S., Rajesha N., Manasa G., Narayan Choudhary. 2021. "Indian English Raw Speech Corpus -Kannada Variant" in Compendium of Linguistic Resources in Indian Languages, Central Institute of Indian Languages, Mysore.pp.58-65.
6. Ramamoorthy L., Narayan Kumar Choudhary, Bharatha Raju A., Rejitha KS, Rajesha N., Manasa, G.2021 Indian English Raw Speech Corpus - Kannada Variant Central Institute of Indian Languages, Mysore.978-81-948885-9-8.

16 CHHATTISGARHI RAW SPEECH CORPUS

Satyaendra Kumar Awasthi, Ankita Tiwari, Narayan Kumar Choudhary

16.1 INTRODUCTION

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

As the mother tongue of an estimated 17 million people, Chhattisgarhi is of great cultural and historical significance to the state of Chhattisgarh. It also serves as a medium of expression for the people of Chhattisgarh, reflecting their rich traditions, folklore, and way of life. According to the Indian government, Chhattisgarhi is an eastern dialect of Hindi and is mostly spoken in the Indian state of Chhattisgarh. As of the 2011 Census, 1,62,45,190 people were estimated to speak Chhattisgarhi. Since 2007, Chhattisgarh has recognised it as one of its official languages.

Central India's Chhattisgarh state was established on November 1st, 2000, following its split from Madhya Pradesh. The Chhattisgarhi language's traditional name is Kosali or Dakshin Kosali, which has historical importance because the Chhattisgarh area was formerly known as Dakshin Kosal. The language has a lengthy and complex history, and it has evolved and changed through time in a number of ways. It has continued to play a significant role in the Chhattisgarh region's cultural and linguistic environment despite these developments.

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. The Chhattisgarhi raw speech corpus is a wide-ranging dataset that represents the vibrant, dynamic linguistic legacy of Chhattisgarhi.

16.2 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 multifarious needs of the research and development community. To ensure that the corpus is good for an ASR, the continuous speech is recorded in a natural environment. Two different content types of speech are collected while building this corpus.

The Chhattisgarhi raw speech corpus is made up of recordings of native Chhattisgarhi speakers from various parts of the state of Chhattisgarh, and it represents a wide range of Chhattisgarhi varieties as they are spoken in various locations by diverse speakers. There are several audio for each content category, and each audio has different material. Each speaker from various

agegroups recites prompt text extracts of literary and news texts. A minimum of 2000 words of read speech is recorded for each speaker. In terms of data size and time length, read speech makes up the majority of the speech corpora. Below are comprehensive explanations of each of the content types:

16.2.1 CREATIVE TEXT

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

16.2.2 NEWS TEXT

To maintain an accurate representation of formal spoken Chhattisgarhi from native speakers, news text extracts were recorded. This content type includes recordings of Chhattisgarhi articles, editorials, and news that have been published from Chhattisgarhi newspapers, monthly magazines, and some online sources.

16.2.3 GENDER AND AGE BALANCE

Three age groups have been chosen. The categories are "16 to 20 years", "21 to 50 years" and "Above 50 years". An effort has been made to maintain the corpus is balanced in terms of age and gender.

16.2.4 COLLECTION OF DATA

Speech data comprising of 140 speakers was collected through fieldwork. The fieldwork covered the areas from Chhattisgarh (Central), during January 15 - 21, 2023, by six investigators namely *Satyaendra Kumar Awasthi, Srishti Singh, Saurabh Varik, Shantanu Kumar Ankita Tiwari* and *Rupesh Pandey*. The Investigators visited several colleges, universities and organisations in order to gather the data. There are 140 speakers in the Chhattisgarhi speech data, The locations covered by the different speakers are mentioned are, Balod, Baloda Bazar, Bemetara, Bilaspur, Dhamtari, Durg, Gariaband, Janjgir-Champa, Kabirdham, Khairagarh-Chhuikhadan-Gandai, Korba, Mungeli, Mahasamund, Raipur, Sakti, Rajnandgaon, Sarangarh-Bilaigarh.

Informants are made aware of why exactly the data is being collected. The informants are also made aware of the potential benefits of the data to the wider community. Once the informant is aware of all this information and is ready to give the data, consent is acquired in writing along with certain personal details which makes a part of metadata.

Every effort is made to reduce any additional noise as much as possible. Before the actual recording of the text, test recordings are conducted. Informants are allowed to read the dataset earlier before recording so that they can get familiar with the content of the text. While audio

recording the read-speech content types the informants were told to read the Text as natural as reading a book or newspaper possible

16.2.5 TECHNICAL SPECIFICATIONS FOR COLLECTING DATA

The LDC-IL data is recorded using Roland EDIROL Recorder. It is a 24-bit Linear PCM (R-09) Recorder. Recording was done at the Sample Rate: 48.0 KH in 16bit wav format using rechargeable batteries.

16.2.6 GUIDELINES FOLLOWED WHILE AUDIO RECORDING

- Minimum 5 cm to 25 cm distance between the microphone and the speaker
- Recorder should not be placed orthogonally but it should be placed diagonally.
- Do not move the recorder during recording. Fix the recorder upon a fixed plane if possible.
- Maintaining fixed distance between the recorder and speaker

After each recording, it is suggested to the investigator to ensure that the recorded data is received properly. The investigator may oblige the informant's request if they also want to hear the data.

16.3 METADATA

The value of speech data can be determined according to the quality of metadata obtained. It is imperative to maintain metadata of the entire data collection for linguistic analysis. A brief of each of these 21 fields is given in the table below:

Sl.	Legend	Description
1.	Language	Name of the Language
2.	Speaker ID	Each speaker has a unique ID. However, this is within the language.
3.	Gender	Note gender, whether it is male, female or other.
4.	Age Group	Three age groups of 16 to 20, 21 to 50, and 50+
5.	Mother Tongue	Mother tongue of the informant.
6.	Dialect	An attempt has been made to cover all the dialects of the language as agreed upon in the academia of the language experts and linguists.
7.	State	Name of the Indian state/province of the speaker
8.	District	Name of the district of Indian state/province of the speaker
9.	Education	Highest educational qualification of the speaker.
10.	Place: Elementary Education	Place of elementary education. This usually corresponds to the early childhood experiences which happen to more than often affect the way a language spoken.
11.	Place	Place provides the information that where the speech data collected.
12.	Date	Date when the recording took place.
13.	Environment	Indoor/Outdoor environment in which the recording has been done.
14.	Content Type	This corresponds to the notation of the content types noted above.
15.	Content ID	This corresponds to the ID of the text being read out.
16.	Audio File Name	The name of every Audio file is distinct based on speaker and content types
17.	Recorded Text	Text (Datasets) of the recorded speech.
18.	Duration	This is related to the duration of audio in milliseconds.
19.	User By	Person who handled the data-metadata association in corpus.
20.	Script	The script in which the content has been provided to the informant for reading.
21.	Investigator	Name of the Investigator.

Table 29: Metadata fields and their description

16.4 TEXT-SPEECH MAPPING AND NAMING CONVENTIONS

After the completion of field work, the speech segmentation and mapping audio with corresponding text and metadata is done. Each recording is named in accordance with its metadata information like language name, gender, age group, speaker id and content id etc. A Typical LDC-IL naming convention for Chhattisgarhi Speech Data is shown below.

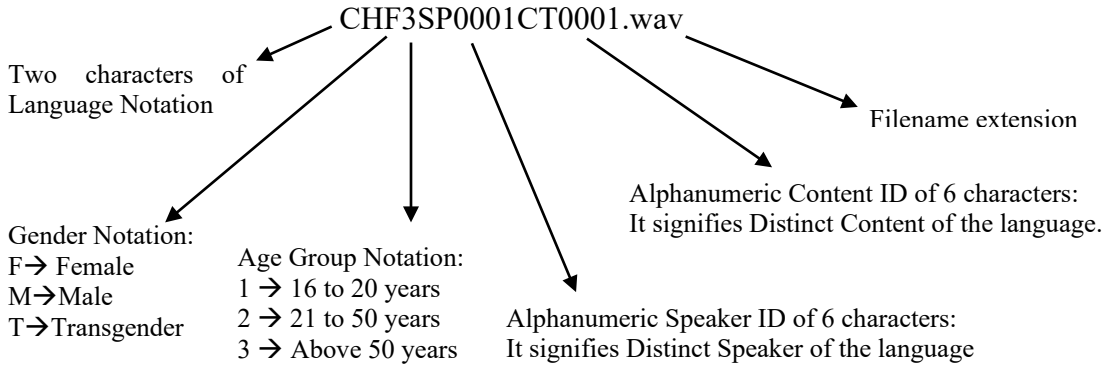

Figure 71: LDC-IL Naming Convention of Chhattisgarhi Speech Data

16.5 SUMMARY OF THE CORPUS

The Speech data has 140 distinct speakers with the duration of 77:58:16.389 (hh:mm:ss.ms)
The distribution of duration across each content type is as follows

ContentType	Gender	AgeGroup	Duration (hh:mm:ss.ms)		
Creative Text	Female	16 to 20	00:39:35.528	21:45:55.196	49:00:23.142
		21 to 50	19:26:08.905		
		Above 50	01:40:10.763		
	Male	16 to 20	00:14:59.795	27:14:27.946	
		21 to 50	22:31:55.540		
		Above 50	00:00:16.970		
News Text	Female	16 to 20	00:43:36.620	10:05:56.501	28:57:53.247
		21 to 50	08:31:57.691		
		Above 50	00:50:22.190		
	Male	16 to 20	01:25:36.546	18:51:56.746	
		21 to 50	14:42:09.963		
		Above 50	02:44:10.237		

Table 30: Distribution of duration across each content type of Chhattisgarhi

The distribution of number of speakers across content type and region is as follows.
Among 140 distinct speakers Creative Text content type is read by 129 speakers and its distribution is as follows.

Gender	Age Group	Speakers	Total Speakers
Female	16 To 20	3	54
	21 to 50	47	
	Above 50	4	
Male	16 To 20	2	75
	21 to 50	62	
	Above 50	11	

Table 31: Distribution of number of speakers across Chhattisgarhi Creative Text

Among 140 distinct speakers News Text content type is read by 94 speakers and its distribution is as follows.

Gender	Age Group	Speakers	Total Speakers
Female	16 To 20	3	34
	21 to 50	28	
	Above 50	3	
Male	16 To 20	3	60
	21 to 50	48	
	Above 50	9	

Table 32: Distribution of number of speakers across Chhattisgarhi News

16.6 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 Adil, Satyabhama 2003. Chhattisgarhi Bhasha aur Sahitya, Vikalp Prakashan Raipur pp. 2-31

5 Census 2011. https://censusindia.gov.in/

17 CHHATTISGARHI RAW TEXT CORPUS

Ankita Tiwari, Satyaendra Kumar Awasthi, Narayan Kumar Choudhary

17.1 RAW TEXT CORPUS: AN OVERVIEW

A corpus is a vast collection of appropriate representation of expressions in a language, either in written or spoken form. The electronic text corpus is a collection of texts from various languages that have been chosen for linguistic research as a source of data. Textual materials best reflect a language or linguistic variant according to external standards. Text Corpus is one of the key resources for language technology.

A corpus serves as a fundamental resource for numerous modules within a variety of applications, such as grammar checkers and spell checkers commonly used in word processing software. Indian languages often present formidable challenges in the realm of Natural Language Processing (NLP) and Artificial Intelligence (AI). The field of language technology frequently encounters obstacles due to the scarcity of language resources. To foster the development of language technology for mass-application tools, there is a consistent demand for the long-term availability of substantial linguistic data. However, it is imperative that this data is collected, organised, and stored in a manner that accommodates the diverse needs of tech developers.

Over the years, many efforts have been made to generate text corpus in Indian languages and various organisations, including government agencies, academic institutions, and private bodies, have made their contributions. The text corpus generated under the LDC-IL ambit is intended to bridge the gap by offering more and more electronic data for the NLP/AI and language technology community. Therefore, promoting technology and increasing its availability could mark a turning point for wider use of Indian languages through technology.

17.2 INTRODUCTION TO CHHATTISGARHI

LDC-IL extended its efforts to promote non-scheduled languages apart from the 22 scheduled languages, commencing with the collection of written material in Chhattisgarhi which has the ISO639-3 Notation of hne 639-3[1]. The prime goal of LDC-IL is to gather the written material to generate a quality text corpus to emphasise the significance of mother tongues.

A corpus effectively showcases the linguistic intricacies and distinctive features of a language when it achieves two critical criteria: substantial size and a genuine representation of diverse domains. Text corpus of a language is an asset to scientifically explore the language and locate reliable evidence for features of the language, thus helping us comprehend its explicit attributes.
Chhattisgarhi, a tongue of approximately 17 million people, carries profound cultural and historical significance within the region of Chhattisgarh. It goes beyond being merely a means of communication; Chhattisgarhi serves as a powerful conduit for the people of this region to convey their time-honoured traditions and captivating stories. Recognized by the Indian

[1] https://iso639-3.sil.org/code/hne

government as an eastern variant of Hindi, Chhattisgarhi primarily finds its voice in the state of Chhattisgarh. According to the 2011 Census, an estimated 16.2 million individuals were identified as Chhattisgarhi speakers.[2]

Since the year 2007, Government of Chhattisgarh has bestowed official status upon Chhattisgarhi, acknowledging its importance in the state's cultural tapestry. Additionally, Chhattisgarhi shares remarkable linguistic and cultural affinities with other languages such as Bagheli, Bhojpuri, Magahi, Jharkhandi (Nagpuriya and/or Sadari Korwa), Odia, Dravidian, Marathi, Bundeli, and Bhojpuri. This linguistic diversity reflects the rich mosaic of cultures and traditions that thrive in this vibrant corner of India. (Pathak & Verma 2018)

The Chhattisgarhi Raw Text Corpus endows an unrivalled window in documenting the colloquialisms, idioms, regional vocabularies, and grammar that are essential to establishing frameworks for linguistic processing. The Chhattisgarhi Raw Text Corpus is an extensive repository encapsulating the viable linguistic elements of Chhattisgarhi textual materials.

17.3 CHHATTISGARHI RAW TEXT CORPUS CATEGORIZATION

The LDC-IL text corpus illustrates language use as it manifests in day-to-day life rather than providing speculative or idealised instances. In order to fulfill these criteria, it is necessary to collect an enormous volume of data from various domains. Otherwise, frequently occurring contents might come from a certain lexicon or style. A large volume of data provides somewhat accurate results of what occurs frequently and what occurs rarely in a language.

In terms of form, function, content, and features, each corpus text source differs from the other. The Chhattisgarhi Raw Text Corpus captures the data from the Aesthetic and Official Document domains. The texts are extracted from various literary works that capture a vast range of literary terms. The textual materials are collected from the standard textbooks that are descriptive in nature. These textbooks demonstrate the style of Chhattisgarhi of a particular time period. Apart from standard textbooks, we also collected texts from creative writings to generate the Chhattisgarhi Raw Text Corpus. It comprises novels, plays, short stories, children's story books, mythology, culture, essays, biographies and folk tales etc.

#	Category	SubCategory	Word Count	Character Count
1	Aesthetics	Biographies	37586	187668
2	Aesthetics	Culture	92413	437741
3	Aesthetics	Folk Tales	26475	121735
4	Aesthetics	Literary Texts	32036	160133
5	Aesthetics	Literature-Children's Literature	7837	36380
6	Aesthetics	Literature-Novels	256246	1210840
7	Aesthetics	Literature-Plays	53123	255168
8	Aesthetics	Literature-Short Stories	248097	1158677
9	Aesthetics	Mythology	142408	699003
10	Official Document	Administration	4689	25343

Table 33: Domains and their sub-categorisation of Chhattisgarhi

[2] https://censusindia.gov.in/nada/index.php/catalog/42458

17.4 LDC-IL TEXT CORPUS METADATA

The metadata information is the integral part of the entire data for linguistic analysis. The collected data is structured in accordance with its information called metadata that involves the data's category, sub-category, text title, author, source, publisher, year of publication, page numbers, etc. This information is quite helpful for users to access data in the database or repository. Metadata provides the detailed information; how the data was initially created and what it contains. The following table shows a brief description of LDC-IL metadata information.

#	Legend	Description
1	Filename	Represented by the "docID" tag in the XML files. This is a unique file number across the datasets.
2	Project Description	The LDC-IL team collected texts from the field in the form of books, newspapers, magazines, and journals for the Chhattisgarhi data by visiting universities, colleges, publishing companies, and individual authors.
3	Sampling Description	The corpus may employ this information for verifiable proof. It will provide details about related book page numbers selected for the corpus. (In the case of Chhattisgarhi, due to the limited availability of text resources, all the resources made available to us were selected in entirety to be a part of the text corpus.)
4	Category	Specifies the domain of the text.
5	Subcategory	Specifies the sub-domain of the text.
6	Text	Specifies the type of the source text i.e. whether its origin is a book, a magazine or a newspaper.
7	Title	Specifies the title of the source text. It contains mostly books but if magazines or newspapers occur, their respective articles are provided here.
8	Volume	Specifies volume number of the title, if any.
9	Issue	Specifies issue number of the title, if any.
10	Text Type	Is mostly blank however sometimes it is used to provide the broad topic of the news items e.g. whether it is a political news or editorial or sports news etc.
11	Headline	This information is a verifiable proof for the corpus. This is normally the heading of the chapter of the selected sample. Gives the fine-tuned information of the topic present in particular file.
12	Author	Specifies the name of the author.
13	Editor	Specifies the name of the editor.
14	Translator	Specifies the name of the translator.
15	Words	Specifies the total number of words in the file.
16	Letters	Specifies the total number of UTF-8 characters in the file.
17	Publishing Place	Specifies the place where the title was published.
18	Publisher	Specifies the name of the publisher.
19	Published Year	Specifies the publication year.
20	Index	Is the index number or ID of the file. It is noted inside the XML file. It is mostly the same as the file name.
21	Date	Date when the file was digitised/inputted.
22	Input	Name of the Data Inputter, if the file has been typed.
23	Proof	Name of the Proof reader.
24	Language	Name of the language.
25	Script	Name of the script in which the text is written.

Table 34: Metadata Legends for LDC-IL Text Data[3]

[3] https://data.ldcil.org/upload/pubs/LDCIL_Release_Documentation.pdf

17.5 CHHATTISGARHI TEXT CORPUS NAMING CONVENTIONS

Collected Chhattisgarhi hardcopies in Devanagari script are digitised in to XML format, adhering to Unicode standards. Each corpus file has a unique filename, indexing it. These files are usually excerpts from specific book titles, named per language and source conventions. Naming convention uses notations and a five-digit number for distinct filenames.

For example, the text taken from the Chhattisgarhi book for LDC-IL Chhattisgarhi Text Corpus always starts with 'CH' followed by 5-digit numbers which is continuous, whereas text collected from Chhattisgarhi Magazine will start with 'CHM' followed by 5 digit numbers. If the source is from a Newspaper then 'CHN' notation will be followed where as if the News is taken from the Web source 'CHNW' will be used as a notation.

In certain cases, if the book is chaptered, the headline of each chapter changes, to capture the change of the topic. If the language experts wish to break the sampling of a book into different smaller files, then the filename will get attached with a roman small letter suffixed and enclosed in braces.

Such file names could be 'CH00001(a)', 'CH00002(b)', 'CH00001(c)', 'CH00001(d)' etc.

17.6 METHODOLOGY

This section outlines the sequence of activities, beginning with data collection, followed by data processing methodologies, culminating in the creation of the final corpus compilation.

17.6.1 DATA SAMPLING

For languages in which there is ample data, LDCIL selects a few chapters from each book as part of the corpus to maintain uniformity in terms of domains and variety. In the case of Chhattisgarhi, due to the limited availability of literary text resources, all the resources made available to us were collected in entirety to be a part of the text corpus, avoiding the process of data sampling.

17.6.2 DATA COLLECTION

A team of 6 resource persons namely, Dr Satyendra Awasthi, Ankita Tiwari, Shantanu Jha, Saurav Varik, Rupesh Pandey and Dr Srishti Singh was sent to Raipur and Bilaspur Districts of Chhattisgarh to collect data from the field in January 2023. The team visited the Universities, Colleges, Publication houses and individual authors, and collected texts in the form of books, Newspapers, Magazines and Journals from the field. The team also contacted some authors over e-mails and phone calls and managed to receive books from them through India Post.

17.7 DATA PROCESSING

This subsection describes each step involved in the processing of the collected data.

17.7.1 PROCESSING RAW DATA

The books underwent a scanning process where each page's image was assigned a distinct identifier. These images were subsequently uploaded onto the LDC-IL Data portal. The complete dataset was then categorised into multiple tasks based on character count, typically around 30,000 characters per task. These tasks were subsequently imported into the LDC-IL digitization platform, which possesses the capability to perform Optical Character Recognition (OCR) and extract the text content from each file.

17.7.2 SELECTION OF LANGUAGE EXPERTS AND DIGITIZATION PROCESS

The Chhattisgarhi text corpus is digitized and review is done by freelance language expert empanlled at CIIL by evaluating the performance of each freelancer for expertise on four levels; 1) spelling 2) grammar 3) punctuation and 4) spacing accuracy. The entire dataset was digitised by the empanelled language experts using the same CIIL digitization platform. Once the Digitization process was complete, each task went through a rigorous review process to check the authenticity of the work using the same platform, before finalising the data.

17.7.3 DATA WAREHOUSING AND CORPUS CREATION

Once the review process is complete, the reviewed data is exported from the digitisation platform and the entire data is obtained in CSV format. The data is then warehoused using the in-house tool called Kanaja, along with the metadata information such as Book ID, page number, author information etc.

17.7.4 COPYRIGHT CONSENT

Copyright consent was sought from the copyright holders of the Chhattisgarhi texts during the process of data collection. Some of the copyright holders gave their consent later on via email. This copyright consent allows the Central Institute of Indian Languages (CIIL) to use electronic excerpts from the source text or book, excluding images or diagrams, at no cost. CIIL can use these extracts for language technology development related to the Chhattisgarhi, citing the source when distributing the text for research or commercial purposes. CIIL cannot sell the extracted content separately without prior consent but may use it commercially as part of a larger corpus for language technology development. The copyright owner retains all rights; CIIL has the consent to use its electronic form of text.

17.8 OVERVIEW OF REPRESENTED DOMAINS

The size of LDC-IL Chhattisgarhi Raw Text Corpus is 8,97,450 words corresponding to 42,76,693 characters, collected from 35 books.

17.8.1 Aesthetics

The Aesthetics category of Chhattisgarhi text corpus covers 9 sub-categories bearing a total of 8,97,450 words. The representational details are given in the following chart

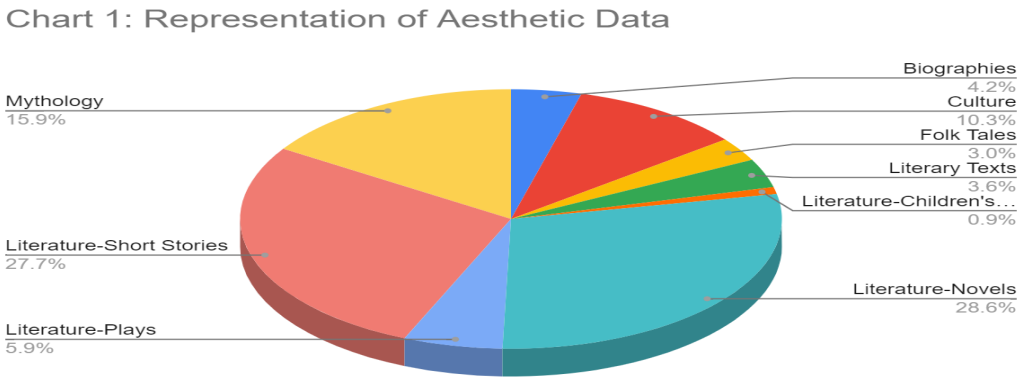

Figure 72: Representation of Aesthetic Data of Chhattisgarhi

17.9 REFERENCES

6 Narayan Choudhary. 2019. Linguistic Resources for AI/NLP in Indian Languages. Central Institute of Indian Languages, Mysuru. https://data.ldcil.org/upload/pubs/LDCIL_Release_Documentation.pdf

7 Ethnologue.(n.d.).hne|ISO-639-3. Retrieved 09 15, 2023, from SIL:

8 https://iso639-3.sil.org/code/hne

9 Vinay Kumar Pathak,& Vinod Kumar Verma. (2018). Chhattisgarhi ka Sampurna Vyakaran. Bilaspur: Vadanya Publication.

10 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