ಕನ್ನಡ ವಾಕ್ಯಕ್ರಮೀಕೃತ ವಾಗ್ದತ್ತಾಂಶ

Kannada - Sentence Aligned Speech Corpus

KANNADA SENTENCE ALIGNED SPEECH CORPUS

Annotated, quality language data (both-text & speech) and tools in Indian Languages to Individuals, Institutions and Industry for Research & Development - Created in-house, through outsourcing and acquisition.

Linguistic Data Consortium for Indian Languages
Central Institute of Indian Language
Mysuru, India-570006

CENTRAL INSTITUTE OF INDIAN LANGUAGES
Manasagangotri, Mysuru, Karnataka, India, 570006
www.ciil.org

Title: Kannada Sentence Aligned Speech Corpus
Authors: Vijayalaxmi F. Patil, Chetan Baji, Kavitha Lenin, Reshma S., Rajesha N., Manasa G.,Srikanth D., Stephen Fernandes, Nithin S., Narayan Kumar Choudhary, Shailendra Mohan

ISBN: 978-81-19411-19-1
CIIL Publication No.: 1423

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

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 KANNADA SPEECH ANNOTATION

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

2.1 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. The corpus consists of an audio file for each recording and corresponding textual layer consisting of the phonetically normalised and the orthographically normalised annotation.

2.2 OBSERVATIONS

LDC-IL sentence level speech annotation strictly follows what the speaker pronounces rather than what is in the prompt sheet. The text has been written in the respective language script and the speech is transcribed as much as the script supports. Two or more different pronunciations can be uttered by the same or different speaker for the same word. Even if it is read 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.

2.2.1 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:kṣisuttade* *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]** *ɛa: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[ɛra]* 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.: *ɛastraʧikitsejinda:gijo:* > *ɛastraɛriʧikitsejinda: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* *viṣṇuvard^{h}an* > *viṣṇuvard^{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 >**jo**ːɭu;* ***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; viʤajoːtsava > viʤajoːstava; 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; heʧʃaːgi > eʧʃaːgi; 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:

bʱajoːtpaːdaka > bajotpaːdaka; *ʧʰaːje > ʧaːje;* *bʱaja > baja;*
samartʰaniːja> samartaniːja; *ullaŋgʱane > ullaŋgane* *gʱoːʂisidare >goːʂisidare*

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.

praveːɕisidaːgaluː > praveːsisidaːgaluː *ɕikʂejinda > sikʂejinda* *prasaːda > praɕaːda*
naḍesuttaːraː > naḍeɕuttara *praɕnege > prasnege* *ɕaːntiːsaːntiː*
maːnasikavaːgi > maːnaɕikavaːgi *ɕlaːgʱisi > slaːgisiː* *niʂpakʂa > nispakʂa*
anulakʂisi> anulakʂiɕi

2.3 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 65,533 audio segments from 600 speakers.

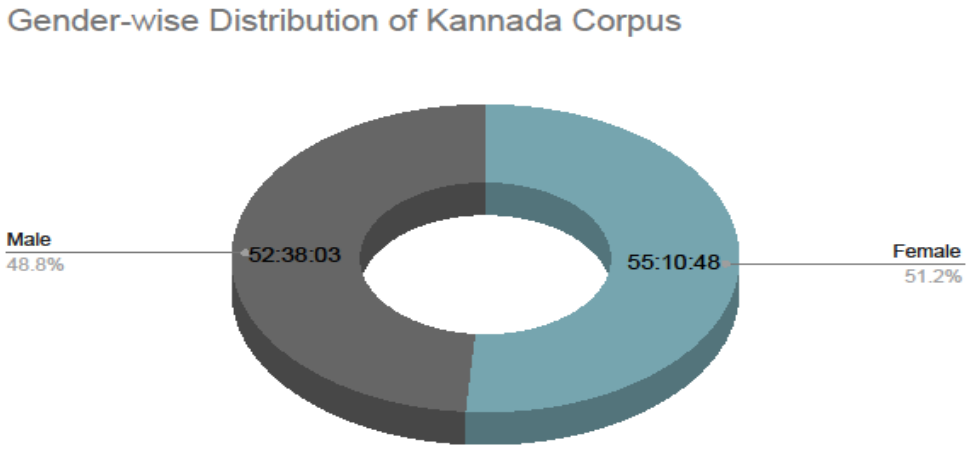

Figure 1: Gender-wise Distribution of Kannada Corpus

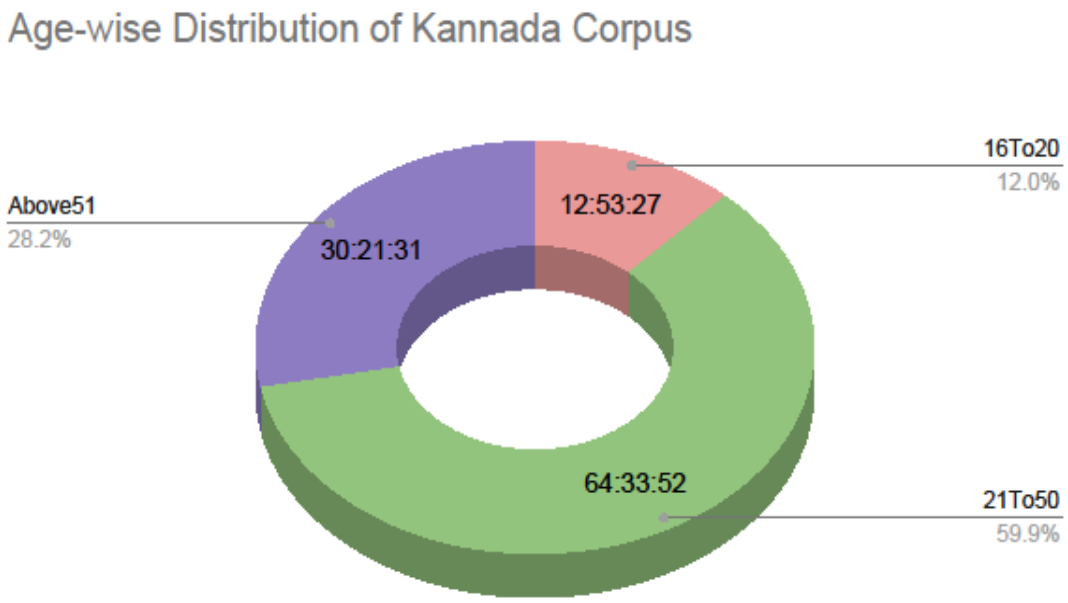

Figure 2: Age-wise Distribution of Kannada Corpus

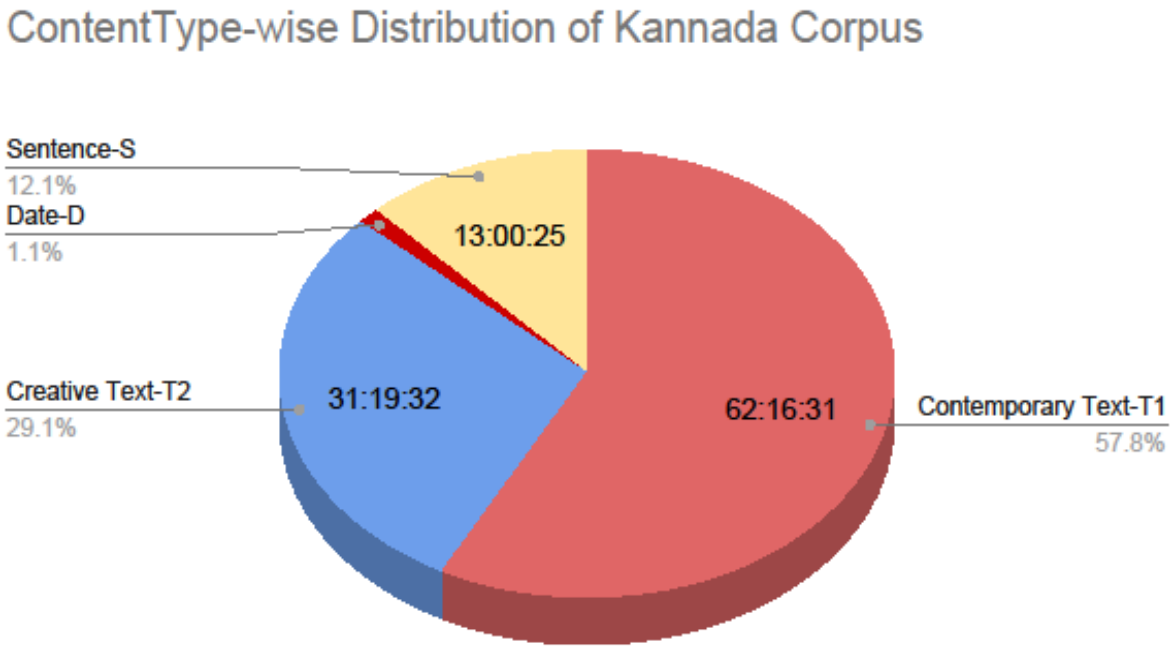

Figure 3: Content Type-wise Distribution of Kannada Corpus

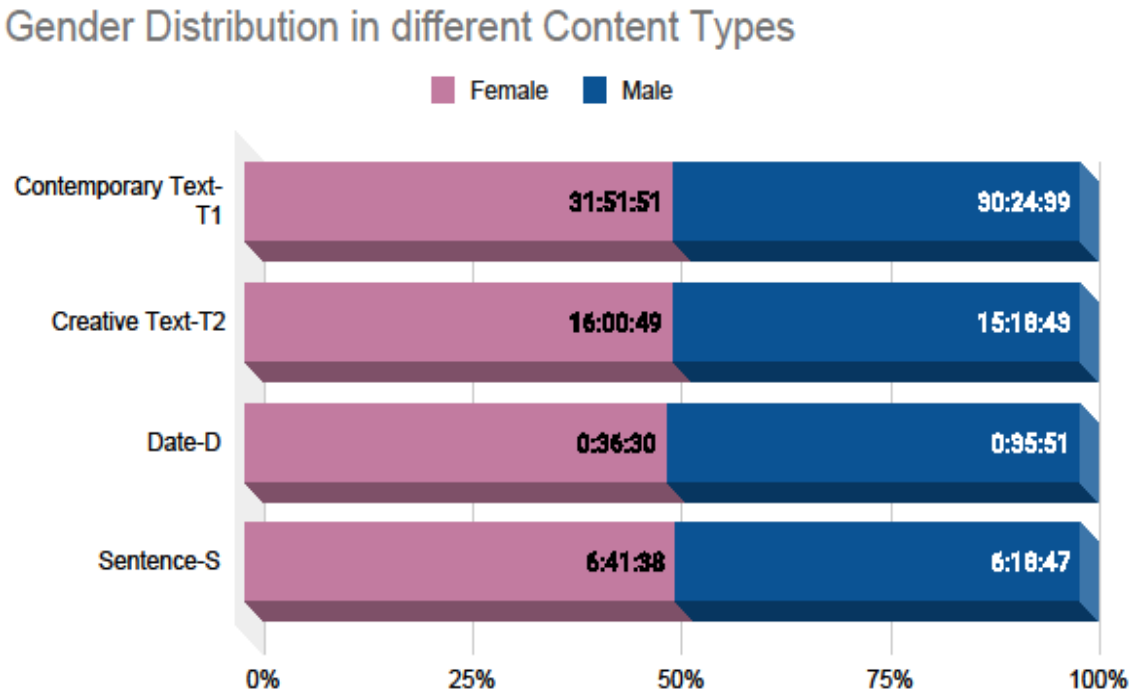

Figure 4: Gender Distribution in different Content Types of Kannada Corpus

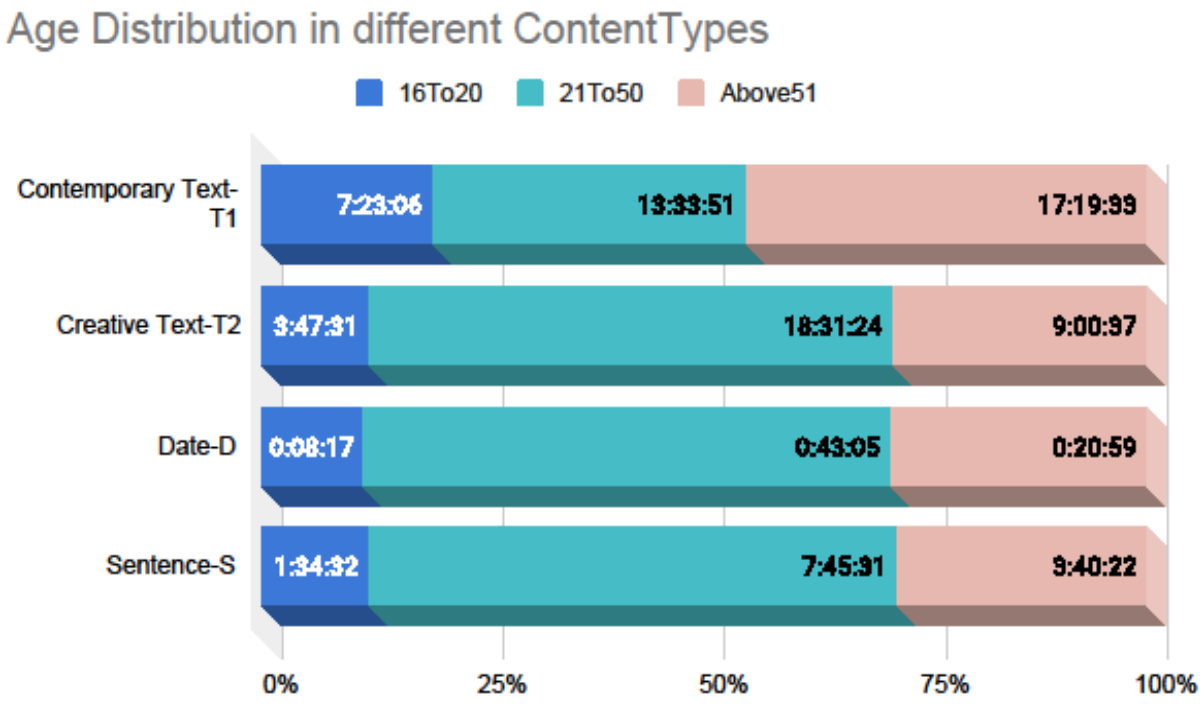

Figure 5: Age Distribution in different Content Types of Kannada Corpus

2.3.1 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 1: Representation of Kannada Sentence Aligned Speech Data Duration

2.4 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 2: Distribution of Speakers of Kannada Sentence Aligned Speech Data

2.5 REFERENCES

1. Choudhary, N. and D. G. Rao. 2020. The LDC-IL Speech Corpora in Proceedings of 23rd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), Yangon, Myanmar, 2020. pp. 28-32, doi: https://doi.org/10.1109/O-COCOSDA50338.2020.9295011
2. Choudhary, N. 2021. LDC-IL: The Indian Repository of Resources for Language Technology. Language Resources & Evaluation. Springer, Vol. 55, Issue 1. doi: https://doi.org/10.1007/s10579-020-09523-3
3. Choudhary, Narayan, Rajesha N., Manasa G. & L. Ramamoorthy. 2019. "LDC-IL Raw Speech Corpora: An Overview" in Linguistic Resources for AI/NLP in Indian Languages. Central Institute of Indian Languages, Mysore. pp. 160-174.
4. Ramamoorthy, L., Narayan Choudhary, 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.