

डोगरी वाक्य संरेखित वाक् कोष



**Dogri Sentence Aligned
Speech Corpus**

DOGRI SENTENCE ALIGNED SPEECH CORPUS



34

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

Dogri Sentence Aligned Speech Corpus

Rajesh N., Manasa G., Dr. Rejitha K. S., Stephen Fernandes, Nithin S., Roopashri M. R.,
Dr. Narayan Kumar Choudhary, Prof. Shailendra Mohan

e-ISBN:

CIIL Publication No.:

First published: AD 2025 March :: Phalguna 1946 Śaka

© Central Institute of Indian Languages, Mysuru 2025

Publisher: Prof. Shailendra Mohan, Director, CIIL

Printing & Publication Unit

Aleendra Brahma, Officer-in-Charge, Printing Press & Publication Unit
B Manju Bhargavi, Unit-in-Charge, Printing Press & Publication Unit
M.N. Chandrashekar, Compositor, Printing Press & Publication Unit

Cover design: Rajesh N

TABLE OF CONTENTS

Table of Contents	i
Figures	i
Tables	i
1 Speech Annotation	1
2 Dogri Speech Annotation	5

FIGURES

Figure 1: Gender-wise Distribution of Dogri Corpus	7
Figure 2: Age-wise Distribution of Dogri Corpus	7
Figure 3: Content Type-wise Distribution of Dogri Corpus	8
Figure 4: Gender Distribution in different Content Types of Dogri Corpus	8
Figure 5: Age group Distribution in different Content Types of Dogri Corpus	8

TABLES

Table 1: Representation of Dogri Sentence Aligned Speech Data Duration	9
Table 2: Distribution of Speakers of Dogri Sentence Aligned Speech Data	9

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 raw 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 DOGRI SPEECH ANNOTATION

Rajasha N., Narayan Kumar Choudhary

2.1 OVERVIEW OF SENTENCE ALIGNED SPEECH CORPUS

Dogri Sentence Aligned Speech Corpus is created by annotating the speech data collected by LDC-IL. A detailed explanation of the Dogri Speech Corpus is available [3] and [4]. LDC-IL Dogri Sentence Aligned Speech Corpus files contain an audio file and its corresponding textual layer. 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
'Dogri_Female_16To20_Contemporary_Text-T1_SP-0043_T1-0043-001.wav'

LDC-IL Sentence Aligned Speech corpus for Dogri 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 contains different way of date formats uttered by the speaker.

2.2 OBSERVATIONS

LDC-IL sentence-level speech annotation strictly follows what the speaker pronounces. 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 or over corrections when the speaker is reading.

There were also variations in how numbers were pronounced. For example, while reading sports news, the speakers read scores of different sports such as cricket, tennis, etc in different ways and they deviated from the standardised way of pronouncing the scores. Similarly, there were some errors in reading large numbers such as thousands or lakhs and also in reading decimals, fractions, etc. Most of the speakers faced difficulty in pronouncing foreign names which frequently appear in sports news. Abbreviations and rarely-used words also influenced the reader's fluency.

2.2.1 PHONETIC ALTERNATION IN DOGRI 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.: də-dina

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.

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.: sətʃi:f > stʃi:f

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

e. Colloquial usage

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

E.g.: bʰa:rət > pa:rt

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.: pətʃa > paʃa

g. Substandard alternation

It has been observed that some speakers have consistently replaced the cluster sounds with nearby or easily articulatable sound.

E.g.: kʃetʃr > ketʃr

2.3 SUMMARY OF THE CORPUS

The total duration of Dogri Sentence Aligned Speech Corpus is 08:32:54 (hh:mm:ss) comprising 5,039 audio segments from 61 speakers. The following figures show the distribution of the corpus with respect to gender, age and content type, respectively. The tables that follow gives a break-up of the corpus in terms of recordings obtained from different kinds of texts and also other demographic details. and the age and gender-wise distribution of all the speakers.

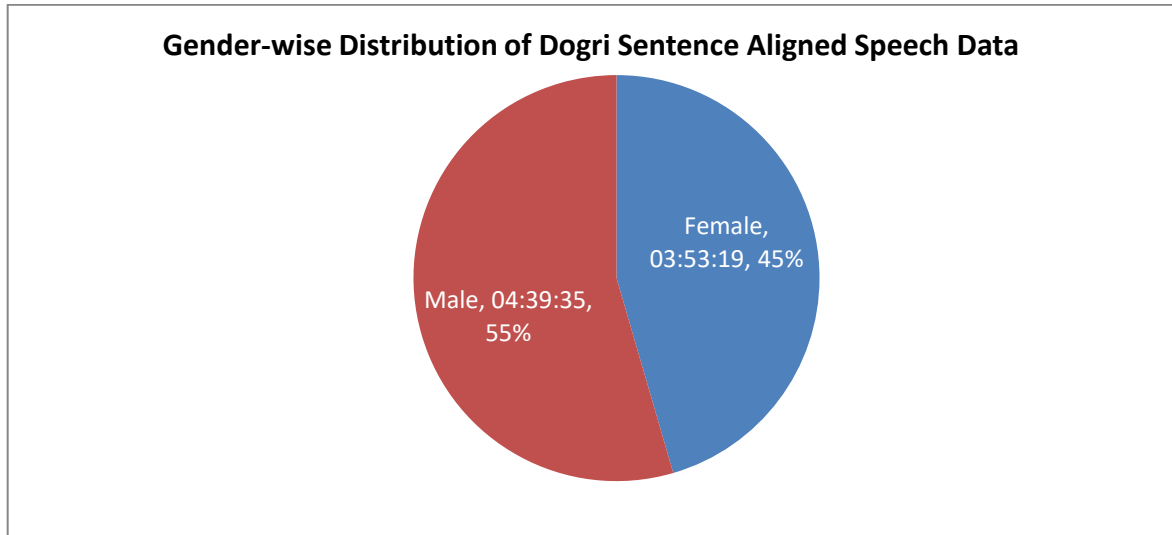


Figure 1: Gender-wise Distribution of Dogri Corpus

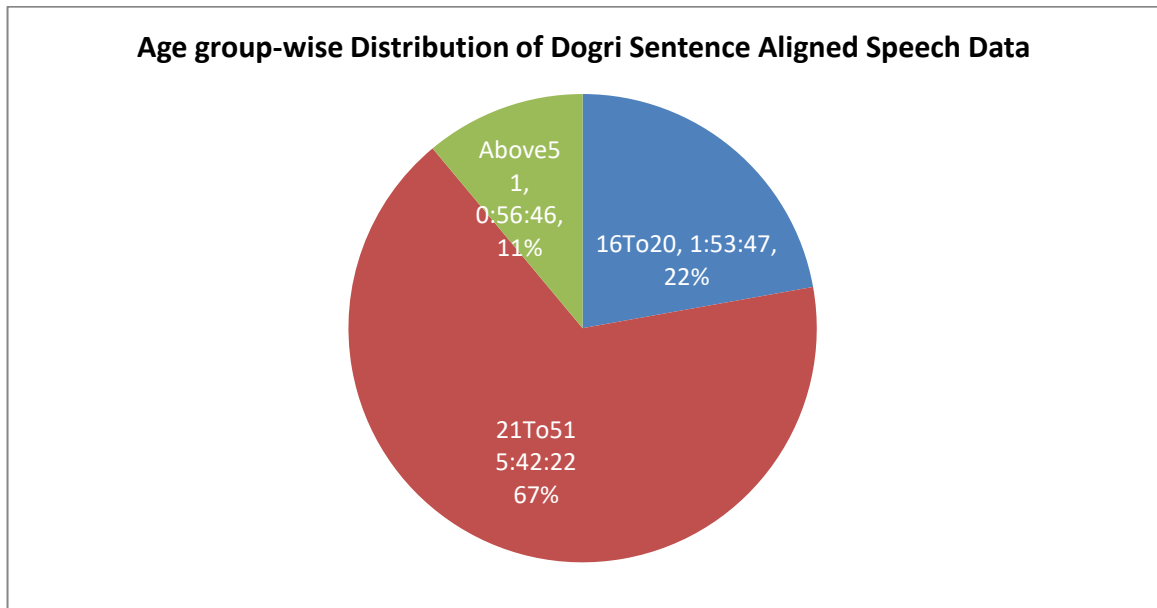


Figure 2: Age-wise Distribution of Dogri Corpus

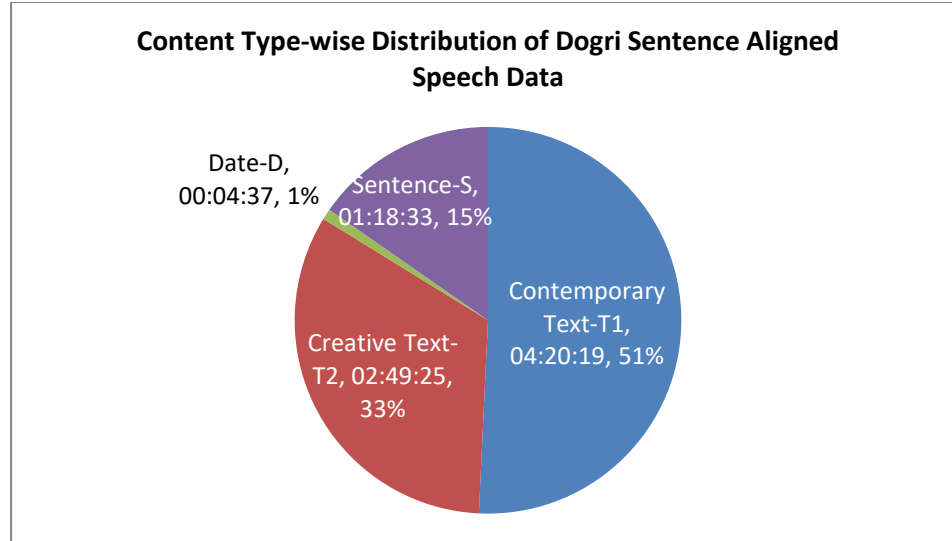


Figure 3: Content Type-wise Distribution of Dogri Corpus

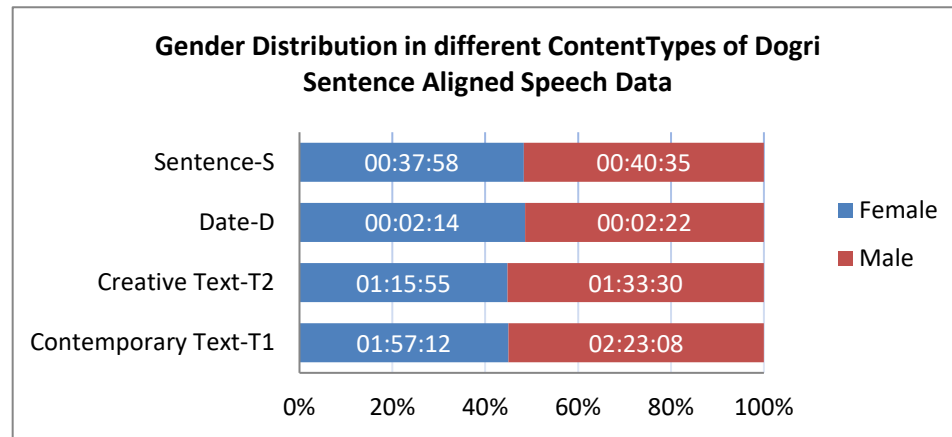


Figure 4: Gender Distribution in different Content Types of Dogri Corpus

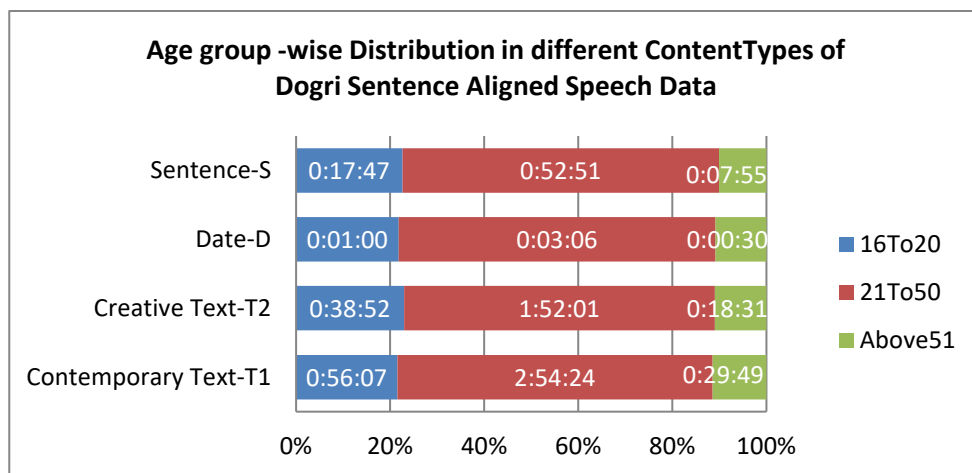


Figure 5: Age group Distribution in different Content Types of Dogri Corpus

2.3.1 DURATION OF DOGRI SENTENCE ALIGNED SPEECH DATA

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

Content Type	Gender	Age Group	Duration (hh:mm:ss.ms)		
Contemporary Text-T1	Female	16To20	00:27:04.637164	01:57:11.807157	04:20:19.703495
		21To50	01:16:47.081918		
		Above51	00:13:20.088075		
	Male	16To20	00:29:02.080940	02:23:07.896338	
		21To50	01:37:36.474275		
		Above51	00:16:29.341124		
Creative Text-T2	Female	16To20	00:22:08.133904	01:15:54.910421	02:49:24.888109
		21To50	00:46:39.926536		
		Above51	00:07:06.849980		
	Male	16To20	00:16:43.964951	01:33:29.977689	
		21To50	01:05:21.556486		
		Above51	00:11:24.456252		
Date-D	Female	16To20	00:00:35.268053	00:02:14.687596	00:04:37.003685
		21To50	00:01:24.840992		
		Above51	00:00:14.578550		
	Male	16To20	00:00:25.223487	00:02:22.316089	
		21To50	00:01:41.610974		
		Above51	00:00:15.481628		
Sentence-S	Female	16To20	00:10:09.902085	00:37:58.074697	01:18:32.645553
		21To50	00:24:05.503304		
		Above51	00:03:42.669308		
	Male	16To20	00:07:37.388053	00:40:34.570856	
		21To50	00:28:45.039539		
		Above51	00:04:12.143265		

Table 1: Representation of Dogri Sentence Aligned Speech Data Duration

2.3.2 SUMMARY OF SPEAKERS

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

Age Group	Female	Male	Total
16To20	8	5	13
21To50	19	23	42
Above51	3	3	6
Total	30	31	61

Table 2: Distribution of Speakers of Dogri Sentence Aligned Speech Data

2.4 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 (ed.), *Linguistic Resource For AI/NLP in Indian Languages*. 2019. Central Institute of Indian Languages, Mysuru. ISBN No 978-81-7343-295-8
4. Choudhary, Narayan (ed.), *Compendium of Linguistic Resources in Indian Languages*. 2021. Central Institute of Indian Languages, Mysuru. ISBN No.: 978-81-948885-6-7
5. Narayan Kumar Choudhary, Sunil Kumar Choudhary, Rajesha N., Manasa G., 2021. Dogri Raw Speech Corpus. Central Institute of Indian Languages, Mysore.
6. Ramamoorthy, L., Narayan Choudhary & Sunil Kumar. 2019. A Gold Standard Dogri Raw Text Corpus. Central Institute of Indian Languages, Mysore