COMPENDIUM OF LINGUISTIC RESOURCES IN INDIAN LANGUAGES

Editor: Narayan Choudhary

CENTRAL INSTITUTE OF INDIAN LANGUAGES
Department of Higher Education,
Ministry of Education, Government of India,
Manasgangotri, Mysore

Compendium of Linguistic Resources in Indian Languages

2021

Central Institute of Indian Languages

Department of Higher Education, Ministry of Education, Government of India,

Manasagangotri, Mysuru

Compendium of Linguistic Resources in Indian Languages

Editor: Narayan Choudhary

Publication No.: 1274

First Published: *AD, 2021 May*
Vaisakha 1943 Saka

Director
Central Institute of Indian Languages
Manasagangotri, Hunsur Road, Mysuru – 570 006, INDIA

Phone: 0091/0821-2515006 (Director) Fax: 0091/0821-2515032
Grams: BHARATI Website: http://www.ciil.org
E-mail: director-ciil@gov.in

For further information contact:

Head, Publication Unit	For Publication orders
Email: publication.kar-ciil@nic.in Ph : 0821-2345026	Contact Publication Unit Ph : 0821-2345182, 09845565614 Email: publication.kar-ciil@nic.in

ISBN No.: 978-81-948885-6-7

Price: Rs. 250/- (print copy) Rs. 0 (for e-copy)

Published by	:	Prof. C. G. Venkatesha Murthy, Director
Head, Press & Publication	:	Prof.Uma Pappuswamy, Professor-cum-Deputy Director
Officer-in-Charge	:	Sri. Aleendra Brahma, Lecturer-cum-JRO
Printing Supervision by	:	Sri. R.Nandeesh, Sri. M.N.Chandrashekar & Sri. H.Manohara
Cover Design	:	Smt. Shweta Ramesh, Artist, Bharatavani
Printed at	:	CIIL, Printing Press, Mysuru

TABLE OF CONTENTS

FOREWORD

C. G. Venkatesha Murthy, Director, CIIL

LDC-IL has been in existence for more than 12 years now. This is the first scheme fully funded by the Government of India and dedicated solely to develop language resources particularly to suit the needs of the language technology development in and across Indian languages. Its inception was well thought with a visionary goal to promote Indian languages in the digital platforms.

Towards meeting these goals, the first set of language resources were released in April 2019 by the Hon'ble Vice President of India. A total of 13 languages were supported at that time. We are happy to release a total of 6 additional languages in that set by releasing the raw text and speech corpora in Assamese, Dogri, Kashmiri, Gujarati, Odia, and Tamil.

English has attained a kind of pseudo-lingua franca in India, specially among the higher echelons of literate people across India. It also continues to function as an associate official language of the Union on India along with its judiciary and higher studies. Indian English is also recognized widely as a variety of English. However, even within Indian English, there are multiple types of English characterized by the various Indian mother tongues prevalent in India. India is home to several most widely spoken languages of the world including Hindi, Bengali, Urdu, Marathi, Telugu, Tamil, Kannada and so on. As per an estimate, among the top 100 spoken languages of the world, the 20 belong to India. English is not a mother tongue in India but is a second language to many in India. As per census report 2011, English is reported as the mother tongue of a total of 259678 people spread across India. This is a negligible figure compared to other mother tongues of India. However, given the importance of English and its widespread use in India, it is always a dire need to study the variety of speech called English in India.

It has been proven that different mother tongues do affect how a second language is spoken. This has led us develop a speech corpus of Indian English as spoken by various speech communities. Towards this goal, LDC-IL collected a speech corpus of Indian English as spoken by the native speakers of Bengali and that of Kannada. These two speech corpora will prove to be test bed for various kinds of linguistic studies on these two varieties of Indian English which will be useful not just for language technology development such as automated speech recognition but also for better linguistic insights in Indian English.

In this release, LDC-IL is also releasing a multilingual dataset containing small samples of 19 scheduled Indian languages extracted from the raw speech corpora of the respective languages. This multilingual raw speech corpus is prepared to suit the needs various tasks such as testing a multilingual ASR or a language identification module. It can also serve the support acoustic and phonetic studies across Indian languages.

A lot has gone into developing these language resources and supporting the Indic languages. We hope the research and development community within India and across the world, including the academia and the industry will reap the benefit out of these and Indian languages will have better artificial intelligence-based support.

1 ASSAMESE RAW TEXT CORPUS

Taznin Hussain, Priyanshe Adhyapak, Rajesha N., Manasa G., Narayan Choudhary

1.1 INTRODUCTION

Assamese or Oxomiya is the language spoken by the natives of the state of Assam in Northeast India. It is also the official language of Assam. It is spoken in some parts of Arunachal Pradesh, Nagaland and in other Northeast Indian states. However, small pockets of Assamese speakers can also be found in Bhutan and Bangladesh. The origin and growth of the Assamese language is not simple and clear. Some writers think that its source is to be found in the Sanskrit or Vedic literature. But, Assamese, a branch of Indo-Aryan along with the cognate languages, Maithili, Bengali and Oriya, developed from Magahi Prakrit. According to linguist Suniti Kumar Chatterji, the Magahi Prakrit in the east gave rise to four Apabhramsa dialects namely-Radha, Vanga, Varendra and Kamarupa. The Kamarupa Apabhramsa gave rise to the Assamese language in the Brahmaputra valley. Though early compositions in Assamese exist from the 13th century yet the earliest relics of the language can be found in paleographic records of the Kamarupa Kingdom from the 5th century to the 12th century. Some compositions also date to the 14th century, during the reign of the Kamata king Durlabh Narayana of the Khen dynasty, when Madhav Kandali composed the Kotha Ramayana.

Assamese text corpus is collected from various libraries in Assam mostly from Guwahati and Jorhat. The greater part of the text has been taken from Guwahati District Library and Jorhat District Library. LDC-IL tried to cover the entire category in its standard list. Some categories like novel, short stories have huge amount of text but some categories like physics, chemistry, economics have very less amount of text. Literary texts are easily available in Assamese but getting scientific text is exceedingly difficult. In some categories, like epigraphy, finance, oceanology, texts are too rare to find in Assamese.

1.2 PERIOD OF ASSAMESE TEXT INCLUDED IN THIS CORPUS

While modern assamese language has a continuum of history starting from 14th century, for the purpose of this corpus, we have focused on modern Assamese writing and no text written prior to 1990 is part of it.

1.3 PECULIARITIES OF ASSAMESE LANGUAGE

The Corpus of Assamese text can be broadly classified into two: literary text and non-literary text. These two explicitly show their differences in terms of frequency of word usage and variety that it brings into corpus. Literary texts are texts in narrative style, and they also contain elements of fiction. Novels, short stories, plays are examples of literary text. Non-literary texts are text whose primary purpose is to convey information. Examples of non-literary texts are text about various scientific or technical subjects, legal documents, articles in academic journals. In literary text, language has emotional elements, cultural information, dialectical variations, ambiguity etc. But technical or scientific terms, foreign words etc. have widely appeared in non-literary texts.

Some of salient features of Assamese language are discussed below.

The Assamese phonemic inventory consists of eight vowels, ten diphthongs and twenty consonants. The velar fricative /x/ is the hallmark of this language. Assamese language is rich in consonant clusters. Eastern Indic languages like Assamese, Bengali, Sylheti, and Oriya do not have a vowel length distinction rather it has a wide set of back rounded vowels. In case of Assamese there are four back rounded vowels.

Gender in Assamese is not grammatically marked. Another important feature of this language is that the verbs in Assamese are negativized by adding /-n/ before the verb with /n/ picking up the initial vowel of the verb.

For example, *lage* which means "want" and to negativize it a negative particle /-na/ is added to the verb like *nalage* meaning "does not want". Assamese has a huge collection of classifiers which are extensively used for different objects. The basic word order in Assamese language is SOV. However, it normally allows scrambling, that is, Assamese, to a certain extent has a flexible word order system.

1.4 ASSAMESE SCRIPT

The Assamese script which is also known as Oxomiya Akhor or Oxomiya Lipi is a variant of the Eastern Nagari script and is also used for Bengali and Bishnupriya Manipuri. In Unicode, it may be referred to as Bangla or Bengali script. The Eastern Nagari script belongs to the Brahmic family of scripts and has a continuous history of development from Nagari script, a precursor of Devanagari.

1.5 PRINCIPLES OF DATA SAMPLING

Assamese text data sampling strictly followed the generic guidelines of LDC-IL text corpus collection which are noted in the generic LDC-IL corpus documentation.

1.6 FIELD WORKS UNDERTAKEN

Assamese text corpus is collected from various libraries in Assam, mostly from Guwahati. The text materials were collected by conducting six field works in the period of 2008 to 2012. The greater part of the text has been taken from different library of Guwahati and Jorhat.

Overall, the following libraries served as the source of the Assamese text corpus:

- J B College, Jorhat.
- Assam Agricultural University, Jorhat.
- District Library, Jorhat.
- Personal Libraries, Jorhat.
- District Library, Guwahati.
- NERLC Library, Guwahati.
- Cotton College Library, Guwahati.
- Personal Libraries, Guwahati.

Collected text materials have been published at various places within Assam and other states of India, including Delhi.

Collecting text data from the field is a difficult job. Most of the libraries do not allow taking huge amount of text from their shelves at a time because it is against their rules and principles. For a particular period, they issue maximum three or four books. Even if the librarian allowed taking many books at a time, the photocopy kiosk had issues as there was a long queue.

Sometime photocopy attendants refused to photocopy randomly selected pages because of the long queue waiting and it took more time for them to turn the pages compared to continuous page photocopying that they are accustomed to. The field worker/linguist had to carry a huge list of photocopy bundles with them, which was many times burdensome to travel with.

Despite all the issues mentioned above, the linguists working on the data collection had to deal with and get going.

1.7 DATA INPUTTING

All the texts have been typed in Unicode using the InScript Keyboard directly onto the XML files. The data sets have been inputted by Ms. Roopa Debi, Rina Sarkar and Golap Borah who are the native speakers of Assamese.

1.8 VALIDATION AND NORMALIZATION WORKSHOPS

A 10-day workshop was conducted at Linguistic Data Consortium for Indian Languages from 24-June-2010 to 05-July-2010 with Ms. Karabi Gogoi (RP), Ms. Rita Sharma (RP), Sewali Deka (RP) and Pikumoni Chutia (RP) as experts. The experts suggested that the Assamese text corpus should remain true to the text and typo errors have been corrected.

1.9 PROOFREADING

Assamese text data has been proofread by internal Resource Persons such as Atreyee Sharma, Ashmrita Gogoi, Jahnobi Kalita, Muslima Begum, Chuchen Dutta, Amrit Upadhyay, Taznin Hussain, Priyanshe Adhyapak, Anupama Rabha, Arpan Jyoti Gogoi, Karishma Hazarika, Bidyut Bezbaruah, Bijoy Krishna Doley, Hemanta Konch, Rini Dehingia, Sharmistha Saikia, Jitu Borah, Rehna Sultana, Tulika Sarmah.

The collected text materials are contemporary and mainly published after 1990.

1.10 TRANSLITERATIONS IN LDC-IL ASSAMESE TEXT CORPUS

For easy reference and to maintain uniformity in metadata, some entries namely the *'Title', 'Headline', 'Author', 'Editor', 'Translator'* are transliterated from Assamese script to Roman letters. Numeric characters are transliterated from Assamese to Hindu-Arabic system. The LDC-IL transliteration scheme of Assamese to Roman is given below:

LDC-IL Transliteration Schema
Assamese characters to Roman and Assamese Numerals to Hindu-Arabic

Vowels and Vowel Signs										
অ	আ	ই	ঈ	উ	ঊ	ঋ	এ	ঐ	ও	ঔ
	া	ি	ী	ু	ূ	ৃ	ে	ৈ	ো	ৌ
a	A	i	I	u	U	x	E	ai	O	au

Consonants				
ক	খ	গ	ঘ	ঙ
ka	kha	ga	gha	ng'a
চ	ছ	জ	ঝ	ঞ
ca	cha	ja	jha	nj'a
ট	ঠ	ড	ঢ	ণ
Ta	Tha	Da	Dha	Na
ত	থ	দ	ধ	ন
ta	tha	da	dha	na
প	ফ	ব	ভ	ম
pa	pha	ba	bha	ma

Symbols		
ং	ঃ	ঁ
M	H	m'

য	ৰ	ল	ৱ	শ	স	ষ	হ	ড়	ঢ়	য়	ৎ
ya	ra	La	wa	sha	sa	Sa	ha	D'a	Dh'a	Ya	t

Numerals (Bengali to Hindu-Arabic)									
০	১	২	৩	৪	৫	৬	৭	৮	৯
0	1	2	3	4	5	6	7	8	9

1.11 OVERVIEW OF REPRESENTED DOMAINS

At LDC-IL, the Assamese text corpus size is: 1,01,27,030 Words drawn from 1,084 different titles. The total Corpus character size is 6,39,50,126 The data can be categorized into two classes: typed+cleaned. The representation of six major domains covered and has been shown in the table below:

Domain	**Domain Word Count**	**Percentage**
Aesthetics	5233452	51.68%
Commerce	66924	0.66%
Mass Media	3354996	33.13%
Official Document	1298	0.01%
Science and Technology	372790	3.68%
Social Sciences	1097570	10.84%
Total	10127030	100.00%

Table 1: Representation of the various domains in Assamese text corpus

Each domain has several sub-domains, the following table shows the representation of main and sub-domains:

1.12 AESTHETICS

The Aesthetic domain of Assamese text corpus covers 30 subdomains bearing a total of 52,33,452 words along with the overall percentage of 51.68% The representational details are given in the table below:

#	Subdomain	Word Count	Percentage (within Subdomain)	Overall Percentage
1	Autobiographies	90101	1.72%	0.89%
2	Biographies	260592	4.98%	2.57%
3	Cinema	70760	1.35%	0.70%
4	Culture	220442	4.21%	2.18%
5	Fine Arts-Dance	80665	1.54%	0.80%
6	Fine Arts-Drawing	8119	0.16%	0.08%
7	Fine Arts-Hobbies	12749	0.24%	0.13%
8	Fine Arts-Music	53977	1.03%	0.53%
9	Fine Arts-Musical Instruments	190	0.00%	0.00%
10	Fine Arts-Sculpture	26746	0.51%	0.26%
11	Folk Tales	4560	0.09%	0.05%
12	Folklore	48052	0.92%	0.47%
13	Handicrafts	1697	0.03%	0.02%
14	Humour	29191	0.56%	0.29%
15	Literary Texts	1074762	20.54%	10.61%
16	Literature-Children's Literature	152813	2.92%	1.51%
17	Literature-Criticism	459582	8.78%	4.54%
18	Literature-Diaries	26657	0.51%	0.26%
19	Literature-Epics	14702	0.28%	0.15%
20	Literature-Essays	32880	0.63%	0.32%
21	Literature-Letters	19603	0.37%	0.19%
22	Literature-Novels	847945	16.20%	8.37%
23	Literature-Plays	86783	1.66%	0.86%
24	Literature-Poetry	9800	0.19%	0.10%
25	Literature-Science Fiction	63482	1.21%	0.63%
26	Literature-Short Stories	1242392	23.74%	12.27%
27	Literature-Speeches	83207	1.59%	0.82%
28	Literature-Text Books (School)	33804	0.65%	0.33%
29	Literature-Travelogues	159367	3.05%	1.57%
30	Mythology	17832	0.34%	0.18%
	Total	5233452	100%	51.68%

Table 2: Representation of Aesthetics

1.13 COMMERCE

The Commerce text corpus covers 8 subdomains bearing a total of 66,924 words along with the overall percentage of 0.66%. The representational details are given in the table below:

#	Sub domain	Word Count	Percentage (within Subdomain)	Overall Percentage
1	Accountancy	5440	8.13%	0.05%
2	Banking	558	0.83%	0.01%
3	Business	7171	10.72%	0.07%
4	Finance	1499	2.24%	0.01%
5	Industry	11974	17.89%	0.12%
6	Management	35423	52.93%	0.35%
7	Share Market	1814	2.71%	0.02%
8	Tourism	3045	4.55%	0.03%
	Total	**66924**	**100%**	**0.66%**

Table 3: Representation of Commerce

1.14 MASS MEDIA

The Mass Media text corpus covers 18 sub-domains bearing a total of 33,54,996 words along with the overall percentage of 33.13%. The representational details are given in the table below:

#	Sub domain	Word Count	Percentage (within Subdomain)	Overall Percentage
1	Article	13508	0.40%	0.13%
2	Business News	3078	0.09%	0.03%
3	Cinema News	727	0.02%	0.01%
4	Classifieds	16099	0.48%	0.16%
5	Discussions	673602	20.08%	6.65%
6	Editorial	288547	8.60%	2.85%
7	General News	1756463	52.35%	17.34%
8	Health	5047	0.15%	0.05%
9	Interviews	42847	1.28%	0.42%
10	Letters	25164	0.75%	0.25%
11	Obituary	3848	0.11%	0.04%
12	Political	179385	5.35%	1.77%
13	Religious/Spiritual News	10230	0.30%	0.10%
14	SMS	1684	0.05%	0.02%
15	Social	70414	2.10%	0.70%
16	Speeches	88151	2.63%	0.87%
17	Sports News	174889	5.21%	1.73%
18	Weather	1313	0.04%	0.01%
	Total	3354996	100%	33.13%

Table 4: Representation of Mass Media

1.15 SCIENCE AND TECHNOLOGY

The Science and Technology text corpus covers 35 sub-domains bearing a total of 3,72,790 words along with the overall percentage of 3.68%. The quantitative representation is shown in the table below:

#	Sub domain	Word Count	Percentage (within Subdomain)	Overall Percentage
1	Agriculture	74524	19.99%	0.74%
2	Astrology	30746	8.25%	0.30%
3	Astronomy	8555	2.29%	0.08%
4	Ayurveda	39037	10.47%	0.39%
5	Biochemistry	1251	0.34%	0.01%
6	Biology	5353	1.44%	0.05%
7	Biotechnology	385	0.10%	0.00%
8	Botany	25290	6.78%	0.25%
9	Chemistry	4391	1.18%	0.04%
10	Computer Sciences	5264	1.41%	0.05%
11	Criminology	2248	0.60%	0.02%
12	Engineering-Chemical	1162	0.31%	0.01%
13	Engineering-Civil	502	0.13%	0.00%
14	Engineering-Electrical	2973	0.80%	0.03%
15	Engineering-Electronics Communication	1343	0.36%	0.01%
16	Engineering-Others	712	0.19%	0.01%
17	Environmental Science	12337	3.31%	0.12%
18	Film Technology	8985	2.41%	0.09%
19	Forestry	13533	3.63%	0.13%
20	Geology	9879	2.65%	0.10%
21	Homeopathy	2684	0.72%	0.03%
22	Horticulture	6256	1.68%	0.06%
23	Logic	2631	0.71%	0.03%
24	Mathematics	3092	0.83%	0.03%
25	Medicine	50832	13.64%	0.50%
26	Microbiology	507	0.14%	0.01%
27	Naturopathy	1179	0.32%	0.01%
28	Physics	12590	3.38%	0.12%
29	Psychology	2438	0.65%	0.02%
30	Sexology	747	0.20%	0.01%
31	Statistics	2284	0.61%	0.02%
32	Textile Technology	10023	2.69%	0.10%
33	Veterinary	2567	0.69%	0.03%
34	Yoga	11184	3.00%	0.11%
35	Zoology	15306	4.11%	0.15%
	Total	372790	100%	3.68%

Table 5: Representation of Science and Technology

1.16 OFFICIAL DOCUMENT

The Official Document text corpus covers 2 sub-domains bearing a total of 1,298 words along with the overall percentage of 0.01%. The representational details are given in the table below:

Sub domain	Word Count	Percentage (within Subdomain)	Overall Percentage
Administration	990	76.27%	0.01%
Police Documents	308	23.73%	0.00%
Total	1298	100%	0.01%

Table 6: Representation of Official Document

1.17 SOCIAL SCIENCE

The Social Science text corpus covers 22 sub-domains bearing a total of 10,97,570 words along with the overall percentage of 10.84%. The representational details are given in the table below:

#	Sub domain	Word Count	Percentage (within Subdomain)	Overall Percentage
1	Anthropology	2784	0.25%	0.03%
2	Archaeology	2860	0.26%	0.03%
3	Demography	3676	0.33%	0.04%
4	Economics	125527	11.44%	1.24%
5	Education	42643	3.89%	0.42%
6	Food and Wellness	19770	1.80%	0.20%
7	Geography	14683	1.34%	0.14%
8	Health and Family Welfare	77750	7.08%	0.77%
9	History	347753	31.68%	3.43%
10	Home Science	9568	0.87%	0.09%
11	Journalism	39815	3.63%	0.39%
12	Law	24695	2.25%	0.24%
13	Linguistics	44567	4.06%	0.44%
14	Personality Development	5566	0.51%	0.05%
15	Philosophy	59189	5.39%	0.58%
16	Physical Education	13677	1.25%	0.14%
17	Political Science	74949	6.83%	0.74%
18	Public Administration	6756	0.62%	0.07%
19	Religion/Spiritual	63025	5.74%	0.62%
20	Sociology	67308	6.13%	0.66%
21	Sports	44452	4.05%	0.44%
22	Textbook (Social Science)	6557	0.60%	0.06%
	Total	1097570	100%	10.84%

Table 7: Representation of Social Science

2 ASSAMESE RAW SPEECH CORPUS

Plabita Bora, Priyanshe Adhyapak, Rajesha N.,Manasa G., Narayan Choudhary

2.1 INTRODUCTION

Assamese, an official language of Assam and is a language that belongs to the Indo-Aryan language family. Its linguistic presence is widely presented in the state of Assam and some parts of Arunachal Pradesh and Nagaland. According to 2011 census, the Assamese Language is spoken by 15 million speakers. It is said to be developed from Kamrupi Prakrit and Old Assamese. At present the script used for Assamese writing is Assamese which is shared also with Bengali and a few other other languages.

The Assamese phonemic inventory consists of total forty-one phonemes that include eight vowels, ten diphthongs and twenty three consonants. The velar fricative /x/ is known as the hallmark of the Assamese language. The language has a complex consonant cluster system. The morphological salience of Assamese lies on its rich case marking system. Known as a tripartite split ergative language, Assamese exhibits the subject+object+verb sentence structure. However, in some contexts, it shows some level of flexibility in its word order. Unlike many other Indo-Aryan languages of India, grammatical gender distinction is not present in Assamese. In this aspect and many others for that matter, Assamese shares many similarities with the Bengali Language.

Assamese a widely spoken language does encounter several dialectal variations. The regional dialects can be broadly divided into two parts - the Eastern Group and the Western Group. The Eastern Group essentially includes the region in and around Sivasagar District. This group is relatively homogenous in nature. On the other hand, the Western Group includes the Western region of the state. This group exhibits heterogeneity. According to some recent linguistic studies, the two dialectal groups can be further divided into four different varieties namely Eastern, also known as Xiboxagoria, Central, Kamrupi and Goalparia Group. Here, the Central Group is spoken in Nagaon, Sonitpur, Morigaon and other adjoining areas; the Kamrupi variety is spoken in Kamrup, Barpeta, Nalbari etc.; and the Goalparia Group is spoken in the Goalpara region.

LDC-IL divided the Assamese speaking areas into these four regions and have collected speech data from each speaker. Following are some of the regions from where the LDC-IL Assamese Speech Data is collected and is listed in the table below:

Region→	Xiboxagoria	Central Assam	Kamrupi	Goalparia
Places →	1. Sivasagar 2. Golaghat 3. Jorhat 4. Dibrugarh 5. Tinsukia	1. Sonitpur 2. Nagaon 3. Morigaon 4. Lakhimpur 5. Dhemaji	1. Guwahati 2. Kamrup Rural 3. Nalbari 4. Baksa	1. Goalpara 2. Dhubri

		6. Darrang	5. Barpeta	

Table 8: Regions and Places Covered for Assamese Speech Data

2.2 DATASET PREPARATION FOR ASSAMESE

For the selected regions, Xiboxagoria, Central Assam, Kamrupi and Goalparia LDC-IL prepared the following dataset by which the prompt sheets were prepared.

Content Type	Count
Creative Text	9
Date	2
Command and Control Words	254
Form and Function Words	267
Phonetically Balanced Words	439
Person Name	500
Place Name	341
Sentences	200

Table 9: LDC-IL Speech Dataset

Distinct News Items were prepared to get the audio recording of contemporary text. It was made sure that each selected news item had minimum 500 words. Each prompt sheet had a distinct news item and selected part of the dataset prepared as follows.

Content Type	Content that Each Typical Prompt Sheet had	Content Selection Type
Contemporary Text	1 Text	Distinct Text
Creative Text	1 text	Random Text selected from dataset*
Sentences	25 Sentences	Random set selected from dataset*
Command and Control Words	30 Words	Random set selected from dataset*
Person Names	20 Words	Random set selected from dataset*
Place Names	10 Words	Random set selected from dataset*
		*randomly selected by machine

Table 10: Table of Contents in LDC-IL Dataset

The Full Set of

1. Phonetically Balanced Vocabulary
2. Form and Function Words

were also carried to the field to get recorded by selected individuals. Once all these preparations were made, the investigator started collecting the data.

The Collection of data is carried out by Atreyee Sharma in 2008 and 2009 in two different phases in the field for different regions and Plabita Bora brought in 6 speakers into the corpus in 2019.

2.3 TRANSLITERATIONS IN LDC-IL ASSAMESE READ CORPUS

For easy reference and uniformity, the recorded text in the metadata file, is transliterated from Assamese to Roman letters. Numeric characters were transliterated from Assamese to Hindu-Arabic system.

The LDC-IL transliteration scheme of Assamese to Roman is given below:

LDC-IL Transliteration Schema
Assamese characters to Roman and Assamese Numerals to Hindu-Arabic

Vowels and Vowel Signs										
অ	আ	ই	ঈ	উ	ঊ	ঋ	এ	ঐ	ও	ঔ
	া	ি	ী	ু	ূ	ৃ	ে	ৈ	ো	ৌ
a	A	i	I	u	U	x	E	ai	O	au

Consonants				
ক	খ	গ	ঘ	ঙ
ka	kha	ga	gha	ng'a
চ	ছ	জ	ঝ	ঞ
ca	cha	ja	jha	nj'a
ট	ঠ	ড	ঢ	ণ
Ta	Tha	Da	Dha	Na
ত	থ	দ	ধ	ন
ta	tha	da	dha	na
প	ফ	ব	ভ	ম
pa	pha	ba	bha	ma

Symbols		
ং	ঃ	ঁ
M	H	m'

য	ৰ	ল	ৱ	শ	স	ষ	হ	ড়	ঢ়	য়	ৎ
ya	ra	La	wa	sha	sa	Sa	ha	D'a	Dh'a	Ya	t

Numerals (Bengali to Hindu-Arabic)									
০	১	২	৩	৪	৫	৬	৭	৮	৯
0	1	2	3	4	5	6	7	8	9

.

2.4 SUMMARY OF THE CORPUS

In the sections below, we provided the tabular details of the different content types of the Assamese raw speech corpus based on various yardsticks which can also be filtered out from the metadata sheet. These figures may be helpful in tuning the corpus for various purposes of the training, testing, and evaluating various algorithms as well as providing useful insights into the datasets. The total duration of the dataset is 54:21:12 (hh:mm:ss) comprising 37,570 audio segments.

2.5 SUMMARY OF THE AUDIO SEGMENTS

The table below shows the total number of Audio Segments and their distribution in the Assamese speech dataset.

LDC-IL Assamese Speech Data Status	Gender →	Female			Male		
	Age Group →	16-20 Years	21-50 Years	50+ Years	16-20 Years	21-50 Years	50+ Years
Content Type	Total Segments	Segments	Segments	Segments	Segments	Segments	Segments
Contemporary Text (News)-T1	304	16	97	41	16	94	40
Creative Text-T2	304	16	97	41	16	94	40
Sentence-S	7593	399	2423	1025	398	2349	999
Date-D	599	32	192	80	32	187	76
Command and Control Words-W1	9118	479	2910	1228	479	2823	1199
Person Name-W2	6081	320	1938	819	319	1886	799
Place Name-W2	3044	160	969	410	160	945	400
Phonetically Balanced-W4	6567	875	2188	876	875	875	878
Form and Function Word-W5	3960	528	1319	528	527	529	529

Table 11: Audio Segments and their Distribution

2.6 DURATION OF THE RAW SPEECH DATA

The table below shows the duration of each of the content type and their distribution across a few factors.

LDC-IL Assamese Speech Data Status	Gender →	Female			Male		
	Age Group →	16-20 Years	21-50 Years	50+ Years	16-20 Years	21-50 Years	50+ Years
Content Type	Total Duration	Duration (hh:mm:ss)	Duration (hh:mm:ss)	Duration (hh:mm:ss)	Duration (hh:mm:ss)	Duration (hh:mm:ss)	Duration (hh:mm:ss)
Contemporary Text (News)-T1	17:23:25	0:50:39	05:18:38	2:24:40	0:56:35	5:26:53	2:26:00
Creative Text-T2	11:44:37	0:38:07	3:32:03	1:35:24	0:41:10	3:42:50	1:35:03
Sentence-S	5:55:29	0:19:56	1:49:53	0:47:06	0:19:49	1:51:35	0:47:13
Date-D	0:33:59	0:01:29	0: 10:57	0:05:08	0:01:21	0:10:13	0:04:51
Command and Control Words-W1	4:56:49	0:15:17	1:34:47	0:40:32	0:13:27	1:30:53	0:41:53
Person Name-W2	5:38:07	0:18:45	1:47:25	0:46:42	0:16:37	1:43:32	0:45:06
Place Name-W2	1:58:33	0:06:08	0: 37:27	0:16:28	0:05:54	0:36:27	0:16:09
Phonetically Balanced-W4	3:41:45	0:33:34	1:25:12	0:27:39	0:20:59	0:25:33	0:28:48
Form and Function-Word-W5	2:28:28	0:25:04	0:54:02	0:19:12	0:13:37	0:17:21	0:19:12

Table 12: Duration of the Collected Data

2.7 DISTINCT SET

The Distinct Set usually contains data which is distinct to each speaker and is rarely repeated. The LDC-IL speech data set contains newspaper extracts, which are read by the each speaker.

2.7.1 The Contemporary Text (News) - T1

The Distinct Text extracts from Newspapers is recorded from the informants to get the speech data of contemporary text. The distribution of the data is as follows:

Age Group	**Total Audio Segments**	**Gender-wise Distribution**		**Region-wise Distribution**							
				Xiboxagoria		**Kamrupi**		**Central Assam**		**Goalparia**	
		Female	**Male**	**Female**	**Male**	**Female**	**Male**	**Female**	**Male**	**Female**	**Male**
16 to 20	32	16	16	8	8	7	3	1	5	0	0
21 to 50	191	97	94	40	38	34	29	23	26	0	1
50+	81	41	40	14	13	15	15	10	9	2	3
Total	304	154	150	62	59	56	47	34	40	2	4

Table 13 Distribution of Assamese Contemporary Text (News) Data

2.8 RANDOM SET

The Random Set comprises of content types which are sampled by machine for each speaker. They are sampled from collection of master data sets available. The random sets are given below.

2.8.1 The Creative Text-T2

The prepared dataset consists of 9 literary pieces. From this dataset, any one randomly selected text was recorded from each informant to get the speech data of creative text. The distribution of the data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution							
				Xiboxagoria		Kamrupi		Central Assam		Goalparia	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 to 20	32	16	16	8	8	7	3	1	5	0	0
21 to 50	191	97	94	40	38	34	29	23	26	0	1
50+	81	41	40	14	13	15	15	10	9	2	3
Total	304	154	150	62	59	56	47	34	40	2	4

Table 14: Distribution of Assamese Creative Text

2.8.2 The Date-D

The answers to 2 questions listed in the dataset to get the date format of the informants. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution							
				Xiboxagoria		Kamrupi		Central Assam		Goalparia	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 to 20	64	32	32	16	16	14	6	2	10	0	0
21 to 50	379	192	187	78	76	68	58	46	51	0	2
50+	156	80	76	28	26	30	28	18	16	4	6
Total	599	304	295	122	118	112	92	66	77	4	8

Table 15: Distribution of Assamese Date Format

2.8.3 The Sentences-S

The Sentences contain a list of 200 sentences that is a representation of almost all the phonemes occurring in Assamese. 25 Randomly selected sentences are recorded from each speaker. The distribution of the data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution							
				Xiboxagoria		Kamrupi		Central Assam		Goalparia	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 to 20	797	399	398	199	200	175	75	25	123	0	0
21 to 50	4772	2423	2349	999	948	850	721	574	655	0	25
50+	2024	1025	999	350	325	375	374	250	225	50	75
Total	7593	3847	3746	1548	1473	1400	1170	849	1003	50	100

Table 16: Distribution of Assamese Sentences

2.8.4 Command and Control Words-W1

The Command and Control Words contain a list of 254 words that is a representation of almost all the command and control words occurring in Assamese. 30 randomly selected words are recorded from the list. The distribution of the data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution							
				Xiboxagoria		Kamrupi		Central Assam		Goalparia	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 to 20	958	479	479	239	240	210	90	30	149	0	0
21 to 50	5733	2910	2823	1200	1139	1020	870	690	784	0	30
50+	2427	1228	1199	418	390	450	449	300	270	60	90
Total	9118	4617	4501	1857	1769	1680	1409	1020	1203	60	120

Table 17: Distribution of Assamese Command and Control Words

2.8.5 10.4.5 Person Names –W2

The Person Names contain a list of 500 popular pan- Indian and regional person names. 20 randomly selected names are recorded from the list. The distribution of the data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution							
				Xiboxagoria		Kamrupi		Central Assam		Goalparia	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 to 20	639	320	319	160	159	140	60	20	100	0	0
21 to 50	3824	1938	1886	799	760	680	580	459	526	0	20
50+	1618	819	799	279	260	300	300	200	179	40	60
Total	6081	3077	3004	1238	1179	1120	940	679	805	40	80

Table 18: Distribution of Assamese Person Names

2.8.6 Place Names-W2

The Place Names contain a list of 341 popular pan- Indian and regional place names. 10 randomly selected names are recorded from the list. The distribution of the data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution							
				Xiboxagoria		Kamrupi		Central Assam		Goalparia	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 to 20	320	160	160	80	80	70	30	10	50	0	0
21 to 50	1914	969	945	400	378	339	290	230	267	0	10
50+	810	410	400	140	130	150	150	100	90	20	30
Total	3044	1539	1505	620	588	559	470	340	407	20	40

Table 19: Distribution of Assamese Place Names

2.9 FULL SET

The Full Set is the master set of a certain data set which is read completely from a few selected speakers in each group. Full Sets are given below:

2.9.1 The Phonetically Balanced Vocabulary-W4

The Phonetically Balanced Vocabulary contains a list of words where almost all the phones of Assamese, have occurred in all the possible positions of a word. In full set all the 439 words are recorded from the informants and these words are uttered by them three times from the list. The distribution of the data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution			
				Xiboxagoria		Central Assam	
		Female	Male	Female	Male	Female	Male
16 to 20	1750	875	875	875	875	0	0
21 to 50	3063	2188	875	1749	875	439	0
50+	1754	876	878	876	0	0	878
Total	6567	3939	2628	3500	1750	439	878

Table 20: Distribution of Assamese Phonetically Balanced Vocabulary

2.9.2 The Form and Function Words-W5

The Form and Function Words contain a list of 266 words which is a representation of almost all the form and function words occurring in Assamese. All the words are recorded from the informants and words are uttered by them three times from the list. The distribution of the data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution			
				Xiboxagoria		Central Assam	
		Female	Male	Female	Male	Female	Male
16 to 20	1055	528	527	528	527	0	0
21 to 50	1848	1319	529	1057	529	262	0
50+	1057	528	529	528	0	0	529
Total	3960	2375	1585	2113	1056	262	529

Table 21: Distribution of Assamese Form and Function words

2.10 NATIVE SPEAKER DISTRIBUTIONS

The following table shows the distribution of native speakers of Assamese, across different regions.

Region-wise Distribution of Native Speakers											
Age Group	Total Native Speakers	Gender-wise Distribution of Native Speakers		Regions							
				Xiboxagoria		Kamrupi		Central Assam		Goalparia	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 to 20	34	17	17	9	9	7	3	1	5	0	0
21 to 50	196	101	95	43	39	34	29	24	26	0	1
50+	83	42	41	15	13	15	15	10	10	2	3
Total	313	160	153	67	61	56	47	35	41	2	4

Table 22: Distribution of Assamese Native Speakers

2.11 MOTHER TONGUE DISTRIBUTION OF THE NATIVE SPEAKERS

The following table shows the distribution of Mother Tongue of the Assamese native speakers in LDC-IL speech data.

Mother Tongue of the Native Speaker	Geographical Dialect Distribution of LDC-IL Assamese Speech Corpus				Total speaker
	Xiboxagoria	Kamrupi	Central Assam	Goalparia	
Assamese	125	101	73	6	305
Bodo	1	-	1	-	2
Miri-Mishing	-	1	1	-	2
Bengali	-	1	-	-	1
Maithili	1	-	-	-	1
Hindi	1	-	-	-	1
Punjabi	-	-	1	-	1
Total	128	103	76	6	313

Table 23: Representation of Mother Tongue Distribution of the Assamese Native Speakers

3 DOGRI RAW SPEECH CORPUS

Rajesha N., Manasa G., Sunil Kumar, Narayan Choudhary

3.1 INTRODUCTION

The Land: Jammu, Kashmir and Ladakh, three geographical entities which were brought together in 1846 by Maharaja Gulab Singh, had distinct languages, cultures and religions. Jammu is located on the map of India on the northern part bordering Punjab in the South, Kashmir in the North, Himachal Pradesh in the East. Tradition has it that the foundation of Jammu town was laid by Jambu Lochan, a scion of the Surya Dynasty of Ayodhya. It is said that Jambu Lochan while hunting at the foothill of the Shivaliks, reached a place where he saw a goat and lion drinking water at the same place. On inquiry he was told that the environment of the place was so pious that none bore any enmity. So impressed by the sight, he decided to set up a town after his name. The town came to be known as Jambupura which later got corrupted to Jammu.

The People-Dogras: The Dogras are of the Aryan race. The settlers in hills that edge the Punjab at all events of those have retained Hindu faith, near the name Dogra and the country they inhabit called Duggar. According to the Imperial Gazetteer of India the origin of the word "Dogras" probably is that the Dogra traditionally existed between the two lakes of Surinsar and Mansar. The people are divided into castes and in the same way, as the Hindus of India with some local variations. Generally, these are partly the race distinction and partly the outcome of occupation which became hereditary.

The Language-Dogri: Dogri, the language of the Dogras, belongs to the Indo-Aryan group and is the first major language of the multi-lingual region i. e. Jammu of the Jammu & Kashmir state. It derives its name from 'Duggar', the ancient title of this region. The earliest reference to Dogri (Duggar) is found in Nuh-Siphir, a *masnavi* written by Amir Khusaro in 1318-19 AD for sultan Kuttubuddin Mubark. *Sindhi O Lahori O Kashmiri O Doggar, Dhur, Samundari Tilangi O Gujar.* Here *doggar* refers to the language of Duggar. Earlier, Dogri had its own script namely "Dogare Akkhar"or "Dogare" based on Takri script which is closely related to the Sharada script employed by Kashmiri language. This script was the script of the official language during the regime of Maharaja Ranbir Singh (1857-1885 AD). After the independence, the state government constituted a committee on 29th October 1953 headed by Sh. Girdhari Lal Dogra. The Committee presented a report and accordingly the state government decided to adopt Devanagri as well as Persian script for Dogri and it was incorporated in the State Constitution in 1957.

Dogri is a morphologically rich language having the pre-dominant word order of Subject-Object-Verb (SOV) with a flexibility to rearrange the constituents as many Indian languages allow. Nouns are generally inflected for number, gender, and case. There are two numbers –singular

and plural; two genders-masculine and feminine; and three cases- simple, oblique, and vocative. The oblique forms occur when a noun or noun phrase is followed by a postposition. Nouns are inflected according to their gender and the word final sound. This language is taught as a subject in the educational institutions of the region right from the 1st to post-graduate classes.

LDC-IL collected speech data from Jammu in the First Phase. The dataset was prepared based on the regions as determined.

3.2 DATASET PREPARATION FOR DOGRI

For the selected regions, Jammu and Kathua districts, LDC-IL prepared the following dataset using which the prompt sheets were prepared. The prompt sheets were in Devanagari Script.

Content Type	Count
Creative Text	6
Date	2
Command and Control Words	314
Most Frequent Words	1000
Form and Function Words	362
Phonetically Balanced Words	1025
Person Name	490
Place Name	370
Sentences	525

Table 24: Representation of Dogri Content Type

Distinct news items were prepared to get the audio recording of contemporary text. Each prompt sheet had a distinct news item and part of the dataset prepared as follows.

Content Type	Content in Each typical prompt sheet	Content selection type
Contemporary Text	1 Text	Distinct Text
Creative Text	1 text	Random Text selected from dataset*
Sentences	25 Sentences	Random set selected from dataset*
Command and Control Words	30 Words	Random set selected from dataset*
Person Names	20 Words	Random set selected from dataset*
Place Names	10 Words	Random set selected from dataset*
Most Frequent Words	30 Words	Random set selected from dataset*
		* selected by machine

Table 25: Representation of Dogri Prompt Sheet

The full set of

1. Phonetically Balanced Vocabulary
2. Form and Function Words
3. 1000 Most Frequent Words

were also carried to the field to get recorded by selected individuals. Once all these preparations were made, the investigator started collecting the data. The Collection of data was carried out by the first author.

3.3 TRANSLITERATIONS IN LDC-IL DOGRI READ CORPUS

For easy reference and uniformity, the recorded text in the metadata file, is transliterated from Dogri (Devanagari) to Roman letters. Numeric characters were transliterated from Dogri (Devanagari) to Hindu-Arabic system.

The LDC-IL transliteration scheme of Dogri (in Devanagari script) to Roman is given below.

LDC-IL Transliteration Schema

Dogri -Devanagari characters to Roman and Dogri Numerals to Hindu-Arabic

Vowels and Vowel Signs																	
अ	आ	इ	ई	उ	ऊ	ऋ	ॠ	ऌ	ॡ	ऍ	ऐ	ए	ऐ	ऑ	ओ	ओ	औ
	ा	ि	ी	ु	ू	ृ	ॄ	ॢ	ॣ	ॅ	े	े	ै	ॉ	ो	ो	ौ
a	A	i	I	u	U	x	X	q	Q	eo	e	E	ai	ao	o	O	au

Consonants				
क	ख	ग	घ	ङ
ka	kha	ga	gha	ng'a
च	छ	ज	झ	ञ
ca	cha	ja	jha	nj'a
ट	ठ	ड	ढ	ण
Ta	Tha	Da	Dha	Na
त	थ	द	ध	न
ta	tha	da	dha	na
प	फ	ब	भ	म
pa	pha	ba	bha	ma

Ayogavaha		
ॅं	ं	ः
M’	M	H

य	र	ल	व	श	ष	स	ह	ळ	क़
ya	ra	la	va	sha	Sa	sa	ha	La	Za

Numerals (Dogri-Devanagari to Hindu-Arabic)									
०	१	२	३	४	५	६	७	८	९
0	1	2	3	4	5	6	7	8	9

3.4 SUMMARY OF THE CORPUS

In the sections below, we provided the tabular details of the different content types of the Dogri raw speech corpus based on various yardsticks which can also be filtered out from the metadata sheet. 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 the data is 17:10:26 (hh:mm:ss) comprising 12,036 audio segments. We plan to augment this dataset in near future.

3.4.1 Summary of the Audio Segments

The table below shows the total number of audio segments and their distribution in the Dogri speech dataset.

LDC-IL Dogri Speech Data Status	Gender →	Female			Male		
	Age Group →	16-20 Years	21-50 Years	50+ Years	16-20 Years	21-50 Years	50+ Years
Content Type	Total Segments	Segments	Segments	Segments	Segments	Segments	Segments
Contemporary Text (News)-T1	60	7	19	3	5	23	3
Creative Text-T2	61	8	19	3	5	23	3
Sentence-S	1527	200	476	75	125	576	75
Date-D	122	16	38	6	10	46	6
Command and Control Words-W1	1830	240	571	90	150	689	90
Person Name-W2	1222	160	382	59	100	461	60
Place Name-W2	609	80	189	30	50	230	30
Most Frequent Word-Part-W3A	1831	240	573	90	150	688	90
Most Frequent Word-FullSet-W3B	2000	0	1000	0	0	1000	0
Phonetically Balanced-W4	2050	0	1025	0	1025	0	0
Form and Function Word-W5	724	362	362	0	0	0	0

Table 26: Representation of Audio Segments of Dogri Raw Speech Data

3.4.2 Duration of the Dogri Raw Speech Data

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

LDC-IL Dogri Speech Data	Gender →	Female			Male		
	Age Group →	16-20 Years	21-50 Years	50+ Years	16-20 Years	21-50 Years	50+ Years
Content Type	Total Duration	Duration (hh:mm:ss)	Duration (hh:mm:ss)	Duration (hh:mm:ss)	Duration (hh:mm:ss)	Duration (hh:mm:ss)	Duration (hh:mm:ss)
Contemporary Text (News)-T1	4:27:51	00:27:36	01:18:21	00:13:49	00:30:00	01:41:10	00:16:55
Creative Text-T2	2:51:42	00:22:24	00:47:13	00:07:06	00:17:09	01:06:18	00:11:32
Sentence-S	1:24:48	00:10:14	00:24:30	00:03:47	00:10:47	00:31:14	00:04:16
Date-D	0:14:07	00:01:45	00:04:26	00:00:45	00:01:15	00:05:11	00:00:45
Command and Control Words-W1	1:24:31	00:10:07	00:24:28	00:04:04	00:09:06	00:32:46	00:04:00
Person Name-W2	1:23:41	00:11:11	00:25:40	00:03:56	00:07:09	00:31:13	00:04:32
Place Name-W2	0:29:10	00:03:41	00:09:40	00:01:18	00:02:34	00:10:31	00:01:26
Most Frequent Word-Part-W3A	1:18:06	00:10:19	00:23:53	00:04:08	00:07:09	00:28:46	00:03:51
Most Frequent Word-FullSet-W3B	1:16:27	00:00:00	00:38:01	00:00:00	00:00:00	00:38:26	00:00:00
Phonetically Balanced-W4	1:50:38	00:00:00	00:38:12	00:00:00	01:12:26	00:00:00	00:00:00
Form and Function Word-W5	0:29:25	00:15:19	00:14:06	00:00:00	00:00:00	00:00:00	00:00:00

Table 27: Representation of Dogri Raw Speech Data Duration

3.5 DISTINCT SET

The distinct set usually contains data which is distinct to each speaker and is rarely repeated. The LDC-IL speech data set contains newspaper extracts which are read by each speaker.

3.5.1 Contemporary Text (News)

Distinct text extracts from newspapers are recorded by the informants to get the Dogri speech data of contemporary text. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution Jammu	
		Female	Male
16 to 20	12	7	5
21 to 50	42	19	23
50+	6	3	3
Total	**60**	29	31

Table 28: Representation of Dogri Contemporary text (news)

3.6 RANDOM SET

The random set data comprises of content types which are sampled randomly for each speaker. They are sampled from collection of master datasets available. The random sets are given below:

3.6.1 Creative Text-T2

One randomly selected text of literature out of 525 texts from the prepared dogri dataset is recorded from the informants to get the speech data of Creative text. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution Jammu	
		Female	Male
16 to 20	13	8	5
21 to 50	42	19	23
50+	6	3	3
Total	61	30	31

Table 29: Representation of Dogri Creative Text

3.6.2 Sentences

The sentences content type contains a list of sentences that is a representation of almost all the phonemes occurring in Dogri. 25 Randomly selected Sentences are recorded from a list of 385 sentences. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution Jammu	
		Female	Male
16 to 20	325	200	125
21 to 50	1052	476	576
50+	150	75	75
Total	1527	751	776

Table 30: Representation of Dogri Sentences

3.6.3 Date Format

The answer of 2 questions is collected from each speaker to get the Dogri date format of the informants. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution Jammu	
		Female	Male
16 to 20	26	16	10
21 to 50	84	38	46
50+	12	6	6
Total	122	60	62

Table 31: Representation of Dogri Date formats

3.6.4 Command and Control Words

The content type of command and control words contains a list of 314 words that is a representation of almost all the command and control words occurring in Dogri. 30 randomly selected words of the list are recorded from each informant. Each word is uttered three times. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution Jammu	
		Female	Male
16 to 20	390	240	150
21 to 50	1260	571	689
50+	180	90	90
Total	1830	901	929

Table 32: Representation of Dogri Command and Control Words

3.6.5 Person Name

The person names include a list of 490 popular pan-Indian and regional person name. 20 randomly selected names are recorded from a list of names. Each name is uttered three times. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution Jammu	
		Female	Male
16 to 20	260	160	100
21 to 50	843	382	461
50+	119	59	60
Total	1222	601	621

Table 33: Representation of Dogri Person Names

3.6.6 Place Name

The place name contains a list of 370 popular pan- Indian and regional place name. 10 randomly selected names are recorded from a list of names. Each name is uttered three times. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution Jammu	
		Female	Male
16 to 20	130	80	50
21 to 50	419	189	230
50+	60	30	30
Total	609	299	310

Table 34: Representation of Dogri Place Names

3.6.7 Most Frequent Word-Part

The most frequent words-part contains a list of 1000 most frequent words occurring in Dogri. 30 randomly selected words of the list are recorded from each informant. Each word is uttered three times. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution Jammu	
		Female	Male
16 to 20	390	240	150
21 to 50	1261	573	688
50+	180	90	90
Total	1831	903	928

Table 35: Representation of Dogri Most Frequent Words-Part

3.7 FULL SETS

The full sets are the master set of certain data sets which are read completely from few selected speakers in each group. The full sets are given as below.

3.7.1 Most Frequent Word- Full

The most frequent words contain a list of 1000 most frequent words. In full set all the 1000 words are recorded from the informant. Each word is uttered three times. It was collected from one Male and one female of 21 to 50 Age group informant.

3.7.2 Phonetically Balanced Vocabulary

The phonetically balanced vocabulary contains a list of words where almost all the phonemes of Dogri language have occurred in all the possible positions of a word. In full set all the 1025 words is recorded from the informant where they uttered those words three times. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution Jammu	
		Female	Male
16 to 20	1025	0	1025
21 to 50	1025	1025	0
Total	**2050**	1025	1025

Table 36: Representation of Dogri Phonetically Balanced Vocabulary

3.7.3 Form and Function Word

The form and function words include a list of 362 words that is a representation of almost all the form and function words occurring in Dogri. All the words are recorded from the informant where they uttered those words three times. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution Jammu	
		Female	Male
16 to 20	362	362	0
21 to 50	362	362	0
Total	**724**	724	0

Table 37: Representation of Dogri Form and Function Wor

4 KASHMIRI RAW SPEECH CORPUS

Hilal Ahmad Dar, Zargar Adil Ahmad, Rajesha N., Manasa G., Narayan Choudhary

4.1 INTRODUCTION

Kashmiri Language belongs to the Dardic group of Indo-Aryan family. It is known by names 'Kashur' and 'Kashmiri'. It is primarily spoken in Kashmir valley and Pir-Panjal range of the Jammu region. As per the census report of 2011, the total number of speakers of Kashmiri language is 67,97,587. It is one of the modern spoken languages of India and the Constitution of India has recognized it as an official language under the Schedule VIII. Kashmiri has a history of varied literature.

The dialects of Kashmiri can be divided into types i.e., regional dialects and social dialects. Apart from the Kashmiri spoken in valley itself, there are other varieties of language that are spoken outside the valley and those varieties are considered as regional dialects of Kashmiri language. These regional dialects consist of Kishtawari, Poguli and Rambani. Kishtawari is spoken in Kishtawar and Doda districts of Jammu and Kashmir state, while Poguli and Rambani are spoken in Pogal Paristan and Ramban areas of the state, respectively. Kashmiri language has three social dialects as well which are known by the names Yamraz, Marak and Kamraz. Yamraz dialect is spoken in the central part of Kashmir valley and it is considered as the standard variety of the language. The central part of valley consists of Srinagar, Budgam and Ganderbal districts. Maraz dialect is spoken in the southern part of the valley. The southern part comprises of Anantnag, Pulwama, Kulgam and Shopian districts. Kamraz dialect is spoken in northern part of the valley which comprises of Baramulla, Kupwara and Bandipora districts. Some minor variations among these varieties have been observed. Most of the variations are phonological in nature. Some of the variations found in these varieties are as follows:

1. Kashmiri spoken in Maraz area usually use /ɽ/ in the final position of a word which is replaced by /r/ in Kashmiri spoken in Srinagar. For example, /guD/ گُڑ (horse) in Maraz area is pronounced as /gur/ گُرin Yamraz and Kamraz areas.
2. The suffix -an is added to the progressive or Indefinite aspect of verb roots in Kashmiri spoken in Maraz, which is replaced by -a:n in the other two varieties. For example, /pakan/ پکن(walking) becomes /paka:n/ پکان in the other two varieties.
3. Kamraz variety of Kashmiri is distinguished from the variety spoken in Maraz as well as Srinagar mainly in the use of intonation and stress.
4. There are some variations in vocabulary items as well in these three different speech varieties of the language, for example, wadun/ (weeping) in Maraz and Yamaraz varieties is /bresun/ in Kamraz dialect. /vızı/ وِزِ (corridor of a house) in Maraz is /voɽ/ وٹ in Yamraz and Kamraz dialects.

Another variation of Kashmiri may be identified by in the Kashmiri being spoken by the Kashmiri Hindus / Kashmiri Pandits. This variety will have more of its vocabulary drawn from Sanskrit words instead of Perso-Arabic as is the case with the speech variety of Muslim community across the valley.

Kashmiri is written both in Perso-Arabic as well as Devanagari. While in the valley it is written mostly in Perso-Arabic script, the practice by the Hindu community is more towards Devanagari. It is note here that Kashmiri was earlier written in Sharada script which is no longer in use. For the current corpus, the script used is Perso-Arabic. However, the same data is available also in Devanagari and can be used simultaneously.

Grierson has classified Kashmiri under the Dardic group of languages. As per his classification Dardic languages consists of three major groups: 1. The Kafir group, 2. The Khowar or Chitrali group, and 3. The Dard group. According to Grierson, "the Dard Group includes Shina, Kashmiri and Kohistani." He considers the Dardic languages to be sub-family of the Aryan languages "neither of Indian nor of Iranian origin, but forming a third branch of Aryan stock, which separated from the parent stem after the branching forth of original of the Indian languages, but before the Iranian languages had developed all their peculiar characteristics" (1906:4). Grierson has further observed that 'Dardic' is only a geographical convention. Morgenstierne (1961) has placed Kashmiri under the Dardic Group of Indo-Aryan languages along with Kishtawari and other dialects which are strongly influenced by Dogri. Fussman (1972) has based his work on Morgenstierne's classification. He has also emphasized that the Dardic is a geographical and not a linguistic expression. It is not in the absence of reliable comparative data about Dardic languages, a geographic or ethnographic label 'Dardic' is frequently used to identity a group of languages or dialects. According to Chaterjee (1963:256) Kashmiri is one of the main languages of Indo Aryan group as Kashmiri has developed like other Indo-Aryan languages out of the Indo-European family like Hindi, Punjabi etc. Other scholars have agreed upon this opinion as well.

LDC-IL collected speech data from Kashmir Valley of the Jammu and Kashmir state. After determining the regions for fieldwork, the dataset is prepared from which the prompt sheets were generated. Places from which LDC-IL Kashmiri Speech Data is collected in the Kashmir Valley are Pulwama, Srinagar, and Anantnag.

4.2 DATASET PREPARATION FOR KASHMIRI

For the different regions of the Kashmir valley, LDC-IL prepared the following dataset by which the prompt sheets were prepared.

Content Type	Count
Created Text	6
Date	2
Command and Control Words	201
Most Frequent Words	1000
Person Name	199
Place Name	300
Sentences	201

Table 38: Representation of Kashmiri Content Type

Distinct News Items were prepared do get the audio recording of contemporary text. It was made sure that each selected news item had minimum 500 words. Each prompt sheet had a distinct news item and selected part of the dataset prepared as follows.

Content Type	Content that Each typical prompt sheet had	Content selection type
Contemporary Text	1 Text	Distinct Text
Creative Text	1 text	Random Text selected from dataset*
Sentences	25 Sentences	Random set selected from dataset*
Command and Control Words	30 Words	Random set selected from dataset*
Person Names	20 Words	Random set selected from dataset*
Place Names	10 Words	Random set selected from dataset*
Most Frequent Words	30 Words	Random set selected from dataset*
		*randomly selected by machine

Table 39: Representation of Kashmiri Prompt Sheet

Once all these preparations were made, the investigator started collecting the data.The collection of data is carried out in three phases for different regions of Kashmir Valley in year 2010 by Shahid Mushtaq Bhat.

4.3 SUMMARY OF THE CORPUS

In the sections below, we provide the tabular details of the different content types of the Kashmiri raw speech corpus based on various yardsticks which can also be filter out from the metadata sheet. 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 data size is of total duration 28:10:07(hh:mm:ss) comprising 16,380 audio segments. We hope to augment the speech data of Kashmiri in near future.

4.4 SUMMARY OF THE AUDIO SEGMENTS

The table below shows the total number of Audio Segments and their distribution in the Kashmiri speech dataset.

LDC-IL Kashmiri Speech Data Status	Gender →	Female			Male		
	Age Group →	16-20 Years	21-50 Years	50+ Years	Below 15 Years	21-50 Years	50+ Years
Content Type	Total Segments	Segments	Segments	Segments	Segments	Segments	Segments
Contemporary Text (News)-T1	147	16	54	6	1	64	6
Creative Text-T2	148	16	55	6	1	64	6
Sentence-S	3704	400	1373	147	25	1608	151
Date-D	281	30	99	12	2	126	12
Command and Control Words-W1	4288	476	1617	164	30	1825	176
Person Name-W2	2065	224	765	82	14	896	84
Place Name-W2	1468	160	548	50	10	640	60
Most Frequent Word-Part-W3A	4279	479	1621	178	30	1793	178

Table 40: Audio Segments and their Distribution

4.5 DURATION OF THE RAW SPEECH DATA

The table below shows the duration of each of the content type and their distribution across few factors.

LDC-IL Kashmiri Speech Data Status	Gender →	Female			Male		
	Age Group →	16-20 Years	21-50 Years	50+ Years	Below 15 Years	21-50 Years	50+ Years
Content Type	Total Duration	Duration (hh:mm:ss)	Duration (hh:mm:ss)	Duration (hh:mm:ss)	Duration (hh:mm:ss)	Duration (hh:mm:ss)	Duration (hh:mm:ss)
Contemporary Text (News) -T1	3:56:57	0:22:02	1:27:11	0:08:30	0:00:46	1:42:53	0:15:35
Creative Text-T2	12:41:33	1:17:51	4:36:55	0:28:27	0:06:17	5:38:35	0:33:28
Sentence-S	2:40:24	0:18:04	0:58:27	0:06:32	0:01:11	1:08:52	0:07:18
Date-D	0:10:36	0:01:03	0:03:39	0:00:30	0:00:06	0:04:46	0:00:32
Command and Control Words-W1	3:04:32	0:19:41	1:05:25	0:06:33	0:01:05	1:22:40	0:09:08
Person Name-W2	1:53:21	0:11:56	0:41:56	0:04:34	0:00:48	0:48:50	0:05:17
Place Name-W2	1:04:37	0:06:46	0:23:29	0:02:08	0:00:28	0:28:45	0:03:01
Most Frequent Word-Part-W3A	2:38:07	0:17:19	0:57:27	0:06:05	0:00:58	1:08:17	0:08:01

Table 41: Duration of Kashmiri Speech Data

4.6 DISTINCT SET

The Distinct Set usually contains data which is distinct to each speaker and is rarely repeated. The LDC-IL speech data set contains newspaper extracts which are read by each speaker.

4.6.1 Contemporary Text (News)

Distinct Text Extracts from newspapers are recorded from the informants to get the speech data of contemporary text. The distribution of data is as follows.

Age Group	Total Audio Segments	Kashmir Valley Region	
		Female	**Male**
Below 16	1	0	1
16 To 20	16	16	0
21 To 50	118	54	64
50+	12	6	6
Total	147	76	71

Table 42: Distribution of Kashmiri Contemporary Text (News) Data

4.7 RANDOM SET

The Random Set data comprises of content types which are sampled by machine for each speaker. They are sampled from collection of master data sets available. The random sets are given below.

4.7.1 The Creative Text-T2

One randomly selected text of literature out of 6 texts from the prepared dataset is recorded from the informants to get the speech data of Creative text. The distribution of data is as follows.

Age Group	Total Audio Segments	Kashmir Valley Region	
		Female	**Male**
Below 16	1	0	1
16 To 20	16	16	0
21 To 50	118	54	64
50+	12	6	6
Total	147	76	71

Table 43: Distribution of Kashmiri Createive Text

4.7.2 The Date-D

The answer to one randomly selected question from the list of 2 questions to get the date format of the informants. The distribution of data is as follows.

Age Group	Total Audio Segments	Kashmir Valley Region	
		Female	**Male**
Below 16	2	0	2
16 To 20	30	30	0
21 To 50	225	99	126
50+	24	12	12
Total	281	141	140

Table 44: Distribution of Kashmiri Date Format

4.7.3 The Sentences-S

The Sentences contain a list of sentences that is a representation of almost all the phonemes occurring in Kashmiri. 25 Randomly selected sentences are recorded from a list of 201sentences. The distribution of data is as follows.

Age Group	Total Audio Segments	Kashmir Valley Region	
		Female	**Male**
Below 16	25	0	25
16 To 20	400	400	0
21 To 50	2981	1373	1608
50+	298	147	151
Total	3704	1920	1784

Table 45: Distribution of Kashmiri Sentences

4.7.4 Command and Control Words-W1

The Command and Control Words contain a list of 201 words that is a representation of almost all the command and control words occurring in Kashmiri. 30 randomly selected words is recorded from a list of words. The distribution of data is as follows.

Age Group	Total Audio Segments	Kashmir Valley Region	
		Female	**Male**
Below 16	30	0	30
16 To 20	476	476	0
21 To 50	3442	1617	1825
50+	340	164	176
Total	4288	2257	2031

Table 46: Distribution of Command and Control Words

4.7.5 Person Names –W2

The Person Names contain a list of 199 popular pan- Indian and regional Person names. 14 randomly selected names are recorded from the list. The distribution of data is as follows.

Age Group	Total Audio Segments	Kashmir Valley Region	
		Female	**Male**
Below 16	14	0	14
16 To 20	224	224	0
21 To 50	1661	765	896
50+	166	82	84
Total	2065	1071	994

Table 47: Distribution of Kashmiri Person Names

4.7.6 Place Names-W2

The Place Names contain a list of 300 popular pan- Indian and regional Place names. 10 randomly selected names are recorded from the list. The distribution of data is as follows.

Age Group	Total Audio Segments	Kashmir Valley Region	
		Female	**Male**
Below 16	10	0	10
16 To 20	160	160	0
21 To 50	1188	548	640
50+	110	50	60
Total	1468	758	710

Table 48: Distribution of Place Names

4.7.7 Most Frequent Words-PART-W3A

The Most Frequent Words-part contains a list of 1000 most frequent words. 30 randomly selected words are recorded from the list. The distribution of data is as follows.

Age Group	Total Audio Segments	Kashmir Valley Region	
		Female	**Male**
Below 16	30	0	30
16 To 20	479	479	0
21 To 50	3414	1621	1793
50+	356	178	178
Total	4279	2278	2001

Table 49: Distribution of Kashmiri Most Frequent Words Part

4.8 NATIVE SPEAKER DISTRIBUTIONS

The distribution of Native Speakers across the regional dialect in LDC-IL Kashmiri Speech corpus is as follows.

Region-wise Distribution of Native Speakers			
Age Group	Total Native Speakers	Native Speakers	
		Female	**Male**
Below 16	1	0	1
16 To 20	16	16	0
21 To 50	121	56	65
50+	12	6	6
Total	150	78	72

Table 50: Distribution of Native Speakers

REFERENCES

1. Chatterjee, S.K.(1954) *Kashmiri language and literature, Kashmir*, IV,, pp. 75-78; For debate on the origin of Kashmiri language See; P. N. K Bamazi, op. cit, pp. 577; G.M. D Sufi, op.cit, pp. 176-181.

2. Grierson, George A. (1910). *Indo-Aryan Family, North Western group: Specimens of the Dardic or Pisaca languages (including Kashmiri).* Linguistic survey of India, vol. 8, Part 2. Calcutta, Reprinted Delhi :MotitalBanarasidas, 1968. 567.

3. Kachru, Braj B. (1969). *Kashmiri and other Dardic languages*. Sebeok, Thomas A. (ED), Current trends in linguistic, vol. 5, pp. 284-306.

4. Koul, Omkar N. and Ruth Laila Schmidt (1984). *Dardistan revised: An examination of relationship between Kashmiri and Shina.* In Koul, Omkar N. and Peter E. Hook (eds.) , Aspects of Kashmiri linguistics. New Delhi: Bahri Publications, pp. 1-2

5 GUJARATI RAW SPEECH CORPUS (MONO RECORDINGS)

Rejitha K.S., Rajesha N., Manasa G., Narayan Choudhary

5.1 INTRODUCTION

Gujarati is one of the major literary languages of India and is the official language of the state of Gujarat and the union territories of Daman and Diu and that of Dadra and Nagar Haveli. Gujarati has developed from "Shaurseni" or "Nagara Apabhramsha" and it is a part of Indo-Aryan language family. Gujarati script is cursive form of Devanagari.

Indo-Aryan speech community was divided into three major groups - Northern, Central and Eastern. The areas which are occupied today by Gujarati, Marwari, Mewati, Jaipuri, Bhili and adjacent dialects were marked off from the central linguistic area. Gradually Gujarati separated from other dialects and achieved the status of a language. This separation occurred around 1200 A.D.

Regional dialect and social dialects are the two types of dialects which linguists derive traditionally. Language change happens through spatial, temporal, and social factors. Sharp boundaries of Gujarati dialect cannot be marked because socio-economic factors, such as caste, ethnicity, education, occupation, social status etc, are overlapped. People of different social classes, occupations or cultural groups in the same community will show variations in speech.

Based on the linguistic features, Gujarati can be separated into three major dialects namely Northern, Central and Southern dialects. The central dialect is considered as the standard dialect of Gujarati. The Northern dialect covers the region between Banas and Sabarmati. The Central dialect covers the region between Sabarmati and Narmada and Southern dialect covers the region beyond Narmada. Moreover, there are some social and occupational dialects. Based on the caste, occupation, social status etc. people use some specialized vocabulary of their own. Tribes of Central and Southern Gujarat, Mer community of Saurashtra and Parsis and Khojas have their own linguistic features and they are preserving some community specific vocabulary items.

For the convenience, LDC-IL considered Gujarati with four dialects namely, South Gujarat, Central Gujarat, North Gujarat, and Saurashtra. The collection of Gujarati Speech Data (Mono Recordings) was carried out by Mona Parekh in 2008 and 2009. LDC-IL published another dataset named "Gujarati Raw Speech Corpus" which is Raw speech corpus of Gujarati, speech in stereo recording that has mutually exclusive speakers from this dataset.

5.2 DATASET PREPARATION FOR GUJARATI (MONO)

LDC-IL prepared the following dataset by which the prompt sheets were prepared.

Content Type	Count
Creative Text	8
Date	2
Command and Control Words	296
Most Frequent Words	1,000
Form and Function Words	232
Phonetically Balanced Words	689
Person Name	543
Place Name	359
Sentences	200

Table 51: LDC-IL Gujarati (Mono) Speech Dataset

Distinct News Items were prepared to get the audio recordings of contemporary text. It was made sure that each selected news item had minimum 500 words. Each prompt sheet had a distinct news item and sleeted part of the dataset prepared as follows:

Content Type	Content that Each Typical Prompt Sheet had	Content Selection Type
Contemporary Text	1 Text	Distinct Text
Creative Text	1 text	Random Text selected from dataset*
Sentences	25 Sentences	Random set selected from dataset*
Command and Control Words	30 Words	Random set selected from dataset*
Person Names	20 Words	Random set selected from dataset*
Place Names	10 Words	Random set selected from dataset*
Most Frequent Words	30 Words	Random set selected from dataset*
		*randomly selected by machine

Table 52: Table of Contents in LDC-IL Dataset

The Full Set of
1. Phonetically Balanced Vocabulary
2. Form and Function Words
3. 1000 Most Frequent Words
were also carried to the field to get recorded by selected individuals. Once all these preparations were made, the investigator started collecting the data.

5.3 TRANSLITERATIONS IN LDC-IL GUJARATI READ CORPUS

For easy reference and uniformity, the recorded text in the metadata file, is transliterated from Gujarati to Roman letters. Numeric characters were transliterated from Gujarati to Hindu-Arabic system.

The LDC-IL transliteration scheme of Gujarati to Roman is given below.

LDC-IL Transliteration Schema
Gujarati to Roman and Gujarati Numerals to Hindu-Arabic

Vowels and Vowel Signs										
અ	આ	ઇ	ઈ	ઉ	ઊ	ઋ	એ	ઐ	ઓ	ઔ
	ા	િ	ી	ુ	ૂ	ૃ	ે	ૈ	ો	ૌ
a	A	i	I	u	U	x	e	ai	o	au

Consonants				
ક	ખ	ગ	ઘ	ઙ
ka	kha	ga	gha	ng'a
ચ	છ	જ	ઝ	ઞ
ca	cha	ja	jha	nj'a
ટ	ઠ	ડ	ઢ	ણ
Ta	Tha	Da	Dha	Na
ત	થ	દ	ધ	ન
ta	tha	da	dha	na
પ	ફ	બ	ભ	મ
pa	pha	ba	bha	ma

Symbols		
ં	ઃ	ઁ
M	H	m'

ય	ર	ળ	લ	વ	શ	ષ	સ	હ
ya	ra	la	La	va	sha	Sa	sa	ha

Numerals (Gujarati to Hindu-Arabic)									
૦	૧	૨	૩	૪	૫	૬	૭	૮	૯
0	1	2	3	4	5	6	7	8	9

5.4 SUMMARY OF THE CORPUS

In the sections below, we provided the tabular details of the different content types of the Gujarati Raw Speech Corpus (Mono Recordings) based on various yardsticks which can also be filtered out from the metadata sheet. 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 data size is of total duration 64:44:02 (hh:mm:ss) comprising 26,223 audio segments.

5.5 SUMMARY OF THE AUDIO SEGMENTS

The table below shows the total number of Audio Segments and their distribution in the Gujarati Raw Speech Corpus (Mono Recordings)

LDC-IL Gujarati Raw Speech Corpus (Mono Recordings)	Gender →	Female			Male		
	Age Group →	16-20 Years	21-50 Years	50+ Years	16-20 Years	21-50 Years	50+ Years
Content Type	Total Segments	Segments	Segments	Segments	Segments	Segments	Segments
Contemporary Text (News)-T1	233	22	79	23	16	66	27
Creative Text-T2	232	22	79	23	16	66	26
Sentence-S	5824	550	1975	575	400	1650	674
Date-D	466	44	158	46	32	132	54
Command and Control Words-W1	6985	660	2367	690	480	1980	808
Person Name-W2	4644	440	1582	461	320	1301	540
Place Name-W2	2322	220	790	230	162	650	270
Phonetically Balanced-W4	4131	689	689	689	688	689	687
Form and Function Word-W5	1386	230	232	231	232	229	232

Table 53: Audio Segments and their Distribution

5.6 DURATION OF THE RAW SPEECH DATA (MONO RECORDINGS)

The table below shows the duration of each of the content type and their distribution across a few factors.

LDC-IL Gujarati Raw Speech Corpus (Mono Recordings)	Gender →	Female			Male		
	Age Group →	16-20 Years	21-50 Years	50+ Years	16-20 Years	21-50 Years	50+ Years
Content Type	Total Duration	Duration (hh:mm:ss)	Duration (hh:mm:ss)	Duration (hh:mm:ss)	Duration (hh:mm:ss)	Duration (hh:mm:ss)	Duration (hh:mm:ss)
Contemporary Text (News)-T1	12:52:46	1:19:26	4:37:23	1:14:07	0:44:48	3:19:57	1:37:05
Creative Text-T2	13:30:15	1:09:50	4:43:31	1:22:45	0:53:56	3:48:13	1:32:00
Sentence-S	7:12:17	0:41:31	2:26:23	0:43:20	0:26:11	2:04:14	0:50:38
Date-D	0:59:31	0:05:53	0:19:18	0:06:14	0:03:34	0:17:21	0:07:11
Command and Control Words-W1	9:43:07	1:01:20	3:13:30	1:06:39	0:33:45	2:42:42	1:05:11
Place Name-W2	3:17:06	0:21:43	1:03:56	0:21:55	0:12:09	0:55:49	0:21:34
Person Name-W2	8:34:44	0:56:52	2:49:10	0:57:29	0:29:42	2:24:24	0:57:07
Phonetically Balanced Vocabulary-W4	6:28:15	1:04:21	1:28:32	1:04:41	0:44:00	0:59:08	1:07:33
Form and Function Word-W5	2:06:01	0:28:54	0:15:23	0:23:26	0:16:56	0:20:28	0:20:54

Table 54: Duration of the Collected Data

5.7 DISTINCT SET

The Distinct Set usually contains data which is distinct to each speaker and is rarely repeated. The LDC-IL speech data set contains newspaper extracts which are read by each speaker.

5.7.1 The Contemporary Text (News) - T1

Distinct Text Extracts from Newspapers are recorded from the informants to get the speech data of contemporary text. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution			
				South Gujarat		Central Gujarat	
		Female	Male	Female	Male	Female	Male
16 To 20	**38**	22	16	7	8	15	8
21 To 50	**145**	79	66	35	20	44	46
50+	**50**	23	27	5	10	18	17
Total	**233**	**124**	**109**	**47**	**38**	**77**	**71**

Table 55: Distribution of Gujarati (Mono) Contemporary Text (News) Data

5.8 RANDOM SET

The Random Set data comprises of content types which are sampled by machine for each speaker. They are sampled from collection of master data sets available. The random sets are given below.

5.8.1 The Creative Text-T2

One randomly selected text of literature out of 8 texts from the prepared dataset is recorded from the informants to get the speech data of Creative text. The distribution of data is as follows.

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution			
				South Gujarat		Central Gujarat	
		Female	Male	Female	Male	Female	Male
16 To 20	**38**	22	16	7	8	15	8
21 To 50	**145**	79	66	35	20	44	46
50+	**49**	23	26	5	9	18	17
Total	**232**	**124**	**108**	**47**	**37**	**77**	**71**

Table 56: Distribution of Gujarati (Mono) Creative Text

5.8.2 The Date-D

The answer to one randomly selected question from the list of 2 questions to get the date format of the informants. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution			
				South Gujarat		Central Gujarat	
		Female	Male	Female	Male	Female	Male
16 To 20	**76**	44	32	14	16	30	16
21 To 50	**290**	158	132	70	40	88	92
50+	**100**	46	54	10	20	36	34
Total	**466**	**248**	**218**	**94**	**76**	**154**	**142**

Table 57: Distribution of Gujarati (Mono) Date Format

5.8.3 The Sentences-S

The Sentences contain a list of sentences that is a representation of all most all the phonemes occurring in Gujarati. 25 Randomly selected Sentences are recorded from a list of 200 sentences. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution			
				South Gujarat		Central Gujarat	
		Female	Male	Female	Male	Female	Male
16 To 20	**950**	550	400	175	200	375	200
21 To 50	**3625**	1975	1650	875	500	1100	1150
50+	**1249**	575	674	125	249	450	425
Total	**5824**	**3100**	**2724**	**1175**	**949**	**1925**	**1775**

Table 58: Distribution of Gujarati (Mono) Sentences

5.8.4 Command and Control Words-W1

The Command and Control Words contain a list of 296 words that is a representation of all most all the command and control words occurring in Gujarati. 30 randomly selected words are recorded from the list. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution			
				South Gujarat		Central Gujarat	
		Female	Male	Female	Male	Female	Male
16 To 20	**1140**	660	480	210	240	450	240
21 To 50	**4347**	2367	1980	1049	600	1318	1380
50+	**1498**	690	808	150	299	540	509
Total	**6985**	**3717**	**3268**	**1409**	**1139**	**2308**	**2129**

Table 59: Distribution of Gujarati (Mono) Command and Control Words

5.8.5 Person Names –W2

The Person Names contain a list of 543 popular pan- Indian and regional person names. 20 randomly selected names are recorded from the list. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution			
				South Gujarat		Central Gujarat	
		Female	Male	Female	Male	Female	Male
16 To 20	**760**	440	320	140	160	300	160
21 To 50	**2883**	1582	1301	702	400	880	901
50+	**1001**	461	540	102	200	359	340
Total	**4644**	**2483**	**2161**	**944**	**760**	**1539**	**1401**

Table 60: Distribution of Gujarati (Mono) Person Names

5.8.6 Place Names-W2

The Place Names contain a list of 359 popular pan- Indian and regional place names. 10 randomly selected names are recorded from the list. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution			
				South Gujarat		Central Gujarat	
		Female	Male	Female	Male	Female	Male
16 To 20	**382**	220	162	70	80	150	82
21 To 50	**1440**	790	650	350	200	440	450
50+	**500**	230	270	50	100	180	170
Total	**2322**	**1240**	**1082**	**470**	**380**	**770**	**702**

Table 61: Distribution of Gujarati (Mono) Place Names

5.9 FULL SET

The Full Set is the master set of certain data set which is red completely from few selected speakers in each group. Full sets are as below.

5.9.1 The Phonetically Balanced Vocabulary-W4

The Phonetically Balanced Vocabulary contains a list of words where all most all the phones of Gujarati have occurred in all the possible positions of a word. In full set all the 689 words are recorded from the informant where they uttered those words three times. The distribution of data is as follows:

Age Group	Total Audio Segments	Region-wise Distribution	
		Central Gujarat	
		Female	Male
16 To 20	**1377**	689	688
21 To 50	**1378**	689	689
50+	**1376**	689	687
Total	**4131**	**2067**	**2064**

Table 62: Distribution of Gujarati (Mono) Phonetically Balanced Vocabulary

5.9.2 The Form and Function Words-W5

The Form and Function Words contain a list of 232 words which is a representation of all most all the form and function words occurring in Indian English. All the words are recorded from the informant where they uttered those words three times. The distribution of data is as follows:

Age Group	Total Audio Segments	Region-wise Distribution	
		Central Gujarat	
		Female	Male
16 To 20	**462**	230	232
21 To 50	**461**	232	229
50+	**463**	231	232
Total	**1386**	**693**	**693**

Table 63: Distribution of Gujarati (Mono) Form and Function words

5.10 NATIVE SPEAKER DISTRIBUTIONS

For Gujarati Raw Speech Corpus (Mono Recordings) a total of 235 speakers were collected in which 124 female speakers and 111 male speakers from two different regions. The distribution of data is as follows:

Region-wise Distribution of Native Speakers							
Age Group	Total Native Speakers	Gender-wise Distribution of Native Speakers		Regions			
				South Gujarat		Central Gujarat	
		Female	Male	Female	Male	Female	Male
16 To 20	**40**	22	18	7	8	15	10
21 To 50	**145**	79	66	35	20	44	46
50+	**50**	23	27	5	10	18	17
Total	**235**	**124**	**111**	**47**	**38**	**77**	**73**

Table 64: Distribution of Gujarati (Mono) Native Speaker

6 GUJARATI RAW SPEECH CORPUS

Rejitha K.S., Rajesha N., Manasa G., Narayan Choudhary

6.1 INTRODUCTION

Gujarati is one of the major literary languages of India and is the official language of the state of Gujarat and the union territories of Daman and Diu and that of Dadra and Nagar Haveli. Gujarati has developed from "Shaurseni" or "Nagara Apabhramsha" and it is a part of Indo-Aryan language family. Gujarati script is cursive form of Devanagari.

Indo-Aryan speech community was divided into three major groups - Northern, Central and Eastern. The areas which are occupied today by Gujarati, Marwari, Mewati, Jaipuri, Bhili and adjacent dialects were marked off from the central linguistic area. Gradually Gujarati separated from other dialects and achieved the status of a language. This separation occurred around 1200 A.D.

Regional dialect and social dialects are the two types of dialects which linguists derive traditionally. Language change happens through spatial, temporal, and social factors. Sharp boundaries of Gujarati dialect cannot be marked because of various socio-economic factors, such as caste, ethnicity, education, occupation, social status etc, are overlapped. People of different social classes, occupations or cultural groups in the same community will show variations in speech.

Based on the linguistic features, Gujarati can be separated into three major dialects namely Northern, Central and Southern dialects. The central dialect is considered as the standard dialect of Gujarati. The Northern dialect covers the region between Banas and Sabarmati. The Central dialect covers the region between Sabarmati and Narmada and Southern dialect covers the region beyond Narmada. Moreover, there are some social and occupational dialects. Based on the caste, occupation, social status etc. people use some specialized vocabulary of their own. Tribes of Central and Southern Gujarat, Mer community of Saurashtra and Parsis and Khojas have their own linguistic features and they are preserving some vocabulary which is related with their community and culture.

For the convenience, LDC-IL considered Gujarati with four dialects namely South Gujarat, Central Gujarat, North Gujart and Saurashtra. The collection of Gujarati Speech data was carried out by Hiren Gadhavi and Purva Dolakia in 2010 and 2012, respectively. LDC-IL published another dataset "Gujarati Raw Speech Corpus (Mono Recordings)" where the speech is in mono recording and has mutually exclusive speakers.

6.2 DATASET PREPARATION FOR GUJARATI

LDC-IL prepared the following dataset by which the prompt sheets were prepared.

Content Type	Count
Creative Text	8
Date	2
Command and Control Words	296
Most Frequent Words	1,000
Form and Function Words	232
Phonetically Balanced Words	689
Person Name	543
Place Name	359
Sentences	200

Table 65: LDC-IL Gujarati Speech Dataset

Distinct News Items were prepared to get the audio recording of contemporary text. It was made sure that each selected news item had minimum 500 words. Each prompt sheet had a distinct news item and sleeted part of the dataset prepared as follows.

Content Type	Content that Each Typical Prompt Sheet had	Content Selection Type
Contemporary Text	1 Text	Distinct Text
Creative Text	1 text	Random Text selected from dataset*
Sentences	25 Sentences	Random set selected from dataset*
Command and Control Words	30 Words	Random set selected from dataset*
Person Names	20 Words	Random set selected from dataset*
Place Names	10 Words	Random set selected from dataset*
Most Frequent Words	30 Words	Random set selected from dataset*
		*randomly selected by machine

Table 66: Table of Contents in LDC-IL Dataset

The Full Set of
1. Phonetically Balanced Vocabulary
2. Form and Function Words
3. 1000 Most Frequent Words
were also carried to the field to get recorded by selected individuals. Once all these preparations were made, the investigator started collecting the data.

6.3 TRANSLITERATIONS IN LDC-IL GUJARATI READ CORPUS

For easy reference and uniformity, the recorded text in the metadata file, is transliterated from Gujarati to Roman letters. Numeric characters were transliterated from Gujarati to Hindu-Arabic system.

The LDC-IL transliteration scheme of Gujarati to Roman is given below.

LDC-IL Transliteration Schema
Gujarati to Roman and Gujarati Numerals to Hindu-Arabic

Vowels and Vowel Signs										
અ	આ	ઇ	ઈ	ઉ	ઊ	ઋ	એ	ઐ	ઓ	ઔ
	ા	િ	ી	ુ	ૂ	ૃ	ે	ૈ	ો	ૌ
a	A	i	I	u	U	x	e	ai	o	au

Consonants				
ક	ખ	ગ	ઘ	ઙ
ka	kha	ga	gha	ng'a
ચ	છ	જ	ઝ	ઞ
ca	cha	ja	jha	nj'a
ટ	ઠ	ડ	ઢ	ણ
Ta	Tha	Da	Dha	Na
ત	થ	દ	ધ	ન
ta	tha	da	dha	na
પ	ફ	બ	ભ	મ
pa	pha	ba	bha	ma

Symbols		
ં	ઃ	ઁ
M	H	m'

ય	ર	ળ	લ	વ	શ	ષ	સ	હ
ya	ra	la	La	va	sha	Sa	sa	ha

Numerals (Gujarati to Hindu-Arabic)									
૦	૧	૨	૩	૪	૫	૬	૭	૮	૯
0	1	2	3	4	5	6	7	8	9

6.4 SUMMARY OF THE CORPUS

In the sections below, we provide the tabular details of the different content types of the Gujarati raw speech corpus based on various yardsticks which can also be filter out from the metadata sheet. 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 data size is of total duration 57:17:08 (hh:mm:ss) comprising 25,712 audio segments.

6.5 SUMMARY OF THE AUDIO SEGMENTS

The table below shows the total number of Audio Segments and their distribution in the Gujarati raw speech dataset.

LDC-IL Gujarati Raw Speech	Gender →	Female			Male		
	Age Group →	16-20 Years	21-50 Years	50+ Years	16-20 Years	21-50 Years	50+ Years
Content Type	Total Segments	Segments	Segments	Segments	Segments	Segments	Segments
Contemporary Text (News)-T1	204	8	53	35	8	66	34
Creative Text-T2	202	8	54	35	8	63	34
Sentence-S	5081	200	1331	876	202	1623	849
Date-D	404	16	104	70	16	130	68
Command and Control Words-W1	6006	239	1509	1049	241	1948	1020
Person Name-W2	4079	160	1079	700	162	1298	680
Place Name-W2	2041	80	542	350	80	649	340
Most Frequent Word-Part-W3A	4236	240	1251	617	240	1139	749
Most Frequent Word-FullSet-W3B	2000	0	0	1000	0	0	1000
Phonetically Balanced-W4	1378	0	0	689	0	0	689

Table 67: Audio Segments and their Distribution

6.6 DURATION OF THE RAW SPEECH DATA

The table below shows the duration of each of the content type and their distribution across a few factors.

LDC-ILGujarati Raw Speech	Gender →	Female			Male		
	Age Group →	16-20 Years	21-50 Years	50+ Years	16-20 Years	21-50 Years	50+ Years
Content Type	Total Duration	Duration (hh:mm:ss)	Duration (hh:mm:ss)	Duration (hh:mm:ss)	Duration (hh:mm:ss)	Duration (hh:mm:ss)	Duration (hh:mm:ss)
Contemporary Text (News)-T1	15:21:28	0:32:07	4:10:41	2:32:01	0:35:44	4:56:08	2:34:47
Creative Text-T2	11:34:29	0:29:35	2:57:18	1:55:51	0:29:41	3:41:06	2:00:58
Sentence-S	5:48:32	0:14:58	1:29:32	1:00:21	0:15:38	1:48:07	0:59:56
Date-D	0:41:39	0:01:41	0:10:23	0:07:14	0:01:36	0:13:29	0:07:16
Command and Control Words-W1	7:17:22	0:23:33	1:47:43	1:10:46	0:28:04	2:11:18	1:15:58
Place Name-W2	2:33:20	0:08:43	0:39:22	0:24:25	0:09:22	0:44:49	0:26:39
Person Name-W2	6:36:02	0:20:39	1:41:06	1:05:40	0:22:24	1:56:27	1:09:46
Most Frequent Word-Part -W3A	5:18:47	0:24:28	1:27:36	0:44:41	0:27:20	1:19:51	0:54:51
Most Frequent Word-FullSet-W3B	1:13:39	0:00:00	0:00:00	0:39:41	0:00:00	0:00:00	0:33:58
Phonetically Balanced-W4	0:51:50	0:00:00	0:00:00	0:26:30	0:00:00	0:00:00	0:25:20

Table 68: Duration of the Collected Data

6.7 DISTINCT SET

The Distinct Set usually contains data which is distinct to each speaker and is rarely repeated. The LDC-IL speech data set contains newspaper extracts which are read by each speaker.

6.7.1 The Contemporary Text (News) - T1

Distinct Text Extracts from Newspapers are recorded from the informants to get the speech data of contemporary text. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution					
				Central Gujarat		South Gujarat		Saurashtra	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	**16**	8	8	0	0	8	8	0	0
21 To 50	**119**	53	66	1	1	42	38	10	27
50+	**69**	35	34	0	1	21	26	14	7
Total	**204**	**96**	**108**	**1**	**2**	**71**	**72**	**24**	**34**

Table 69: Distribution of Gujarati Contemporary Text (News) Data

6.8 RANDOM SET

The Random Set data comprises of content types which are sampled by machine for each speaker. They are sampled from collection of master data sets available. The random sets are given below.

6.8.1 The Creative Text-T2

One randomly selected text of literature out of 8 texts from the prepared dataset is recorded from the informants to get the speech data of Creative text. The distribution of data is as follows.

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution					
				Central Gujarat		South Gujarat		Saurashtra	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	**16**	8	8	0	0	0	0	8	8
21 To 50	**117**	54	63	1	1	10	27	43	35
50+	**69**	35	34	0	1	14	7	21	26
Total	**202**	**97**	**105**	**1**	**2**	**24**	**34**	**72**	**69**

Table 70: Distribution of Gujarati Creative Text

6.8.2 The Date-D

The answer to one randomly selected question from the list of 2 questions to get the date format of the informants. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution					
				Central Gujarat		South Gujarat		Saurashtra	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	**32**	16	16	0	0	0	0	16	16
21 To 50	**234**	104	130	0	2	20	54	84	74
50+	**138**	70	68	0	2	28	14	42	52
Total	**404**	**190**	**214**	**0**	**4**	**48**	**68**	**142**	**142**

Table 71: Distribution of Gujarati Date Format

6.8.3 The Sentences-S

The Sentences contain a list of sentences that is a representation of all most all the phonemes occurring in Gujarati language. 25 Randomly selected Sentences are recorded from a list of 200 sentences. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution					
				Central Gujarat		South Gujarat		Saurashtra	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	**402**	200	202	0	0	0	0	200	202
21 To 50	**2954**	1331	1623	25	23	250	675	1056	925
50+	**1725**	876	849	0	25	350	175	526	649
Total	**5081**	**2407**	**2674**	**25**	**48**	**600**	**850**	**1782**	**1776**

Table 72: Distribution of Gujarati Sentences

6.8.4 Command and Control Words-W1

The Command and Control Words contain a list of 296 words that is a representation of all most all the command and control words occurring in Gujarati. 30 randomly selected words are recorded from the list. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution					
				Central Gujarat		South Gujarat		Saurashtra	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	**480**	239	241	0	0	0	0	239	241
21 To 50	**3538**	1590	1948	30	28	300	810	1260	1110
50+	**2069**	1049	1020	0	30	419	210	630	780
Total	**6087**	**2878**	**3209**	**30**	**58**	**719**	**1020**	**2129**	**2131**

Table 73: Distribution of Gujarati Command and Control Words

6.8.5 Person Names –W2

The Person Names contain a list of 543 popular pan- Indian and regional person names. 20 randomly selected names are recorded from the list. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution					
				Central Gujarat		South Gujarat		Saurashtra	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	**322**	160	162	0	0	0	0	160	162
21 To 50	**2377**	1079	1298	20	20	198	539	861	739
50+	**1380**	700	680	0	20	280	140	420	520
Total	**4079**	**1939**	**2140**	**20**	**40**	**478**	**679**	**1441**	**1421**

Table 74: Distribution of Gujarati Person Names

6.8.6 Place Names-W2

The Place Names contain a list of 359 popular pan- Indian and regional place names. 10 randomly selected names are recorded from the list. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution					
				Central Gujarat		South Gujarat		Saurashtra	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	**160**	80	80	0	0	0	0	80	80
21 To 50	**1191**	542	649	12	9	100	270	430	370
50+	**690**	350	340	0	10	140	70	210	260
Total	**2041**	**972**	**1069**	**12**	**19**	**240**	**340**	**720**	**710**

Table 75: Distribution of Gujarati Place Names

6.8.7 Most Frequent Words-PART-W3A

The Most Frequent Words-part contains a list of 1,000 most frequent words. 30 randomly selected words are recorded from the list. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution			
				South Gujarat		Saurashtra	
		Female	Male	Female	Male	Female	Male
16 To 20	**480**	240	240	0	0	240	240
21 To 50	**2360**	1251	1109	0	0	1251	1109
50+	**779**	0	779	0	30	617	749
Total	**3619**	**1491**	**2128**	**0**	**30**	**2108**	**2098**

Table 76: Distribution of Gujarati Most Frequent Words – Part

6.9 FULL SET

The Full Set is the master set of certain data set which is red completely from few selected speakers in each group. Full sets are as below.

6.9.1 Most Frequent Words-Full-W3B

The Most Frequent Words contain a list of 1000 most frequent words. In full set all the 1000 words are recorded from the informant. The distribution of data is as follows.

:Age Group	Total Audio Segments	Gender-wise Distribution	
		South Gujarat Female	Central Gujarat Male
50+	**2000**	1000	1000

Table 77: Distribution of Gujarati Most Frequent Words – Full

6.9.2 The Phonetically Balanced Vocabulary-W4

The Phonetically Balanced Vocabulary contains a list of words where all most all the phones of Gujarati language have occurred in all the possible positions of a word. In full set all the 689 words are recorded from the informant where they uttered those words three times. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution	
		South Gujarat Female	Central Gujarat Male
50+	**1378**	689	689

Table 78: Distribution of Gujarati Phonetically Balanced Vocabulary

6.10 NATIVE SPEAKER DISTRIBUTIONS

For Gujarati speech data a total of 205 speakers were collected in which 97 female speakers and 108 male speakers from three different regions. The distribution of data is as follows:

Age Group	Total Speakers	Gender-wise Distribution		Regions					
				South Gujarat		Central Gujarat		Saurashtra	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	**16**	8	8	0	0	0	0	8	8
21 To 50	**120**	54	66	10	27	1	1	43	38
50+	**69**	35	34	14	7	0	1	21	26
Total	**205**	**97**	**108**	**24**	**34**	**1**	**2**	**72**	**72**

Table 79: Distribution of Gujarati Native Speakers

7 INDIAN ENGLISH RAW SPEECH CORPUS - BENGALI VARIANT

Rejitha K.S., Rajesha N., Manasa G., Narayan Choudhary

7.1 INTRODUCTION

English language is a blend of Anglo-Saxon which is the prominent language of Britain in Middle Ages. The English language propagated to every corner of the world during the colonial period. The colonisation imposed English as an administrative language in British colonies like Africa, America, Australia, Asia and so on. Even if British English is considered as Standard English, there are varieties like American English, Australian English, Indian English. English emerges as the most visible legacy of British in India because India was under the British rule for almost two centuries and English is a part of education system here. In the streets of all but the remotest villages, it is possible for strangers to communicate, though minimally in English. India has the highest English speaking population second only to the USA.

In India, English has the status of second official language, after Hindi. However, in practice, it is the first language of the Central Government of India as well as the language of the judiciary. English is also used predominantly in higher education, specially in that of scientific education. Through Official Language Act 1963, the second official language status of English was extended to an indefinite period.

The primary areas of English used in India are education, administration, law, mass media, science, and technology. It is used predominantly for trade and commerce. Part XVII of the Constitution of India states that all orders, rules and regulations, and bye-laws etc. shall be in English Language. Language of High Court and Supreme Court is English. Medicine, engineering, technology and all the sciences which are taken mostly from the West and its education systems are in English.

In India, states were pre-dominantly divided based on linguistic geography. English also acts as an inter-language in India, specially among the educated mass. The variations in the phonological features of the speech communities belonging to different mother tongues are obvious because of mother tongue influence. The native language influence is prominent in both the pronunciation and the vocabulary. Some Indian origin words such as veranda, curry, shawl etc. have also attained the universal status in English language.

The phonemic system of Standard English (British or American or any other variety) has several differences with what we call as Indian English. For example, all prominent Indian languages have a phonemic aspiration which at both voiced and voiceless stops. However, this is not true of British or American English. Similarly, Indian English does not have some vowels present in

British or American English (e.g., /ɛ/ and /ɔ/). While there is something that we call as Indian English (Bansal, 1976), it is not well taken and cannot always be encompassing the different phonological variations we have across different linguistic regions within India. While there may a similarity among well trained speakers of Indian English across various linguistic regions, most of the time it may not be true.

This brings us to the peculiar situation of collecting English speech samples from various linguistic regions within India. Bengali is second most spoken language in India after Hindi and is fifth most spoken native language of the world. It has its own phonetic peculiarities that make it different from the other languages of India and this has its own bearing on the English being spoken as a second language by the Bengali native speakers.

The presented corpus of Indian English represents the variety of English as spoken by the Bengali native speakers. The Collection of Indian English Bengali data is carried out in region of West Bengal of India from 28-12-2009 to 09-01-2010.

7.2 DATASET PREPARATION FOR INDIAN ENGLISH – BENGALI VARIANT

LDC-IL prepared the following dataset for which the prompt sheets were prepared.

Content Type	Count
Creative Text	6
Date	2
Command and Control Words	246
Most Frequent Words	1,000
Form and Function Words	154
Phonetically Balanced Words	299
Person Name	350
Place Name	242
Sentences	224

Table 80: LDC-IL Speech Dataset

Distinct news items were prepared to get the audio recording of contemporary text. It was made sure that each selected news item had minimum 500 words. Each prompt sheet has a distinct news item. The selected part of the dataset prepared is as follows:

Content Type	Content that Each Typical Prompt Sheet had	Content Selection Type
Contemporary Text	1 Text	Distinct Text
Creative Text	1 text	Random Text selected from dataset*
Sentences	25 Sentences	Random set selected from dataset*
Command and Control Words	30 Words	Random set selected from dataset*
Person Names	20 Words	Random set selected from dataset*
Place Names	10 Words	Random set selected from dataset*
Most Frequent Words	30 Words	Random set selected from dataset*
		*randomly selected by machine

Table 81: Table of Contents in LDC-IL Dataset

The Full Set of
1. Phonetically Balanced Vocabulary
2. Form and Function Words
3. 1000 Most Frequent Words

were also carried to the field to get recorded by selected individuals. Once all these preparations were made, the investigator started collecting the data.

7.3 SUMMARY OF THE CORPUS

In the sections below, we have provided the tabular details of the different content types of the LDC-IL Indian English - Bengali Variant Raw Speech Corpus based on various yardsticks which can also be filtered out from the metadata sheet. These figures may be helpful in tuning the corpus for various purposes of training, testing, and evaluating various algorithms as well as provides useful insights into the dataset. The data size is of total duration 25:50:17 (hh:mm:ss) comprising 16,044 audio segments.

7.4 SUMMARY OF THE AUDIO SEGMENTS

The table below shows the total number of Audio Segments and their distribution in the Indian English Bengali speech dataset.

LDC-IL Indian English - Bengali Variant	Gender →	Female			Male		
	Age Group →	16-20 Years	21-50 Years	50+ Years	16-20 Years	21-50 Years	50+ Years
Content Type	Total Segments	Segments	Segments	Segments	Segments	Segments	Segments
Contemporary Text (News)-T1	52	5	16	5	5	16	5
Creative Text-T2	52	5	16	5	5	16	5
Sentence-S	1300	125	400	125	125	400	125
Date-D	104	10	32	10	10	32	10
Command and Control Words-W1	2882	366	695	396	366	694	365
Person Name-W2	1040	100	320	100	100	320	100
Place Name-W2	519	50	159	50	50	160	50
Most Frequent Word- Part-W3A	1442	120	482	150	120	450	120
Most Frequent Word-FullSet-W3B	5985	1000	1000	1000	1000	986	999
Phonetically Balanced-W4	1782	297	299	298	298	293	297
Form and Function Word-W5	886	148	148	148	148	146	148

Table 82: Audio Segments and their Distribution

7.5 DURATION OF THE RAW SPEECH DATA

The table below shows the duration of each of the content type and their distribution across a few factors.

LDC-IL Indian English - Bengali Variant	Gender →	Female			Male		
	Age Group →	16-20 Years	21-50 Years	50+ Years	16-20 Years	21-50 Years	50+ Years
Content Type	Total Duration	Duration (hh:mm:ss)	Duration (hh:mm:ss)	Duration (hh:mm:ss)	Duration (hh:mm:ss)	Duration (hh:mm:ss)	Duration (hh:mm:ss)
Contemporary Text (News)-T1	6:03:15	0:30:01	1:48:46	0:34:09	0:33:43	2:01:18	0:35:18
Creative Text-T2	2:41:17	0:14:34	0:53:53	0:18:57	0:14:11	0:42:51	0:16:51
Sentence-S	1:29:35	0:08:54	0:27:48	0:09:06	0:07:57	0:26:37	0:09:13
Date-D	0:08:56	0:00:47	0:02:47	0:00:52	0:00:55	0:02:37	0:00:58
Command and Control Words-W1	3:09:13	0:37:26	0:38:28	0:26:30	0:24:29	0:41:24	0:20:56
Place Name-W2	0:33:56	0:03:55	0:09:48	0:02:54	0:03:18	0:10:50	0:03:11
Person Name-W2	1:30:22	0:09:55	0:26:20	0:07:50	0:08:15	0:28:38	0:09:24
Most Frequent Word-Part- W3A	1:22:38	0:06:44	0:26:44	0:07:50	0:07:20	0:27:24	0:06:36
Most Frequent Word-FullSet W3B	6:01:44	1:38:11	0:43:14	1:00:44	1:09:03	0:40:53	0:49:39
Phonetically Balanced-W4	1:52:21	0:42:12	0:12:45	0:17:30	0:16:02	0:10:18	0:13:34
Form and Function-Word-W5	0:53:54	0:19:26	0:06:10	0:08:51	0:08:00	0:04:45	0:06:42

Table 83: Duration of the Collected Data

7.6 DISTINCT SET

The Distinct Set usually contains data which is distinct to each speaker and is rarely repeated. The LDC-IL speech data set contains newspaper extracts which are read by each speaker.

7.6.1 The Contemporary Text (News) - T1

Distinct Text Extracts from Newspapers are recorded from the informants to get the speech data of contemporary text. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution	
		Female	Male
16 to 20	10	5	5
21 to 50	32	16	16
50+	10	5	5
Total	**52**	**26**	**26**

Table 84: Distribution of Indian English - Bengali Variant Contemporary Text (News) Data

7.7 RANDOM SET

The Random Set data comprises of content types which are sampled by machine for each speaker. They are sampled from collection of master data sets available. The random sets are given below.

7.7.1 The Creative Text-T2

One randomly selected text of literature out of 6 texts from the prepared dataset is recorded from the informants to get the speech data of Creative text. The distribution of data is as follows.

Age Group	Total Audio Segments	Gender-wise Distribution	
		Female	Male
16 to 20	10	5	5
21 to 50	32	16	16
50+	10	5	5
Total	**52**	**26**	**26**

Table 85: Distribution of Indian English - Bengali Variant Creative Text

7.7.2 The Date-D

The answer to one randomly selected question from the list of 2 questions to get the date format of the informants. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution	
		Female	Male
16 to 20	20	10	10
21 to 50	64	32	32
50+	20	10	10
Total	**104**	**52**	**52**

Table 86: Distribution of Indian English - Bengali Variant Date Format

7.7.3 The Sentences-S

The Sentences contain a list of sentences which is a representation of all the phonemes occurring in Indian English. 25 Randomly selected sentences are recorded from a list of 224 sentences. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution	
		Female	Male
16 to 20	250	125	125
21 to 50	800	400	400
50+	250	125	125
Total	**1300**	**650**	**650**

Table 87: Distribution of Indian English - Bengali Variant Sentences

7.7.4 Command and Control Words-W1

The Command and Control Words contain a list of 246 words that is a representation of all most all the command and control words occurring in Indian English. 30 randomly selected words are recorded from the list. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution	
		Female	Male
16 to 20	732	366	366
21 to 50	1389	695	694
50+	761	396	365
Total	**2882**	**1457**	**1425**

Table 88: Distribution of Indian English - Bengali Variant Command and Control Words

7.7.5 Person Names –W2

The Person Names contain a list of 350 popular pan-Indian and regional person names. 20 randomly selected names are recorded from the list. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution	
		Female	Male
16 to 20	200	100	100
21 to 50	640	320	320
50+	200	100	100
Total	**1040**	**520**	**520**

Table 89: Distribution of Indian English - Bengali Variant Person Names

7.7.6 Place Names-W2

The Place Names contain a list of 242 popular pan-Indian and regional place names. 10 randomly selected names are recorded from the list. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution	
		Female	Male
16 to 20	100	50	50
21 to 50	319	159	160
50+	100	50	50
Total	**519**	**259**	**260**

Table 90: Distribution of Indian English - Bengali Variant Place Names

7.7.7 Most Frequent Words-PART-W3A

The Most Frequent Words-part contains a list of 1,000 most frequent words. 30 randomly selected words are recorded from the list. The distribution of the data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution	
		Female	Male
16 to 20	240	120	120
21 to 50	932	482	450
50+	270	150	120
Total	**1442**	**752**	**690**

Table 91: Distribution of Indian English - Bengali Variant Most Frequent Words – Part

7.8 FULL SET

The Full Set is the master set of certain data set which is red completely from few selected speakers in each group. Full sets are as below:

7.8.1 Most Frequent Words-Full-W3B

The Most Frequent Words contain a list of 1000 most frequent words. In full set all the 1000 words are recorded from the informant. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution	
		Female	Male
16 to 20	2000	1000	1000
21 to 50	1986	1000	986
50+	1999	1000	999
Total	**5985**	**3000**	**2985**

Table 92: Distribution of Indian English - Bengali Variant Most Frequent Words – Full

7.8.2 The Phonetically Balanced Vocabulary-W4

The Phonetically Balanced Vocabulary contains a list of words where all most all the phones of Indian English language have occurred in all the possible positions of a word. In full set all the 299 words are recorded from the informant where they uttered those words three times. The distribution of the data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution	
		Female	Male
16 to 20	595	297	298
21 to 50	592	299	293
50+	595	298	297
Total	**1782**	**894**	**888**

Table 93: Distribution of Indian English - Bengali Variant Phonetically Balanced Vocabulary

7.8.3 The Form and Function Words-W5

The Form and Function Words contain a list of 154 words which is a representation of all most all the form and function words occurring in Indian English. All the words are recorded from the informant where they uttered those words three times. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution	
		Female	Male
16 to 20	296	148	148
21 to 50	294	148	146
50+	296	148	148
Total	**886**	**444**	**442**

Table 94: Distribution of Indian English - Bengali Variant Form and Function words

7.9 REFERENCES

1. Pingali Sailaja (2009), *Indian English*, Edinburgh University Press, Edinburgh.
2. Bansal, R.K. 1976. The Intelligibility of Indian English. CIEFL, Hyderabad.

8 INDIAN ENGLISH RAW SPEECH CORPUS - KANNADA VARIANT

Rejitha K.S., Rajesha N., Manasa G., Narayan Choudhary

8.1 INTRODUCTION

English language is a blend of Anglo-Saxon which is the prominent language of Britain in Middle Ages. The English language propagated to every corner of the world during the colonial period. The colonisation imposed English as an administrative language in British colonies like Africa, America, Australia, Asia and so on. Even if British English is considered as Standard English, there are varieties like American English, Australian English, Indian English. English emerges as the most visible legacy of British in India because India was under the British rule for almost two centuries and English is a part of education system here. In the streets of all but the remotest villages, it is possible for strangers to communicate, though minimally in English. India has the highest English speaking population second only to the USA.

In India, English has the status of second official language, after Hindi. However, in practice, it is the first language of the Central Government of India as well as the language of the judiciary. English is also used predominantly in higher education, specially in that of scientific education. Through Official Language Act 1963, the second official language status of English was extended to an indefinite period.

The primary areas of English used in India are education, administration, law, mass media, science, and technology. It is used predominantly for trade and commerce. Part XVII of the Constitution of India states that all orders, rules and regulations, and bye-laws etc. shall be in English Language. Language of High Court and Supreme Court is English. Medicine, engineering, technology and all the sciences which are taken mostly from the West and its education systems are in English.

In India, states were pre-dominantly divided based on linguistic geography. English also acts as an inter-language in India, specially among the educated mass. The variations in the phonological features of the speech communities belonging to different mother tongues are obvious because of mother tongue influence. The native language influence is prominent in both the pronunciation and the vocabulary. Some Indian origin words such as veranda, curry, shawl etc. have also attained the universal status in English language.

The phonemic system of standard English (British or American or any other variety) has several differences with what we call as Indian English. For example, all prominent Indian languages have a phonemic aspiration which at both voiced and voiceless stops. However, this is not true of

British or American English. Similarly, Indian English does not have some vowels present in British or American English (e.g., /ɛ/ and /ɔ/). While there is something that we call as Indian English (Bansal, 1976), it is not well taken and cannot always be encompassing the different phonological variations we have across different linguistic regions within India. While there may a similarity among well trained speakers of Indian English across various linguistic regions, most of the time it may not be true.

This brings us to the peculiar situation of collecting English speech samples from various linguistic regions within India. Bengali is second most spoken language in India after Hindi and is fifth most spoken native language of the world. It has its own phonetic peculiarities that make it different from the other languages of India and this has its own bearing on the English being spoken as a second language by the Bengali native speakers.

The presented corpus of Indian English represents the variety of English as spoken by the Bengali native speakers. The Collection of Indian English Bengali data is carried out in region of Karnataka from 19-12-2009 to 31-01-2010.

8.2 DATASET PREPARATION FOR INDIAN ENGLISH - KANNADA VARIANT

LDC-IL prepared the following dataset for which the prompt sheets were prepared.

Content Type	Count
Creative Text	6
Date	2
Command and Control Words	246
Most Frequent Words	1,000
Form and Function Words	154
Phonetically Balanced Words	299
Person Name	650
Place Name	242
Sentences	224

Table 95: LDC-IL Speech Dataset

Distinct News Items were prepared to get the audio recording of contemporary text. It was made sure that each selected news item had minimum 500 words. Each prompt sheet had a distinct news item and sleeted part of the dataset prepared as follows.

Content Type	Content that Each Typical Prompt Sheet had	Content Selection Type
Contemporary Text	1 Text	Distinct Text
Creative Text	1 text	Random Text selected from dataset*
Sentences	25 Sentences	Random set selected from dataset*
Command and Control Words	30 Words	Random set selected from dataset*
Person Names	20 Words	Random set selected from dataset*
Place Names	10 Words	Random set selected from dataset*
Most Frequent Words	30 Words	Random set selected from dataset*
		*randomly selected by machine

Table 96: Table of Contents in LDC-IL Dataset

The Full Set of

1. Phonetically Balanced Vocabulary
2. Form and Function Words
3. 1000 Most Frequent Words

were also carried to the field to get recorded by selected individuals. Once all these preparations were made, the investigator started collecting the data.

8.3 SUMMARY OF THE CORPUS

In the sections below, we provide the tabular details of the different content types of the Indian English - Kannada Variant Raw Speech Corpus based on various yardsticks which can also be filtered out from the metadata sheet. These figures may be helpful in tuning the corpus for various purposes of training, testing, and evaluating various algorithms as well as provides useful insights into the dataset. The data size is of total duration 23:43:04 (hh:mm:ss) comprising 14,455 audio segments.

8.4 SUMMARY OF THE AUDIO SEGMENTS

The table below shows the total number of Audio Segments and their distribution in the Indian English – Kannada Variant Speech dataset.

LDC-IL Indian English - Kannada Variant	Gender →	Female			Male		
	Age Group →	16-20 Years	21-50 Years	50+ Years	16-20 Years	21-50 Years	50+ Years
Content Type	Total Segments	Segments	Segments	Segments	Segments	Segments	Segments
Contemporary Text (News)-T1	52	7	14	5	5	16	5
Creative Text-T2	58	7	14	5	11	16	5
Sentence-S	1522	174	350	125	349	400	124
Date-D	106	14	28	10	12	32	10
Command and Control Words-W1	2543	456	665	396	396	480	150
Person Name-W2	2040	140	280	450	750	320	100
Place Name-W2	762	70	140	50	292	160	50
Most Frequent Word-Part-W3A	1563	210	420	150	149	484	150
Most Frequent Word-FullSet-W3B	3999	1000	1000	999	1000	0	0
Phonetically Balanced-W4	1194	299	297	299	299	0	0

Form and Function Word-W5	616	154	154	154	154	0	0

Table 97: Audio Segments and their Distribution

8.5 DURATION OF THE RAW SPEECH DATA

The table below shows the duration of each of the content type and their distribution across a few factors.

LDC-IL Indian English - Kannada Variant	Gender →	Female			Male		
	Age Group →	16-20 Years	21-50 Years	50+ Years	16-20 Years	21-50 Years	50+ Years
Content Type	Total Duration	Duration (hh:mm:ss)	Duration (hh:mm:ss)	Duration (hh:mm:ss)	Duration (hh:mm:ss)	Duration (hh:mm:ss)	Duration (hh:mm:ss)
Contemporary Text (News)-T1	7:19:31	1:19:43	1:50:13	0:33:16	0:45:59	2:04:34	0:45:46
Creative Text-T2	3:57:15	0:30:00	0:56:20	0:18:07	0:53:34	0:56:58	0:22:16
Sentence-S	1:54:10	0:12:27	0:28:28	0:09:10	0:28:11	0:26:54	0:09:00
Date-D	0:04:32	0:00:38	0:01:27	0:00:30	0:00:26	0:01:10	0:00:21
Command and Control Words-W1	1:55:43	0:20:53	0:31:31	0:17:09	0:17:20	0:21:19	0:07:31
Place Name-W2	0:39:43	0:03:29	0:07:59	0:02:31	0:15:24	0:07:38	0:02:42
Person Name-W2	2:38:49	0:10:55	0:23:46	0:33:59	0:59:55	0:22:39	0:07:35
Most Frequent Word-Part-W3A	1:09:10	0:09:07	0:21:18	0:06:31	0:06:17	0:18:55	0:07:02
Most Frequent Word-FullSet-W3B	2:49:55	0:41:50	0:37:11	0:51:00	0:39:54	0:00:00	0:00:00
Phonetically Balanced-W4	0:49:21	0:10:10	0:11:31	0:15:51	0:11:49	0:00:00	0:00:00
Form and Function Word-W5	0:24:55	0:05:07	0:06:04	0:07:47	0:05:57	0:00:00	0:00:00

Table 98: Duration of the Collected Data

8.6 DISTINCT SET

The Distinct Set usually contains data which is distinct to each speaker and is rarely repeated. The LDC-IL speech data set contains newspaper extracts which are read by each speaker.

8.6.1 The Contemporary Text (News) - T1

Distinct Text Extracts from Newspapers are recorded from the informants to get the speech data of contemporary text. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution	
		Female	Male
16 to 20	12	7	5
21 to 50	30	14	16
50+	10	5	5
Total	**52**	**26**	**26**

Table 99 Distribution of Contemporary Text (News) Data

8.7 RANDOM SET

The Random Set data comprises of content types which are sampled by machine for each speaker. They are sampled from collection of master data sets available. The random sets are given below.

8.7.1 The Creative Text-T2

One randomly selected text of literature out of 6 texts from the prepared dataset is recorded from the informants to get the speech data of Creative text. The distribution of data is as follows.

Age Group	Total Audio Segments	Gender-wise Distribution	
		Female	Male
16 to 20	18	7	11
21 to 50	30	14	16
50+	10	5	5
Total	**58**	**26**	**32**

Table 100: Distribution of Indian English Kannada Creative Text

8.7.2 The Date-D

The answer to one randomly selected question from the list of 2 questions to get the date format of the informants. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution	
		Female	Male
16 to 20	26	14	12
21 to 50	60	28	32
50+	20	10	10
Total	**106**	**52**	**54**

Table 101: Distribution of Indian English Kannada Date Format

8.7.3 The Sentences-S

The sentences section contains a list of sentences that is the representation of all phonemes occurring in Indian English. 25 Randomly selected sentences are recorded from a list of 224 sentences. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution	
		Female	Male
16 to 20	523	174	349
21 to 50	750	350	400
50+	249	125	124
Total	**1522**	**649**	**873**

Table 102: Distribution of Indian English Sentences

8.7.4 Command and Control Words-W1

The Command and Control Words contain a list of 246 words that is a representation of all most all the command and control words occurring in Indian English. 30 randomly selected words are recorded from the list. The distribution of the data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution	
		Female	Male
16 to 20	852	456	396
21 to 50	1145	665	480
50+	546	396	150
Total	**2543**	**1517**	**1026**

Table 103: Distribution of Indian English Command and Control Words

8.7.5 Person Names –W2

The Person Names contain a list of 650 popular pan-Indian and regional person names. 20 randomly selected names are recorded from the list. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution	
		Female	Male
16 to 20	890	140	750
21 to 50	600	280	320
50+	550	450	100
Total	**2040**	**870**	**1170**

Table 104: Distribution of Kannada Person Names

8.7.6 Place Names-W2

The Place Names contain a list of 242 popular pan-Indian and regional place names. 10 randomly selected names are recorded from the list. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution	
		Female	Male
16 to 20	362	70	292
21 to 50	300	140	160
50+	100	50	50
Total	**762**	**260**	**502**

Table 105: Distribution of Kannada Place Names

8.7.7 Most Frequent Words-PART-W3A

The Most Frequent Words-part contains a list of 1,000 most frequent words. 30 randomly selected words are recorded from the list. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution	
		Female	Male
16 to 20	359	210	149
21 to 50	904	420	484
50+	300	150	150
Total	**1563**	**780**	**783**

Table 106: Distribution of Indian English Most Frequent Words – Part

8.8 FULL SET

The Full Set is the master set of certain data set which is red completely from few selected speakers in each group. Full sets are as below.

8.8.1 Most Frequent Words-Full-W3B

The Most Frequent Words contain a list of 1000 most frequent words. In full set all the 1000 words are recorded from the informant. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution	
		Female	Male
16 to 20	2000	1000	1000
21 to 50	1000	1000	0
50+	999	999	0
Total	**3999**	**2999**	**1000**

Table 107: Distribution of Indian English Most Frequent Words – Full

8.8.2 The Phonetically Balanced Vocabulary-W4

The Phonetically Balanced Vocabulary contains a list of words where all most all the phones of Indian English Language have occurred in all the possible positions of a word. In full set all the 299 words are recorded from the informant where they uttered those words three times. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution	
		Female	Male
16 to 20	598	299	299
21 to 50	297	297	0
50+	299	299	0
Total	**1194**	**895**	**299**

Table 108: Distribution of Indian English Phonetically Balanced Vocabulary

8.8.3 The Form and Function Words-W5

The Form and Function Words contain a list of 154 words which is a representation of all most all the form and function words occurring in Indian English. All the words are recorded from the informant where they uttered those words three times. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution	
		Female	Male
16 to 20	308	154	154
21 to 50	154	154	0
50+	154	154	0
Total	**616**	**462**	**154**

Table 109: Distribution of Form and Function words

8.9 REFERENCES

1. Pingali Sailaja (2009), *Indian English*, Edinburgh University Press, Edinburgh.
2. Bansal, R.K. 1976. The Intelligibility of Indian English. CIEFL, Hyderabad.

9 ODIA RAW SPEECH CORPUS

Santosh Kumar Mohanty, Rajesha N., Manasa G., Narayan Choudhary

9.1 INTRODUCTION

Odia (formerly Oriya) is one of the most prominent languages among the Indo-Aryan language family and designated as a Classical Language status by the Government of India. It is mainly spoken in the state of Odisha and in some of the bordering States like West Bengal, Jharkhand, Chhattisgarh, and Andhra Pradesh. Odia is the administrative language of the state of Odisha and the second official language of Jharkhand. The Odia language uses a script by its own name which originated from the Bramhi script and it also has been used as a regional writing-system for Sanskrit and different tribal languages of Odisha.

Odia is a syllabic alphabet wherein all consonants have an inherent vowel embedded within. The diacritics (which can be appearing before, above, below or after the consonant they belong to) are used to change the form of the inherent vowel. When vowels appear at the beginning of a syllable, they are written as independent letters. Similarly, when certain consonants occur together, special conjunct symbols are used to combine the essential parts of each consonant symbols.

The state of Odisha was formed by integrating Odia speaking areas. The regions were previously administrated by princely states and British Presidencies. The education level, mother tongue and the language used by previous administration play a role in characterizing the variety of Odia spoken in Odisha. Therefore, Odia language is influenced by the English, Arabic, Persian, Hindi, Bangla, Marathi, Dravidian, and Mundari languages etc.

Regional varieties (dialect) of Odia may be tentatively and roughly divided into five categories (Dash, 2005:205). They are, North-Coastal (Odia spoken in the districts of Balasore, Bhadrak etc.), Central-Coastal (spoken in the districts of Cuttack, Puri, Khordha, Jajapur and Jagatsingpur etc.), South-Coastal (spoken in the districts of Ganjam, Gajapati etc.), North-and-Central-Western or briefly Western (spoken in the districts of Sambalpur, Baragada, Balangiri, Sonepur etc.) and South-Western (spoken in the districts of Kalahandi, Rayagada, Nabarangpur etc.). The Standard Odia is mainly based on Central-Coastal dialect for extra linguistic reasons.

LDC-IL has collected data from two regional varieties i.e., Central-Coastal and North-Coastal. In future LDC-IL may collect data from other regions of Odisha.

9.2 DATASET PREPARATION FOR ODIA

For the selected Central and Northern regions of Odisha LDC-IL prepared the following dataset using which the prompt sheets were prepared.

Content Type	Count
Creative Text	6
Date	2
Command and Control Words	500
Most Frequent Word	1000
Form and Function Word	167
Phonetically Balanced Word	605
Person Name	499
Place Name	295
Sentence	252

Table 110: LDC-IL Odia Speech Dataset

Distinct news items were prepared to get the audio recording of contemporary text. It was made sure that each selected news item had minimum 500 words. Each prompt sheet had a distinct news item and selected part of the dataset prepared as follows.

Content Type	Content that Each Typical Prompt Sheet had	Content Selection Type
Contemporary Text	1 Text	Distinct Text
Creative Text	1 text	Random Text selected from dataset*
Sentences	25 Sentences	Random set selected from dataset*
Command and Control Words	30 Words	Random set selected from dataset*
Person Name	20 Words	Random set selected from dataset*
Place Name	10 Words	Random set selected from dataset*
Most Frequent Word	30 Words	Random set selected from dataset*
		*randomly selected by machine

Table 111: Table of Contents in LDC-IL Dataset

The Full Set of

1. Phonetically Balanced Vocabulary
2. Form and Function Words
3. 1000 Most Frequent Words

were also carried to the field to get recorded by selected individuals. Once all these preparations were made, the investigator started collecting the data.

The collection of data is carried out in three phases for different regions of Odisha through field works by Raja Kumar Naik (2009), Pramod Kumar Rout (2010 and 2012) and Kshirod Kumar Das (2010).

9.3 TRANSLITERATIONS IN LDC-IL ODIA READ CORPUS

For easy reference and uniformity, the recorded text in the metadata file, is transliterated from Odia to Roman letters. Numeric characters were transliterated from Odia to Hindu-Arabic system.

The LDC-IL transliteration scheme of Odia to Roman is given below.

LDC-IL Transliteration Schema

Odia Characters to Roman and Odia Numerals to Hindu-Arabic

Vowels and Vowel Signs										
ଅ	ଆ	ଇ	ଈ	ଉ	ଊ	ଋ	ଏ	ଐ	ଓ	ଔ
	ା	ି	ୀ	ୁ	ୂ	ୃ	େ	ୈ	ୋ	ୌ
a	A	i	I	u	U	x	e	ai	o	au

Consonants				
କ	ଖ	ଗ	ଘ	ଙ
ka	kha	ga	gha	ng'a
ଚ	ଛ	ଜ	ଝ	ଞ
ca	cha	ja	jha	nj'a
ଟ	ଠ	ଡ	ଢ	ଣ
Ta	Tha	Da	Dha	Na
ତ	ଥ	ଦ	ଧ	ନ
ta	tha	da	dha	na
ପ	ଫ	ବ	ଭ	ମ
pa	pha	ba	bha	ma

Ajogabaaha		
ଂ	ଃ	ଁ
M	H	m'

ଯ	ୟ	ର	ଳ	ଲ	ୱ ୱ	ଶ	ଷ	ସ	ହ
ya	Ya	ra	la	La	wa	sha	Sa	sa	ha

Numerals (Odia to Hindu-Arabic)									
୦	୧	୨	୩	୪	୫	୬	୭	୮	୯
0	1	2	3	4	5	6	7	8	9

9.4 SUMMARY OF THE CORPUS

In the sections below, we provide the tabular details of the different content types of the Odia raw speech corpus based on various yardsticks which can also be filter out from the metadata sheet. 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 data size is of total duration 138:06:18 (hh:mm:ss) comprising 73,418 audio segments of 89.GB size.

9.5 SUMMARY OF THE AUDIO SEGMENTS

The table below shows the total number of Audio Segments and their distribution in the Odia speech dataset.

LDC-IL Odia Speech Data Status	Gender →	Female			Male		
	Age Group →	16-20 Years	21-50 Years	50+ Years	16-20 Years	21-50 Years	50+ Years
Content Type	Total Segments	Segments	Segments	Segments	Segments	Segments	Segments
Contemporary Text (News)-T1	449	23	141	61	25	138	61
Creative Text-T2	450	23	142	61	25	138	61
Sentence-S	11248	575	3550	1524	625	3449	1525
Date-D	900	46	284	122	50	276	122
Command and Control Words-W1	13499	690	4259	1831	750	4140	1829
Person Name-W2	8998	460	2839	1220	500	2759	1220
Place Name-W2	4496	230	1420	607	250	1380	609
Most Frequent Word-Part-W3A	8994	450	2728	1320	510	2697	1289
Most Frequent Word-FullSet-W3B	10989	998	4994	0	1000	3997	0
Phonetically Balanced-W4	10438	1218	3656	635	1240	3054	635
Form and Function Word-W5	2957	426	1007	172	339	841	172

Table 112: Audio Segments and their distribution of Odia Speech Data

9.6 DURATION OF THE RAW SPEECH DATA

The table below shows the duration of each of the content type and their distribution across a few factors.

LDC-IL Odia Speech Data Status	Gender →	Female			Male		
	Age Group →	16-20 Years	21-50 Years	50+ Years	16-20 Years	21-50 Years	50+ Years
Content Type	Total Duration	Duration (hh:mm:ss)	Duration (hh:mm:ss)	Duration (hh:mm:ss)	Duration (hh:mm:ss)	Duration (hh:mm:ss)	Duration (hh:mm:ss)
Contemporary Text (News)-T1	**42:49:56**	1:58:15	12:55:57	6:32:22	2:20:56	12:27:12	6:35:12
Creative Text-T2	**19:43:50**	0:55:42	6:05:47	2:49:02	1:03:14	5:54:39	2:55:22
Sentence-S	**08:22:57**	0:24:58	2:39:48	1:12:45	0:24:37	2:29:21	1:11:26
Date-D	**1:27:49**	0:04:16	0:27:02	0:12:34	0:04:35	0:26:58	0:12:21
Command and Control Words-W1	**14:18:49**	0:43:14	4:30:55	1:57:10	0:46:22	4:30:53	1:50:13
Place Name-W2	**5:01:40**	0:15:13	1:34:43	0:40:08	0:17:16	1:36:25	0:37:54
Person Name-W2	**13:22:45**	0:40:14	4:14:10	1:53:49	0:41:23	4:06:31	1:46:35
Most Frequent Word-Part-W3A	**9:40:04**	0:31:47	2:54:28	1:19:38	0:36:07	3:06:54	1:11:10
Most Frequent Word-FullSet-W3B	**10:21:04**	1:02:06	4:21:42	0:00:00	0:58:46	3:58:30	0:00:00
Phonetically Balanced-W4	**10:05:10**	1:16:44	3:14:58	0:38:27	1:13:44	3:04:16	0:37:01
Form and Function-Word-W5	**02:52:14**	0:24:34	0:52:03	0:09:52	0:22:24	0:53:26	0:09:55

Table 113: Duration of the collected Odia Speech Data

9.7 DISTINCT SET

The Distinct Set usually contains data which is distinct to each speaker and is rarely repeated. The LDC-IL speech data set contains newspaper extracts which are read by each speaker.

9.7.1 The Contemporary Text (News) - T1

Distinct Text Extracts from newspapers are recorded from the informants to get the speech data of contemporary text. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution			
				Central		Northern	
		Female	Male	Female	Male	Female	Male
16 to 20	48	23	25	21	25	2	0
21 to 50	279	141	138	137	132	4	6
50+	122	61	61	61	55	0	6
Total	**449**	**225**	**224**	**219**	**212**	**6**	**12**

Table 114: Distribution of Odia Contemporary Text (News) Data

9.8 RANDOM SET

The Random Set data comprises of content types which are sampled by machine for each speakers. They are sampled from collection of master data sets available. The random sets are given below.

9.8.1 The Creative Text-T2

One randomly selected text of literature out of 6 texts from the prepared dataset is recorded from the informants to get the speech data of Creative Text. The distribution of data is as follows.

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution			
				Central		Northern	
		Female	Male	Female	Male	Female	Male
16 to 20	48	23	25	21	25	2	0
21 to 50	280	142	138	137	132	5	6
50+	122	61	61	61	55	0	6
Total	**450**	**226**	**224**	**219**	**212**	**7**	**12**

Table 115: Distribution of Odia Creative Text

9.8.2 The Date-D

The answer to one randomly selected question from the list of 2 questions to get the date format of the informants. Both the question and answer are available in the data. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution			
				Central		Northern	
		Female	Male	Female	Male	Female	Male
16 to 20	96	46	50	42	50	4	0
21 to 50	560	284	276	274	264	10	12
50+	244	122	122	122	110	0	12
Total	**900**	**452**	**448**	**438**	**424**	**14**	**24**

Table 116: Distribution of Odia Date Format

9.8.3 The Sentence-S

The content type of Sentence contains a list of sentences that is the representation of all most all the phonemes occurring in Odia. 25 Randomly selected sentences are recorded from a list of 252 sentences. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution			
				Central		Northern	
		Female	Male	Female	Male	Female	Male
16 to 20	1200	575	625	525	625	50	0
21 to 50	6999	3550	3449	3425	3299	125	150
50+	3049	1524	1525	1524	1375	0	150
Total	**11248**	**5649**	**5599**	**5574**	**5299**	**175**	**300**

Table 117: Distribution of Odia Sentence

9.8.4 Command and Control Words-W1

The Command and Control Words contain a list of 500 words that is a representation of all most all the command and control words occurring in Odia. 30 randomly selected words are recorded from the list. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution			
				Central		Northern	
		Female	Male	Female	Male	Female	Male
16 to 20	1440	690	750	630	750	60	0
21 to 50	8399	4259	4140	4110	3960	149	180
50+	3660	1831	1829	1831	1649	0	180
Total	**13499**	**6780**	**6719**	**6571**	**6359**	**209**	**360**

Table 118: Distribution of Odia Command and Control Words

9.8.5 Person Name –W2

The Person Name contains a list of 499 popular pan- Indian and regional person names. 20 randomly selected names are recorded from the list. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution			
				Central		Northern	
		Female	Male	Female	Male	Female	Male
16 to 20	960	460	500	420	500	40	0
21 to 50	5598	2839	2759	2739	2640	100	119
50+	2440	1220	1220	1220	1100	0	120
Total	**8998**	**4519**	**4479**	**4379**	**4240**	**140**	**239**

Table 119: Distribution of Odia Person Name

9.8.6 Place Name-W2

The Place Name contains a list of 295 popular pan- Indian and regional place names. 10 randomly selected names are recorded from the list. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution			
				Central		Northern	
		Female	Male	Female	Male	Female	Male
16 to 20	480	230	250	210	250	20	0
21 to 50	2800	1420	1380	1370	1320	50	60
50+	1216	607	609	607	549	0	60
Total	**4496**	**2257**	**2239**	**2187**	**2119**	**70**	**120**

Table 120: Distribution of Odia Place Name

9.8.7 Most Frequent Words-PART-W3A

The Most Frequent Word-part contains a list of 1000 most frequent words. 30 randomly selected words are recorded from the list. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution			
				Central		Northern	
		Female	Male	Female	Male	Female	Male
16 to 20	960	450	510	450	510	0	0
21 to 50	5425	2728	2697	2728	2577	0	120
50+	2609	1320	1289	1320	1289	0	0
Total	**8994**	**4498**	**4496**	**4498**	**4376**	**0**	**120**

Table 121: Distribution of Odia Most Frequent Word – Part

9.9 FULL SET

The Full Set is the master set of certain data set which is red completely from few selected speakers in each group. Full sets are as below.

9.9.1 Most Frequent Words-Full-W3B

The Most Frequent Words contain a list of 1000 most frequent words. In full set all the 1000 words are recorded from the informant. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution in Central Dialect	
		Female	Male
16 to 20	1998	998	1000
21 to 50	8991	4994	3997
Total	**10989**	**5992**	**4997**

Table 122: Distribution of Odia Most Frequent Word – Full

9.9.2 The Phonetically Balanced Vocabulary-W4

The Phonetically Balanced Vocabulary contains a list of words where all most all the phones of Odia language have occurred in all the possible positions of a word. In full set all the 605 words are recorded from the informants where they uttered those words three times. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution in Central Dialect	
		Female	Male
16 to 20	2458	1218	1240
21 to 50	6710	3656	3054
50+	1270	635	635
Total	**10438**	**5509**	**4929**

Table 123: Distribution of Odia Phonetically Balanced Vocabulary

9.9.3 The Form and Function Word-W5

The Form and Function Word contain a list of 167 words which is a representation of all most all the form and function words occurring in Odia. All the words are recorded from the informant where they uttered those words three times. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution in Central Dialect	
		Female	Male
16 to 20	765	426	339
21 to 50	1848	1007	841
50+	344	172	172
Total	**2957**	**1605**	**1352**

Table 124: Distribution of Odia Form and Function word

9.10 NATIVE SPEAKER DISTRIBUTIONS

The following table shows the distribution of native speakers of Odia, across different regions.

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution			
				Central		Northern	
		Female	Male	Female	Male	Female	Male
16 to 20	55	27	28	25	28	2	0
21 to 50	293	149	144	144	138	5	6
50+	126	63	63	63	57	0	6
Total	**474**	**239**	**235**	**232**	**223**	**7**	**12**

Table 125: Distribution of Odia Native Speaker

REFERENCE

Dash, Gaganendra Nath. 2005. “Paschimanchaliya Odia” in *Nirbachita Prabandha Sankalana*, Cuttack: Vidyapuri.

10 TAMIL RAW SPEECH CORPUS

Srikanth D., Amudha R., Rajesha N., Manasa G., Narayan Choudhary

10.1 INTRODUCTION

Tamil is one of the prominent languages among the Dravidian language family. Tamil is widely spoken in the state of Tamil Nadu, Union Territory of Pondicherry, Sri Lanka, in East-Asian countries like Burma, Malaysia, Singapore, Indonesia, Indio china, Fiji, in South-Africa, British Guinea and in islands like Mauritius and Madagascar etc. Within India, it is the official language of the Indian state of Tamil Nadu and the Indian Union Territory of Pondicherry. Tamil also has an official language Sri Lanka and Singapore. The Tamil language uses a script by its own name which has originated from the Brahmi script. The language is highly agglutinative in nature. Compared to Sanskrit and other Indian languages.

Scholars have described six broad categories of geographical dialects. They are Chennai Tamil (Chengalpet, Tiruvallur, Chennai, Kanchipuram, and Pondicherry) Kongu Tamil (Tirupur, Nilgiris, Coimbatore, Erode), Kumari Tamil (Tamil spoken in coastal region of Kanyakumari), Madurai Tamil (Madurai, Sivaganga, Ramanathapuram, Dindigul, Teni and Virudunagar), Nellai Tamil (Tirunelveli and Tuticorin), Salem Tamil (Dharmapuri, Karur, Namakkal and Salem), Thanjai Tamil (Thanjavur Trichy, Pudukottai, Perambaur, Aruiyalur, Nagapattinam, Tiruvarur and Cuddalore) and Vellore Tamil (Vellore, Tiruvannamalai and Viluppuram). Of course, each one of these consists of sub-dialects which have their own distinctive features. Many of these distinctions occur because the dialects are strongly influenced by their neighbouring languages.

Chennai Tamil is a combination of Tamil, Telugu, Kannada, Urdu, and English. Kongu Tamil is a unique variety of Tamil as this dialect is very respectful and uses a unique kind of vocabulary in a respectful manner as standard Tamil. Kumari Tamil will be very tough to understand for someone who speaks Madras Bashai or Kongu Tamil, but is understandable to speakers of Madurai Tamil, Tirunelveli Tamil, or even Malayalam and this involves a lot of tongue rolling i.e., retroflex sounds. Madurai Tamil is also considered one of the purest Tamil dialects spoken in Tamil Nadu along with the Central Tamil dialect. Tirunelveli Tamil is also famously known as Nellai Tamil and is easily understood if someone knows Madurai Tamil well. Salem Tamil and Thanjai Tamil are easily understood among all Tamil speakers.

LDC-IL divided the Tamil speaking areas into these eight regions and collected speech data from six regions. Total speech data of 452 speakers, were collected from Kumari, Madurai, Nellai, Salem and Thanjai regions.

10.2 DATASET PREPARATION FOR TAMIL

For the selected regions, Kongu, Kumari, Madurai, Nellai, Salem and Thanjai. LDC-IL prepared the following dataset by which the prompt sheets were prepared.

Content Type	Count
Creative Text	8
Date	2
Command and Control Words	370
Most Frequent Words	1000
Form and Function Words	598
Phonetically Balanced Words	565
Person Name	572
Place Name	348
Sentences	228

Table 126: LDC-IL Speech Dataset

Distinct News Items were prepared to get the audio recording of contemporary text. It was made sure that each selected news item had minimum 500 words. Each prompt sheet had a distinct news item and sleeted part of the dataset prepared as follows.

Content Type	Content that Each Typical Prompt Sheet had	Content Selection Type
Contemporary Text	1 Text	Distinct Text
Creative Text	1 text	Random Text selected from dataset*
Sentences	25 Sentences	Random set selected from dataset*
Command and Control Words	30 Words	Random set selected from dataset*
Person Names	20 Words	Random set selected from dataset*
Place Names	10 Words	Random set selected from dataset*
Most Frequent Words	30 Words	Random set selected from dataset*
		*randomly selected by machine

Table 127: Table of Contents in LDC-IL Dataset

The Full Set of

1. Phonetically Balanced Vocabulary
2. Form and Function Words
3. 1000 Most Frequent Words

were also carried to the field to get recorded by selected individuals. Once all these preparations were made, the investigator started collecting the data.

The Collection of data is carried out in three phases for different regions. First phase carried out in Madurai, Salem, Thanjai region in 2008 by S.Thennarasu. In Kumari, Madurai, Nellai region in 2009 by L R Prem Kumar. In Kongu, Madurai, Salem region from 2009 by R Prabagaran. Some of the data are collected by D Srikanth in 2018

10.3 TRANSLITERATIONS IN LDC-IL TAMIL READ CORPUS

For easy reference and uniformity, the recorded text in the metadata file, is transliterated from Tamil to Roman letters. Numeric characters were transliterated from Tamil to Hindu-Arabic system.

The LDC-IL transliteration scheme of Tamil to Roman is given below.

LDC-IL Transliteration Scheme
Tamil characters to Roman and Tamil Numerals

Vowels and Vowel Signs											
ஆ	இ	ஈ	உ	ஊ	எ	ஏ	ஐ	ஒ	ஓ	ஔ	ஃ
ா	ி	ீ	ு	ூ	ெ	ே	ை	ொ	ோ	ௌ	
A	i	I	u	U	e	E	ai	o	O	au	H

Consonants									
க	ங	ச	ஞ	த	ண	ட	ந	ப	ம
ka	ng'a	ca	nj'a	ta	Na	Ta	na	pa	ma

ய	ர	ற	ல	வ	ள	ழ	ன
ya	ra	Ra	la	va	La	Za	n'a

Tamil Grantha Consonants			
ஜ	ஷ	ஸ	ஹ
ja	sha	sa	ha

Numerals (Tamil to Hindu-Arabic)									
০	ক	২	৩	৪	৫	৬	৭	৮	৯
0	1	2	3	4	5	6	7	8	9

10.4 SUMMARY OF THE CORPUS

In the sections below, we provide the tabular details of the different content types of the Tamil raw speech corpus based on various yardsticks which can also be filter out from the metadata sheet. 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 data size is of total duration 139:11:41 (hh:mm:ss) comprising 60,287 audio segments.

10.5 SUMMARY OF THE AUDIO SEGMENTS

The table below shows the total number of Audio Segments and their distribution in the Tamil speech dataset.

LDC-IL Tamil Speech Data Status	Gender →	Female			Male		
	Age Group →	16-20 Years	21-50 Years	50+ Years	16-20 Years	21-50 Years	50+ Years
Content Type	Total Segments	Segments	Segments	Segments	Segments	Segments	Segments
Contemporary Text (News)-T1	433	22	144	48	18	145	56
Creative Text-T2	429	22	143	48	18	143	55
Sentence-S	10764	552	3586	1199	449	3602	1376
Date-D	842	44	279	92	36	283	108
Command and Control Words-W1	12882	660	4291	1406	538	4341	1646
Person Name-W2	8755	457	2907	938	365	2985	1103
Place Name-W2	4002	204	1330	455	159	1310	544
Most Frequent Word-Part-W3A	12813	662	4213	1408	539	4341	1650
Most Frequent Word-FullSet-W3B	2000	0	1000	0	0	1000	0
Phonetically Balanced-W4	3860	564	563	435	991	871	436
Form and Function Word-W5	3507	593	1194	592	0	535	593

Table 128: Audio Segments and their Distribution

10.6 DURATION OF THE RAW SPEECH DATA

The table below shows the duration of each of the content type and their distribution across a few factors.

LDC-IL Tamil Speech Data Status	Gender →	Female			Male		
	Age Group →	16-20 Years	21-50 Years	50+ Years	16-20 Years	21-50 Years	50+ Years
Content Type	Total Duration	Duration (hh:mm:ss)	Duration (hh:mm:ss)	Duration (hh:mm:ss)	Duration (hh:mm:ss)	Duration (hh:mm:ss)	Duration (hh:mm:ss)
Contemporary Text (News)-T1	58:01:47	03:08:11	21:03:01	04:54:00	02:14:05	21:17:56	05:24:34
Creative Text-T2	14:21:31	00:40:55	04:36:11	01:35:17	00:47:15	05:00:29	01:14:24
Sentence-S	14:51:03	00:46:12	05:01:18	01:24:44	00:37:10	05:15:49	01:45:50
Date-D	01:20:17	00:03:54	00:26:01	00:08:07	00:03:05	00:28:34	00:10:36
Command and Control Words-W1	12:57:06	00:40:59	04:22:40	01:14:16	00:32:53	04:30:28	01:35:50
Place Name-W2	03:57:29	00:11:59	01:19:08	00:24:55	00:09:59	01:19:15	00:32:13
Person Name-W2	10:34:38	00:35:04	03:31:43	01:02:22	00:26:55	03:39:47	01:18:47
Most Frequent Word-Part-W3A	11:14:05	00:35:39	03:44:22	01:02:48	00:29:47	03:58:46	01:22:43
Most Frequent Word-FullSet-W3B	02:26:05	00:00:00	01:14:45	00:00:00	00:00:00	01:11:20	00:00:00
Phonetically Balanced-W4	04:55:10	00:45:27	00:38:05	00:31:13	01:11:46	01:09:15	00:39:24
Form and Function-Word-W5	04:40:29	00:53:17	01:37:21	00:47:00	00:00:00	00:43:42	00:39:09

Table 129: Duration of the Collected Data

10.7 DISTINCT SET

The Distinct Set usually contains data which is distinct to each speaker and is rarely repeated. The LDC-IL speech data set contains newspaper extracts which are read by each speaker.

10.7.1 The Contemporary Text (News) - T1

Distinct Text Extracts from Newspapers are recorded from the informants to get the speech data of contemporary text. The distribution of data is as follows:

Age Group	16 to 20		21 to 50		Above 51		Total Female	Total Male	Total
Gender	Female	Male	Female	Male	Female	Male			
Kongu Tamil	5	2	48	38	9	6	62	46	108
Kumari Tamil	2	4	9	9	2	2	13	15	28
Madurai Tamil	2	0	25	23	12	7	39	30	69
Nellai Tamil	3	5	13	14	8	14	24	33	57
Salem Tamil	0	2	3	15	15	25	18	42	60
Thanjai Tamil	10	5	46	46	2	2	58	53	111

Table 130: Distribution of Contemporary Text (News) Data

10.8 RANDOM SET

The Random Set data comprises of content types which are sampled by machine for each speaker. They are sampled from collection of master data sets available. The random sets are given below.

10.8.1 The Creative Text-T2

One randomly selected text of literature out of 6 texts from the prepared dataset is recorded from the informants to get the speech data of Creative text. The distribution of data is as follows.

Age Group	16 to 20		21 to 50		Above 51		Total Female	Total Male	Total
Gender	Female	Male	Female	Male	Female	Male			
Kongu Tamil	5	2	48	37	9	6	62	45	107
Kumari Tamil	2	4	9	9	2	2	13	15	28
Madurai Tamil	2	0	23	23	12	7	37	30	67
Nellai Tamil	3	5	13	13	8	14	24	32	56
Salem Tamil	0	2	3	15	15	24	18	41	59
Thanjai Tamil	10	5	47	46	2	2	59	53	112

Table 131: Distribution of Tamil Creative Text

10.8.2 The Date-D

The answer to one randomly selected question from the list of 2 questions to get the date format of the informants. The distribution of data is as follows:

Age Group	16 to 20		21 to 50		Above 51		Total Female	Total Male	Total
Gender	Female	Male	Female	Male	Female	Male			
Kongu Tamil	10	4	92	74	16	12	118	90	208
Kumari Tamil	4	8	18	18	4	4	26	30	56
Madurai Tamil	4	0	47	44	24	14	75	58	133
Nellai Tamil	6	10	24	25	16	26	46	61	107
Salem Tamil	0	4	6	30	30	48	36	82	118
Thanjai Tamil	20	10	92	92	2	4	114	106	220

Table 132: Distribution of Tamil Date Format

10.8.3 The Sentences-S

The Sentences contain a list of sentences that is a representation of all most all the phonemes occurring in Tamil. 25 Randomly selected Sentences are recorded from a list of 228 sentences. The distribution of data is as follows:

Age Group	16 to 20		21 to 50		Above 51		Total Female	Total Male	Total
Gender	Female	Male	Female	Male	Female	Male			
Kongu Tamil	127	49	1186	950	224	150	1537	1149	2686
Kumari Tamil	50	100	225	225	50	50	325	375	700
Madurai Tamil	50	0	600	575	300	177	950	752	1702
Nellai Tamil	75	125	324	326	200	350	599	801	1400
Salem Tamil	0	50	75	374	375	599	450	1023	1473
Thanjai Tamil	250	125	1176	1152	50	50	1476	1327	2803

Table 133: Distribution of Tamil Sentences

10.8.4 Command and Control Words-W1

The Command and Control Words contain a list of 370 words that is a representation of all most all the command and control words occurring in Tamil. 30 randomly selected words are recorded from the list. The distribution of data is as follows:

Age Group	16 to 20		21 to 50		Above 51		Total Female	Total Male	Total
Gender	Female	Male	Female	Male	Female	Male			
Kongu Tamil	148	59	1409	1139	240	180	1797	1378	3175
Kumari Tamil	60	120	270	270	60	60	390	450	840
Madurai Tamil	60	0	720	691	360	210	1140	901	2041
Nellai Tamil	90	150	361	391	240	420	691	961	1652
Salem Tamil	0	60	90	449	446	716	536	1225	1761
Thanjai Tamil	302	149	1441	1401	60	60	1803	1610	3413

Table 134: Distribution of Tamil Command and Control Words

10.8.5 Person Names –W2

The Person Names contain a list of 572 popular pan- Indian and regional person names. 20 randomly selected names are recorded from the list. The distribution of data is as follows:

Age Group	16 to 20		21 to 50		Above 51		Total Female	Total Male	Total
Gender	Female	Male	Female	Male	Female	Male			
Kongu Tamil	100	38	930	760	159	120	1189	918	2107
Kumari Tamil	40	80	181	180	40	40	261	300	561
Madurai Tamil	40	0	490	475	240	141	770	616	1386
Nellai Tamil	60	100	256	261	159	280	475	641	1116
Salem Tamil	0	40	61	301	296	479	357	820	1177
Thanjai Tamil	217	107	989	1008	44	43	1250	1158	2408

Table 135: Distribution of Tamil Person Names

10.8.6 Place Names-W2

The Place Names contain a list of 347 popular pan- Indian and regional place names. 10 randomly selected names are recorded from the list. The distribution of data is as follows:

Age Group	16 to 20		21 to 50		Above 51		Total Female	Total Male	Total
Gender	Female	Male	Female	Male	Female	Male			
Kongu Tamil	50	20	458	380	80	60	588	460	1048
Kumari Tamil	20	40	90	90	20	20	130	150	280
Madurai Tamil	20	0	230	216	120	72	370	288	658
Nellai Tamil	30	40	131	130	76	136	237	306	543
Salem Tamil	0	20	29	149	143	239	172	408	580
Thanjai Tamil	84	39	392	345	16	17	492	401	893

Table 136: Distribution of Tamil Place Names

10.8.7 Most Frequent Words-PART-W3A

The Most Frequent Words-part contains a list of 1,000 most frequent words. 30 randomly selected words are recorded from the list. The distribution of data is as follows:

Age Group →	16 to 20		21 to 50		Above 51		Total Female	Total Male	Total
Gender →	Female	Male	Female	Male	Female	Male			
Kongu Tamil	152	60	1373	1137	239	180	1764	1377	3141
Kumari Tamil	60	120	270	271	60	60	390	451	841
Madurai Tamil	60	0	719	690	360	210	1139	900	2039
Nellai Tamil	90	149	391	408	240	420	721	977	1698
Salem Tamil	0	60	90	450	449	720	539	1230	1769
Thanjai Tamil	300	150	1370	1385	60	60	1730	1595	3325

Table 137: Distribution of Tamil Most Frequent Words – Part

10.9 FULL SET

The Full Set is the master set of certain data set which is read completely from few selected speakers in each group. Full sets are as below.

10.9.1 Most Frequent Words-Full-W3B

The Most Frequent Words contain a list of 1000 most frequent words. In full set all the 1000 words are recorded from the informant. The distribution of data is as follows:

Age Group →	16 to 20		21 to 50		Above 51		Total Female	Total Male	Total
Gender →	Female	Male	Female	Male	Female	Male			
Thanjai Tamil	0	0	1000	0	0	1000	1000	1000	2000

Table 138: Distribution of Tamil Most Frequent Words – Full

10.9.2 The Phonetically Balanced Vocabulary-W4

The Phonetically Balanced Vocabulary contains a list of words where almost all the phones of Tamil language have occurred in all the possible positions of at least one of the words present in the list. In full set all the 565 words are recorded from the informant where they uttered those words three times. The distribution of data is as follows:

Age Group	16 to 20		21 to 50		Above 51		Total Female	Total Male	Total
Gender	Female	Male	Female	Male	Female	Male			
Thanjai Tamil	564	991	563	871	435	436	1562	2298	3860

Table 139: Distribution of Tamil Phonetically Balanced Vocabulary

10.9.3 The Form and Function Words-W5

The Form and Function Words contain a list of 598 words which is a representation of all most all the form and function words occurring in Tamil. All the words are recorded from the informant where they uttered those words three times. The distribution of data is as follows:

Age Group →	16 to 20		21 to 50		Above 51		Total Female	Total Male	Total
Gender →	Female	Male	Female	Male	Female	Male			
Salem Tamil	0	0	598	0	0	0	598	0	598
Thanjai Tamil	593	0	596	535	592	593	1781	1128	2909

Table 140: Distribution of Form and Function words

10.10 SPEAKER DISTRIBUTIONS

The following table shows the distribution of native speakers of Tamil, across different regions, who contributed to the speech data as part of this release.

Age Group →	16 to 20		21 to 50		Above 51		Total Female	Total Male	Total
Gender →	Female	Male	Female	Male	Female	Male			
Kongu Tamil	5	2	48	38	9	6	62	46	108
Kumari Tamil	2	4	9	9	2	2	13	15	28
Madurai Tamil	2	23	25	0	12	7	39	30	69
Nellai Tamil	3	5	13	14	8	14	24	33	57
Salem Tamil	0	2	4	15	15	25	19	42	61
Thanjai Tamil	12	7	51	51	4	4	67	62	129

Table 141: Distribution of Tamil Native Speakers

11 MULIT-LINGUAL RAW SPEECH CORPUS

Rajesha N., Manasa G., Narayan Choudhary

11.1 INTRODUCTION

The LDC-IL Multi-Lingual Raw Speech Corpus dataset is extracted from the raw speech corpora published by LDC-IL in various Indian languages. The Multi-Lingual speech dataset sampling is taken from the content type of 'Creative Text-T2' There are three age groups selected from the LDC-IL datasets. They are, '16 to 20 years', '21 to 50 years' and 'above 50 years'. For more details about how the data is collected from the field, coverage, etc., please refer the overview of speech corpora (Choudhary, et.al, 2019) and specific language documentation available at the LDC-IL Data distribution portal (https://data.ldcil.org). This dataset is built to address the needs of some applications like language identifier modules where multiple language samples are a requirement, to explore cross-linguistic variations and diatopic comparison to determine what generalizations are possible about the types of variable features, to build multilingual phoneme set and models etc.

11.2 CONTENT TYPE

This Multi-Lingual speech dataset sampling is taken from the content type of 'Creative Text-T2'. 'Creative Text-T2' is extracted mainly from literary sources. It is used to capture literary terms. Creative Texts are Stories or Essays collected from books. It exhibits the language style of the period from which the text is taken.

The creative text of the LDC-IL Speech dataset comprises of essays or short stories. One of these essays or short stories, selected randomly from a data set, is assigned to a speaker for reading out. The same story may be read out by multiple speakers.

11.3 TECHNICAL SPECIFICATIONS

The LDC-IL data is recorded using Roland EDIROL Recorder. It is a 24-bit Linear PCM (R-09) Recorder, at the sample rate is 48.0 KH, with 16 bit WAV recording mode. The audio segments are recorded using rechargeable batteries or alkaline batteries.

11.4 TEXT - SPEECH MAPPING AND NAMING CONVENTIONS

The collected data is segmented and mapped with its corresponding text and metadata. Each recording is named in accordance with its metadata information like Language Name, Speaker id, Content id, Gender, Age, Content type etc.

The Naming convention of LDC-IL Multi-Lingual Raw Speech Corpus dataset is as follows:
LDC_IL_Scheduled_<Language>_<Gender>_<Age Group>_<Content Type>_<Speaker ID>_<ContentID>

A Typical LDC-IL naming convention for Speech corpus is shown bellow.

LDC-IL_Scheduled_Bodo_Female_16To20_Creative Text-T2_SP-0021_T2-0004.wav
LDC-IL_Scheduled_Bodo_Female_16To20_Creative Text-T2_SP-0021_T2-0004.txt

Where .wav extension represents the audio and the .txt represents corresponding metadata file.

11.5 METADATA

The value of speech data can be determined according to the quality of the metadata obtained. It is imperative to maintain metadata of the entire data collection for linguistic analysis. A brief of each of these 25 fields/legends is given in the table below:

SL	Legend	Description
1	Language	Name of the Language
2	SpeakerID	Each speaker has a unique identity language. However, this is within the language. If one is working on speech corpus from more than one language, the IDs may get repeated.
3	ContentType	This corresponds to the notation of the content types noted above.
4	ContentID	This corresponds to the ID of the text being read out.
5	Gender	Denotes gender, whether it is male, female, or other.
6	AgeGroup	Three age groups of 16 to 20, 21 to 50, and 50+
7	Dialect	Notes the dialect of the language. An attempt has been made to cover all the dialects of the language as agreed upon in the academia of the language experts.
8	ReadInScript	The script in which the content has been provided to read in.
9	RecordingEnvironment	A brief info on the environment in which the recording has been done.
10	PowerSource	The source of the power using which the recording was done. It may be Li-ion, NiCd or Alkaline batteries.
11	RecorderManufacturer	Manufacturer of the recorder.
12	RecorderType	Type of the recorder. It is mostly 24-bit Linear PCM (R-09).
13	SamplingFrequency	Sampling frequency. It is mostly 48.
14	BitPerSample	Bit per sample. It is mostly 16-bit.
15	Channel	How many channels. All LDC-IL data is stereo. Only data set is mono which is segregated and constitutes a separate dataset of its own.
16	State	Name of the Indian state/province to which the speaker belongs to.
17	District	Name of the Indian district to which the speaker belongs to.
18	Place	Name of the place to which the speaker belongs to.
19	MotherTongue	Mother tongue of the speaker. It is note that data has been taken from people who professo to speak the language. However, it may be that the speaker uses the target language as a second or third language. However, if the speaker confidently says (and it is also verified by other speakers of the community), some samples have been taken from these types of users as well. This adds to the variety of the speech data collected.
20	EducationalQualification	Highest educational qualification of the speaker.
21	PlaceOfElementaryEducation	Place of the elementary education. This usually corresponds to the early childhood experiences which happen to more than often affect the way a language spoken.
22	RecordingDate	Date when the recording took place.
23	Investigator	Name of the Investigator.
24	RecordedText	Text of the recorded speech (in the script of the language).
25	TextInRoman	Text of the recorded speech (in the Roman transliteration as per the LDC-IL transliteration scheme. If the text is long (as is the case with T1 and T2 content

		types), a reference of the corresponding file is given.)

Table 142: Metadata Legends and their Description

11.6 TRANSLITERATION SCHEME

The recorded text is provided in the native language as well as in the transliterated format of Roman (except for Kashmiri and Urdu) The transliteration schema for each language are given in the LDC-IL transliteration schema document available at https://ldcil.org/Tools/CorporaToolsPackage/LDC-IL%20Transliteration%20Scheme.pdf.

11.7 SUMMARY OF THE CORPUS

The LDC-IL multi-lingual raw speech has 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 data size is of 62.2 GB with the total duration 97:43:54 (hh:mm:ss) comprising 1,916 Speakers.

The table below shows the distribution of each language in terms of total number of Speakers, Size and Duration in LDC-IL Multi-Lingual Raw Speech Corpus

Language	Female			Male			Total		
	Duration (hh:mm:ss)	Speakers	Size (in GB)	Duration (hh:mm:ss)	Speakers	Size (in GB)	Duration (hh:mm:ss)	Speakers	Size (in GB)
Assamese	2:33:40	68	1.64	2:34:33	64	1.65	5:08:13	132	3.30
Bengali	2:38:34	56	1.59	2:47:32	61	1.69	5:26:06	117	3.29
Bodo	2:30:39	42	1.61	2:41:04	40	1.72	5:11:43	82	3.34
Dogri	1:16:44	30	0.84	1:35:00	31	1.01	2:51:44	61	1.84
Gujarati	2:32:10	45	1.63	2:30:40	42	1.61	5:02:50	87	3.25
Hindi	2:37:28	44	1.66	2:30:18	44	1.57	5:07:46	88	3.23
Kannada	2:37:06	45	1.68	2:32:50	48	1.63	5:09:56	93	3.32
Kashmiri	2:32:26	30	1.63	2:39:46	29	1.71	5:12:12	59	3.34
Konkani	2:50:24	62	1.82	2:41:25	62	1.74	5:31:49	124	3.57
Maithili	2:46:28	54	1.71	2:53:31	50	2.00	5:39:59	104	3.48
Malayalam	2:38:16	68	1.69	2:28:17	61	1.59	5:06:33	129	3.29
Manipuri	2:15:42	29	1.45	2:44:43	32	1.76	5:00:25	61	3.22
Marathi	2:38:26	56	1.70	2:41:57	58	1.73	5:20:23	114	3.43
Nepali	2:51:09	44	1.83	2:58:41	52	1.91	5:49:50	96	3.75
Odia	2:38:24	63	1.70	2:32:10	60	1.63	5:10:34	123	3.33
Punjabi	2:41:13	67	1.72	2:35:40	62	1.66	5:16:53	129	3.40
Tamil	2:35:24	78	1.57	2:45:20	70	1.66	5:20:44	148	3.24
Telugu	2:06:18	24	1.33	3:00:40	38	1.93	5:06:58	62	3.27
Urdu	2:20:22	53	1.50	2:48:54	54	1.81	5:09:16	107	3.31

Table 143: Summary of Multi-Lingual Raw Speech Corpus