Linguistic Resources for AI/NLP in Indian Languages

2019

Central Institute of Indian Languages

Department of Higher Education, Ministry of Human Resource and Development, Government of India,

Manasgangotri, Mysore

VICE –PRESIDENT OF INDIA

Message

Language is much more than a means of communication. It is essence of every civilization, the reflection of its history, its culture, its traditions and its evolution. For languages to survive and thrive, they have to be continually enriched and nourished.

We live in an age of Information Technology, where technology and human lives are inextricably interwoven. We must make use of the infinite possibilities offered by information technology to safeguard nature and promote our languages. To achieve this objective, it is imperative to develop interfaces between technology and languages. A lot more needs to be done to improve the technological support available to Indian Languages.

The resources required to develop language technology and artificial intelligence based tools have not been readily available for Indian languages. To fill this gap, the Government of India launched the scheme of Linguistic Data Consortium for Indian Languages (LDC-IL) in 2008 and has been preparing high quality linguistic resources since then in all the scheduled languages of India.

I congratulate the Central Institute of Indian Languages (CIIL) for its efforts in bringing out 31 large text and speech datasets in 19 scheduled Indian languages. It is good to know that around 50 more datasets with more fine-grained annotations are also set to be released within this year. More updates and new datasets will keep on coming afterwards that will help capture in-depth structures of these languages. These datasets are the largest corpora for these languages available so far in the public domain.

I am very happy to launch the Data Distribution Portal of LDC-IL (http://data.ldcil.org) which will provide a seamless, easy and quick way of requesting and availing various types of datasets helpful for the development of Natural Language Processing (NLP) and Artificial Intelligence based technologies in Indian languages, including technologies such as automatic dictation, speech recognition, language understanding, machine translation, grammar and spelling checks and so on.

I am also happy to know that these datasets are available for free to the academic and not-for-profit research organizations. The release of these resources marks the beginning of new era for the availability of cutting edge IT tools in Indian languages. Thereby filling the digital divide by breaking the language barrier in the digital domain.

M. Venkaiah Naidu

Linguistic Resource For AI/NLP in Indian Languages
Editor: Narayan Choudhary

Publication No.: 1196

First Published: April, 2019
Chaitra, 1941

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 978-81-7343-295-8

Price: 250/-

Published by : Prof. D. G. Rao, 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 : Sri. H. Manohara, Artist
Printed at: CIIL, Printing Press, Mysuru

TABLE OF CONTENTS

FOREWORD

D. G. Rao, Director, CIIL

Since past three decades, information technology has been the buzzword at the national and international level. Exchange of information happens at the click of a button now. The world is moving at a much faster rate than it used to do just 20 years back. All this has happened because tele-communication is made easy, instant and accessible to all.

While we talk of communication, the first mode of it is the natural, human language in its various forms as text, speech, sign language or other modes. Language is not only mode of communication, it is also an identity for human race.

It is an irony that the issue of language endangerment has taken prominence along the same time when the information technology took prominence in the human lives. This is a clear indication that information technology has a role to play in the rather faster pace of language endangerment being witnessed at present across the globe.

It is evident that the problem caused by IT will also find solution in IT. The default language for IT based tools have been English (of the US variant, to be more specific) for several decades. Support for other languages came very late and has remained still so less that default language for IT professionals have become English across the globe. One cannot expect to be become an IT professional without having a fair or working knowledge of English because all of the computer programming languages use English as their primary language. There is no programming language compiler that can support any text other than ASCII (a short for American Standard Code for Information Interchange, a language encoding system that supports only English alphabets). These are some of the advances that have been towards the hegemony of English language all over the world and it has established English language as a common international language.

But this has also caused an irreparable damage to other languages of the world by way making all other languages of the world secondary to English.

India is home to hundreds of languages with several languages having tens of millions of speakers. IT support for these languages have been negligible until recently. Some support that started in the meanwhile have been meagre even though India is on the path of becoming a digital economy and pushing hard towards the digital eco-system.

It is understood that major challenge faced before the software developer communities in the IT sector is lack of resources in the Indian languages. The content on the internet has been very low or negligible that has put a restriction on the developers community towards providing support for more and more of Indian languages.

Linguistic Data Consortium for Indian Languages (LDC-IL), a scheme of the Government of India implemented by Central Institute of Indian Languages was started to create such linguistic resources that will provide impetus towards development of higher level language technologies in Indian languages.

In the last ten years, the scheme has developed largest resources in almost all of the scheduled languages of India that contain text and speech corpora as well as higher level linguistic annotations on them.

The task of creating the text corpora in 18 languages (i.e. Bengali, Bodo, Dogri, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Maithili, Malayalam, Manipuri, Nepali, Marathi, Odia, Punjabi, Tamil, Telugu and Urdu) that are being released now has been an uphill one. Even though all of the above languages are major languages of India, having more than 5 million speakers (with Hindi being even the third most spoken languages of the world), the electronic content in these languages have different kinds of bottlenecks.

While for some languages we have got almost nil or negligible text, for a few others, the text data have been sampled but some cleaning is required before it is ready for release. The corpus generated are mostly typed as representative text in various domains are not available in electronic format. The text are also proof-read to give it a cleaner look and make it readily useful for people working on the real world applications.

These raw corpora will be helpful in creating various language models for different types of language technologies including lexicon generation, grammatical structure modelling, concordancers, spelling corrections and so on.

13 speech corpora, including the languages of Bengali, Bodo, Hindi, Kannada, Konkani, Maithili, Malayalam, Manipuri, Marathi, Punjabi, Telugu, Urdu and Nepali, are being released at this time. More languages are getting readied and will be made available on the LDC-IL data distribution portal very soon.

This is to note that for most the languages, there are no corpora available at all and the LDC-IL initiative will be the first initiative towards the electronic resource creation in these languages. For the languages where the corpora are available, the release of the LDC-IL datasets is going to boost it in a big way as this would be largest corpora in its segment to be released for these languages.

Apart from the raw corpora, LDC-IL has also invested in Parts of Speech Annotation in most of the above mentioned languages and a fair size of PoS annotation has already been done. Some validation works are required after which these datasets will also be released.

Only raw speech datasets are part of the release at present and the sentence level annotation and word level annotations are also being readied for release in the near future.

One of the bottlenecks that we surpassed in the last two years is to get copyright issues resolved as most of the text that are part of the text corpora are extracted via sampling method from published

books which are often copyrights of other organizations including public and private ones. Thousands of letters seeking permission from authors/respective copyright holders were written across India and elsewhere and copyrights were sought. Despite the 2016 order of the Supreme Court on this issue, this was considered as a bottleneck and prevented the Institute from releasing the datasets to the general public. With a pathbreaking decisions being taken, it is hoped that this will pave the way for new datasets in more, smaller languages being created with lesser hassles.

The Project Advisory Committee sat twice in 2018 to finalize the licensing and pricing policies and we are glad this has finally been completed and approved. This is further going to prove another path-breaking decision in the government that will create a milestone towards the development of language technologies in Indian languages.

As the data portal is ready and licensing policies finalized, we hope that there will be a further impetus towards meeting the needs of the technology development community at a much faster rate than ever for Indian languages. The data portal will also provide a platform for other institutes/bodies who want to commercialize or distribute their datasets/resources through this portal and help promote Indian languages in the IT field with greater fervour.

1 LDC-IL RAW TEXT CORPORA: AN OVERVIEW

Narayan Choudhary, L. Ramamoorthy

1.1 INTRODUCTION

This is a generic documentation of the LDC-IL raw text corpus which applies to all the languages covered in LDC-IL unless otherwise specified. However, this does not give the specifics of a language dataset.

The objective of language technology is to utilize the facilities of computer, to scientifically analyze language for retrieving verifiable proofs about properties of a language that enable the understanding of multi-dimensional nature of a language. Corpus of a language reflects the nature of the language. The larger and the more representative a corpus, the better it shows its nature.

A corpus is a large collection of language manifestation duly representing its aspects, mainly in text or spoken form. In case of sign language it is the collection of signs in visual form. The electronic text corpus is a collection of pieces of language text in electronic form, selected in accordance with the external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research. Corpora are one of the major resources for language technology. Computers offer advantages like searching, selecting, sorting and formatting, which eases the language studies. Computers can avoid human bias in an analysis, thus making the result more reliable. Corpus serves as the basis for a number of research tasks within the field of Corpus Linguistics. It is the main resource for many modules of various applications like grammar checkers, spell checkers used in word editors etc. Indian languages often pose difficult challenges for developer community in Natural Language Processing/Artificial Intelligence. The technology developers building mass-application tools/products have for long been calling for availability of linguistic data on a large scale. However, the data should be collected, organized and stored in a manner that suits different groups of technology developers.

Over the years, a lot of efforts have been made to develop text corpora in Indian languages and several agencies have made contributed towards this including the government organizations, academic institutions as well as private bodies. However, the constant greed of more and more electronic data as required by the contemporary machine learning oriented technology models have proved that the data is still not sufficient for all the scheduled languages of India.

Linguistic Data consortium for Indian Languages (LDC-IL) is one of the Government of India initiatives to develop linguistic corpora in Indian languages. Approved as a scheme in 2007 by the Ministry of Human Resource & Development, Government of India, LDC-IL started functioning at Central Institute of Indian Languages (CIIL), Mysore from April 15, 2008 when human resources got recruited for this scheme. The mission statement for this project is to develop "***Annotated, quality language data (both-text & speech) and tools in Indian Languages to Individuals, Institutions and Industry for Research & Development - Created in-house, through outsourcing and acquisition[1].***"

[1] Extract from the *Detailed Project Report* of LDC-IL.

The text datasets created under the LDC-IL ambit strives to fill the gap and provide more and more of electronic data for the NLP and language technology community such that the Indian languages get a boost and more of IT applications are available in these languages.

1.2 LDC-IL APPROACH OF SAMPLING

Developing a written text corpus involves various factors like size of corpus, representativeness, quality of the text, determination of target users, selection of time-span, selection of documents etc. The data for the LDC-IL corpus are collected from books of general interest, textbooks, magazines, newspapers and Government documents of the contemporary text. The data is collected in accordance with prior set of criteria and with the convenience of material such as availability, proper format etc.

As a corpus is supposed to be representative of the language, there is no need to collect all the text from a given book. The representativeness of the corpus depends on a range of different kinds of text categories included in the corpus. LDC-IL corpora try to cover a wide range of text categories that could be representative of the language or language variety under consideration. Corpus representativeness and balance is closely associated with sampling.

LDC-IL collected text corpus from different sources. They are mainly books, magazines, and newspapers. The books are from literature and knowledge text books, magazines and newspapers are web crawled, or keyed in text or both.The newspaper and magazines are great resource of words which are hard to find in books because of the scarcity of those domain specific books in Indian languages.

LDC-IL has different Sampling approach over while extracting text from these three sources.

1.2.1 Sampling Approach for Books

The books were identified so that the representation of different domains can be catered.After identifying the books,the next step is to extract typically 10 pages of text from it. LDC-IL follows a sampling method to collect the pages from a book. For example, if the book has 100+ pages we collect every 10^{th} page and if the book has 200+ pages we collect every 20^{th} page of the book. If the selected page contains pictures, tables etc, then its next or previous page, which may have the text content, will be chosen for the corpus. Even though one may find rare cases where partial or whole book is selected for the corpus, since the total corpus is going to be very large, such rare cases may not have an impact on balance of corpus. While selecting the book, the LDC-IL's motive is to select from wide variety of domains so that corpus can cover large part of vocabulary and should not miss out certain domain specific words.

Other generic principles that have been normally followed in the sampling tasks across languages are as follows:

- Contents containing obnoxious or vulgar texts have been avoided.
- New editions of the old books having a writing style prior to 1990 were not preferred. Rarely we may have text extracts from such books published prior to 1990 to ensure that the writing style is contemporary.
- For all texts containing short stories, sampling has been made by considering the short stories as a single entity and not based on the whole book containing all the short stories i.e. each page starting with a new short story have been sampled instead of the usual sampling method based on page numbers of the book.

- The data sampling personnel carried the category and sub-category list for ready reference in the field.
- Text extracts containing poems and formulae have been avoided.
- Pages containing diagrams, tables or figures have been avoided.
- Books containing less than 50 pages are not part of sampling.
- Texts having very small font have been enlarged during photocopying to make it look like 10 to 12 font size.
- If the text contains content other than the intended language, those texts have been avoided if the other language content is longer than one sentence.

1.2.2 Sampling Approach for Magazines

In case of magazine textsare small and from different domains so the whole magazine is to be considered to be included in corpus discarding advertisements, image captions, and tables etc. Magazine corpus usually includes different types of texts like cookery, health, cinema, stories, contemporary articles, etc.

1.2.3 Sampling Approach for newspaper

The newspaper corpus is contemporary text in nature. The text may contain political news, editorials, sports news etc.The news data does not have literary flourish. The news stories are on many unfamiliar domains, religious ideas, scientific principles etc. that have to be conveyed to the common people. So, it is expected that the writerswould have captured these domains in a simple and meaningful way. Such write-ups have proper usage of vocabulary, correct language structure and effective phraseology. The newspaper articles may use colloquial, non-standard terms or jargons to attract the readers. The words used need to be expressive and represents the feeling and attitude towards the events. To cover such nuance of the language the newspaper are sampled to be part of the text corpus.

The News items of the paper is sampled based on the domains, classifieds, very small news snippets were avoided. Usually much of the newspaper is keyed.

1.3 LDC-IL TEXT CORPUS CATEGORIZATION

The LDC-IL corpus shows how people naturally use the language and it does not give imaginary, idealized examples. To satisfy this requirements we needed large amount of data otherwise the frequent items will be from some specific vocabulary or a particular style. Quantitative data gives somewhat accurate results of what occurs frequently and what occurs rarely in the language.

Each text source of corpus is different from others in form, function, content and features.This gives room to classify corpora into different categories. LDC-IL maintains a standard list of categories for which the text is to be collected. LDC-IL Identifies six major categories namely '*Aesthetics*', '*Commerce*', '*Mass Media*', '*Official Document*', '*Science and Technology*', '*Social Sciences*'. These categories are further classified into 128 minor categories or sub-categories to cover various domains.

1.3.1 Aesthetics

The Aesthetics category is one of the largest contributors to the LDC-IL corpus. This category contains sub-domains from Literature and Fine-arts. The text extracts are from literary sources. It is used to capture literature terms. Aesthetics text is collected from collected from books. The text is probably any standard

text which is descriptive in nature. It exhibits the language style of a particular period from which the text is taken. It is an extract of creative writing. It is made up of stories based on fiction, essays on various topics etc. These write-ups are mostly self-expressions of the writer. It captures the flow of language of the writer of the literary text.

The subdomains that are identified for mark-up in corpus under the Aesthetics is given below:

Aesthetics				
Fine Arts-Dance	Literary Texts	Literature-Novels	Autobiographies	Folk Tales
Fine Arts-Drawing	Literature-Criticism	Literature-Plays	Biographies	Folklore
Fine Arts-Hobbies	Literature-Diaries	Literature-Poetry	Cinema	Mythology
Fine Arts-Music	Literature-Essays	Literature-Epics	Culture	Photography
Fine Arts-Sculpture	Literature-Letters	Literature-Speeches	Handicrafts	Humour
Fine Arts-Musical Instruments	Literature-Children'sLiterature	Literature-Text Books (School)	Literature-Travelogues	Literature-Science Fiction
				Literature-Short Stories

Table 1-1: Subcategories of the Category Aesthetics

1.3.2 Commerce

The trade is a part of the society. It exists and operates in association with various groups in society such as customer, suppliers, competitors, banks and financial institutions, Government agencies, trade unions. The trade domain has many domain specific words which need to be part of the corpus. The trade related books will bring such texts to the corpus.

The Subdomains that are identified for mark-up in corpus under the Commerce is given below:

Commerce					
Industry	Accountancy	Share Market	Banking	Business	Career and Employment
Management	Finance	Tourism			

Table 1-2: Subcategories of the Category Commerce

1.3.3 Mass Media

Media is an integral part of everyday life for many people all over the world, at work and in the home. The text from this domain is contemporary in nature. The text may contain political news, editorials, or sports news. The major source of the Mass Media text category is newspaper; it contains words which are used in day-to-day life. Structurally, the language of mass media contains exposition, argument, description and narration. It includes different types of write up; consists of structures with different patterns, words and styles.All this is written in a language in which everyone can relate and understand. Some of the media prints are in the form of conversation or question answers. This data usually contains an interviewer and an interviewee .They usually consist dialogues. The interviewee may be a celebrity or a renowned personality from cinema, politics etc. The words used in such text are usually more personal and simple.

The Subdomains that are identified for mark-up in corpus under the Mass Media is given below:

Mass Media					
Article	Classifieds	General News	Obituary	SMS	Religious/Spiritual News
Business News	Discussions	Interviews	Political	Social	Sports News
Cinema News	Editorial	Letters	Speeches	Weather	Health

Table 1-3: Subcategories of the Category Mass Media

1.3.4 Official Document

The usage of language in official documents is highly standard, unambiguous, straight forward and structurally modified. The communication intended in official documents are intended about some action, or some enquiry or proceedings of some assemblies. This text usually it is to get the due representation of such domain specific terminologies of administration, official document category is included.

The Subdomains that are identified for mark-up in corpus under the Official Document is given below:

Official Document			
Administration	Legislature	Parliamentary/Assembly Debates	Police Documents

Table 1-4: Subcategories of the Category Official Documents

1.3.5 Science and Technology

The science and technology domain contains text extracts from various scientific books, articles of magazines, journals etc. These texts are also called as knowledge texts. The language structure and usage of words are different from the language of day-to-day life. The terminologies that are from this domain will have highest number of loan words because the subject in the text is usually global. To get the due representation of such domain specific terminologies, the Science and Technology category is included.

The Subdomains that are identified for mark-up in corpus under the Science and Technology is given below:

Science and Technology					
Agriculture	Biotechnology	Engineering-Civil	Forestry	Medicine	Statistics
Architecture	Botany	Engineering-Electrical	Geology	Micro Biology	Astrology
Textile Technology	Educational Psychology	Engineering-Electronics Communication	Text Book (Science)	Computer Sciences	Language Technology
Chemistry	Naturopathy	Engineering-Mechanical	Horticulture	Oceanology	Veterinary
Ayurveda	Criminology	Engineering-Others	Astronomy	Physics	Film Technology
Bio Chemistry	Homeopathy	Environmental Science	Logic	Psychology	
Biology	Yoga	Engineering-Chemical	Mathematics	Sexology	Zoology

Table 1-5: Subcategories of the Category Science and Technology

1.3.6 Social Sciences

Language is a medium for creation and maintenance of human society so language in social sciences category correlates the linguistic features of the dynamic society. Human development and reformation happening in different communalcontext hence all the social knowledge and reality could be reflected in this text category.

The Subdomains that are identified for mark-up in corpus under the Social Sciences is given below:

Social Sciences					
Anthropology	Food and Wellness	Personality Development	Physical Education	Text Book (Social Science)	Philosophy
Archaeology					Journalism
Demography	Fisheries	Library Science	Law	Sports	Geography
Economics					Religion / Spiritual

Education	Home	Political	Public	Health and Family	Sociology
Epigraphy	Science	Science	Administration	Welfare	Linguistics

Table 1-6: Subcategories of the Category Social Sciences

1.4 LDC-IL TEXT DATA ENCODING AND FORMAT

The collected data should be encoded in a machine readable form for further analysis. While storing the data one has to keep some standards so that the data is easy to store and retrieve in long term.The encoding being used in LDC-IL Text corpus is Unicode and stored in XML format.Large scale language resource depends on the metadata. Metadata is an authentic source to prove the quality of the data. Metadata should have the subject information, source information and encoding information.

The selected text along with metadata information is indexed with a five digit unique number to get keyed-in. Each text fragment of selected book is typed as corpus file with xml extension. The given unique Index number gets prefixed with the LDC-IL notations which make the filename of the XML file. Sometimes the XML file names carry small case alphabets enclosed in braces. This is done if the book title carries different type of textual topics, so that each chapter, in the selected book title which may be related to different topics, chapters etc., can be differentiated. This helps the text content get categories based on the context.

1.5 LDC-IL TEXT CORPUS METADATA

It is imperative to maintain metadata of the entire data collection for linguistic analysis.The collected data are arranged with its metadata information such as its category, subcategory, title of the text, author name, source, publisher name, year of publication, page numbers etc. This information helps the users to retrieve the data easily from the database/repository. Metadata gives authenticity to the text by way of providing the details of how the data was created in the first instance and what is its content about. The following table shows the legend used in the metadata and provides description of them.

#	Legend	Description
1	Filename	Represented by "docID" tag in the XML files. This is a unique file number across the datasets.
2	ProjectDescription	This gives a brief of the project under which the file was generated. As CIIL has been involved into corpus creation over a long period time, including before the inception of LDC-IL scheme, there might be some data for a few languages which might have come from different projects e.g. the CIIL Corpus or CIIL-KHS corpus. This field indicates the source of the project.
3	SamplingDescription	This information is a verifiable proof for the corpus. It will have the information of selected page numbers of the book for corpus.
4	Category	Specifies the domain of the text.
5	Subcategory	Specifies the sub-domain of the text.
6	Text	Specifies the type of the source text i.e. whether its origin is a book, a magazine or a newspaper.
7	Title	Specifies the title of the source text. It contains mostly books but if magazines or newspapers occur, their respective are provided here.
8	Volume	Specifies volume number the title, if any.
9	Issue	Specifies issue number the title, if any.
10	TextType	Is mostly blank however sometimes it is used to provide the broad topic of the news items e.g. whether it is a political news or editorial or sports news etc.

11	Headline	This information is a verifiable proof for the corpus. This is normally the heading of the chapter of the selected sample. Gives the fine tuned information of the topic present in particular file.
12	Author	Specifies the name of the author.
13	Editor	Specifies the name of the editor.
14	Translator	Specifies the name of the translator.
15	Words	Specifies the total number of words in the file.
16	Letters	Specifies the total number of UTF8 characters in the file.
17	PublishingPlace	Specifies the place where the title was published.
18	Publisher	Specifies the name of the publisher.
19	PublishedYear	Specifies the publishing year.
20	Index	Is the index number or ID of the file. It is noted inside the XML file. It is mostly the same as the file name.
21	Date	Date when the file was digitized/inputted.
22	Input	Name of the Data Inputter, if the file has been typed.
23	Proof	Name of the Proof reader.
24	Language	Name of the language.
25	Script	Name of the script the text is written in.

Table 1-7: Metadata Legends for LDC-IL Text Data

Typical Metadata Mark-ups in a text corpus file structure is given below.

<?xml version="1.0" ?>			
<?xml-stylesheet type="text/css" href="home.css"?>			
<Doc id="mal-w-media-	ML00172	"	lang="Malayalam">
<Header type="text">			
<encodingDesc>			
<projectDesc>	CIIL-Malayalam Corpora, Monolingual Written Text		</projectDesc>
<samplingDesc>	Simple written text only has been transcribed. Diagrams, pictures and tables have been omitted. Samples taken from page 30-31,50-51,70-71,94-95,114-115,132-133,152-153,172-173,192-193,210-211		</samplingDesc>
</encodingDesc>			
<sourceDesc> <biblStruct> <source>			
	<category>	Aesthetics	</category>
	<subcategory>	Literature-Novel	</subcategory>
	<text>	Book	</text>
	<title>	Kalapam	</title>
	<vol>		</vol>
	<issue>		</issue>
</source>			
<textDes>			
	<type>		</type>
	<headline>		</headline>
	<author>	ShashiTharoor	</author>
	<ediotr>		</editor>
	<translator>	Thomas George	</translator>
	<words>	2745	</words>
</textDes>			
<imprint>			
	<pubPlace>	India-Kottayam	</pubPlace>

	<publisher>	DC Books	</publisher>
	<pubDate>	2006	</pubDate>
</imprint>			
<idno type="CIIL code">		Kerala University Campus Library- 13535	</idno>
<index>		ML00172	</index>
</biblStruct> </sourceDesc>			
<profileDesc> <creation>			
	<date>	26-Apr-2010	</date>
	<inputter>	Remya K	</inputter>
	<proof>		</proof>
</creation>			
<langUsage>		Malayalam	</langUsage>
<ScriptUsage>		Malayalam	</ScriptUsage>
<wsdUsage>			
<writingSystem id="ISO/IEC 10646"> Universal Multiple-Octet Coded Character Set (UCS). </writingSystem>			
</wsdUsage>			
<textClass>			
<channel mode="w">		Print	</channel>
<domain type="public">			</domain>
</textClass> </profileDesc> </Header>			
<text> <body>			
</text> </body> < /Doc>			

1.6 LDC-IL TEXT CORPUS AND NAMING CONVENTIONS

The selected hardcopies were marked for sampling and given to typists by concerned language experts.LDC-IL has built an in-house corpus developing application and stores it in a repository database. The samples get typed in xml format through a software application built for it in LDC-IL. Each sampling is a corpus file and gets typed and saved in Unicode standards. Each corpus file has unique filename. One can say the corpus is indexed through filenames.Typically each corpus file is an extract of a book of a particular title. The LDC-IL corpus file name follows certain naming convention. The naming convention is based on language and source of text.Every scheduled language has a notation for each kind of source of corpus. The notation is prefixed to a five digit number to create a unique corpus filename.

The LDC-IL notations for Indian Scheduled languages are given below.

#	Language	ISO 639Language Code	Script	Notation as per Source of Corpus			
				Book	Magazine	News Paper	News Web
1	Assamese	asm	Assamese	AS	ASM	ASN	ASNW
2	Bengali	ben	Bengali	BE	BEM	BEN	BENW
3	Bodo	brx	Devanagari	BD	BDM	BDN	BDNW
4	Dogri	doi	Devanagari	DG	DGM	DGN	DGNW
5	Gujarati	guj	Gujarati	GJ	GJM	GJN	GJNW
6	Hindi	hin	Devanagari	HN	HNM	HNN	HNNW
7	Kannada	kan	Kannada	KA	KAM	KAN	KANW
8	Kashmiri	kas	Persio-Arabic	KS	KSM	KSN	KSNW
9	Konkani	kok	Devanagari	KO	KOM	KON	KONW
10	Maithili	mai	Devanagari	MT	MTM	MTN	MTNW
11	Malayalam	mal	Malayalam	ML	MLM	MLN	MLNW
12	Manipuri	mni	Bengali/MeeteiMayek	MN	MNM	MNN	MNNW

13	Marathi	mar	Devanagari	MA	MAM	MAN	MANW
14	Nepali	nep	Devanagari	NP	NPM	NPN	NPNW
15	Odia	ori	Odia	OD	ODM	ODN	ODNW
16	Punjabi	pan	Gurmukhi	PN	PNM	PNN	PNNW
17	Sanskrit	san	Any Script	SA	SAM	SAN	SANW
18	Santali	sat	OlChiki	SN	SNM	SNN	SNNW
19	Sindhi	snd	Persio-Arabic / Devanagari	SI	SIM	SIN	SINW
20	Tamil	tam	Tamil	TA	TAM	TAN	TANW
21	Telugu	tel	Telugu	TE	TEM	TEN	TENW
22	Urdu	urd	Persio-Arabic	UR	URM	URN	URNW

Consider the example of Malayalam,The text taken from Malayalam book for LDC-IL Malayalam Text Corpus always starts with 'ML' followed by 5 digit numbers which is continuous, where as text collected from Malayalam Magazine starts with 'MLM' followed by 5 digit numbers. If the source is from Newspaper then 'MLN' notation will be followed where as if the News is taken from Web source 'MLNW' will be used as notation.

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

Such filenames could be 'ML00001(a)', 'ML00001(b)', 'ML00001(c)', 'ML00001(d)' etc.

1.7 PROOF READING

Once it is in digital form, the same is proofread so that it is free from any kind of typographical errors. Proofing is the next process of corpus building. Since the typed corpus may carry errors because of various reasons like speed of the typist and typist not belonging to the language community, the proofing is done by the language experts.

While proofing of a corpus file is done in LDC-IL, the following things are taken care of

1. Removing the poetic text, if any poem or poetic structure occurs within the running text
2. If there are incomplete sentences typed (generally at the end of the paragraph) the sentence is removed up to the logical ending of the previous sentence.
3. Verifying the difference between the visargaha and colon ' : ' symbol, and to ensure that the correct symbol/punctuation is used in the correct place.
4. During Content cleaning focus stays on the corrections of typographical errors and spacing. If there is a space preceding a punctuation mark, space is removed, unless it is there in the actual text itself (i.e. hard copy of the text).
5. If there is any mismatch between the hard copy and the input corpus file, it is ensured that the corpus file should be faithful to hard copy.
6. It is ensured that the Title, Author, Headline fields of the XML files is written in Roman using the LDC-IL transliteration scheme. The LDC-IL Transliteration scheme can be referred on the LDC-IL website. Also, the LDC-IL transliteration tool from Roman to Indian Scripts and vice versa is available for download on the LDC-IL website.
   Link to download LDC-IL Transliteration Scheme:

http://ldcil.org/Tools/CorporaToolsPackage/LDC-IL%20Transliteration%20Scheme.pdf

Link to download the LDC-IL Transliteration Tool (.exe file):
http://ldcil.org/Tools/LDC-IL%20Transliterator.zip

Proof reading is used to correct clear cases of spelling mistakes, splitting sentences or words, removing unnecessary repeated paragraphs, sentences, phrases, words. Moreover, it includes removing unwanted texts from the corpus such as foreign script sentences and incorrect use of ungrammatical sentences.

1.8 COPYRIGHT

Anyone intending to put together a corpus for commercial purposes must always obtain the permission from the publishers of the source texts. Many commercially available corpora contain texts from a large number of sources and obtaining permission to use these can be a very cumbersome and financially costly process. However, LDC-IL took up the task and managed to get the consent of most of the copyright holders or has at least communicated to them that the text extracts from their sources are being used in the language sampling task which may also be used commercially.

Considering LDC-IL is a government initiative taken up in the larger public interest and the corpus is used for the development of language, most of the publishers and authors generously agreed to archive the samples of their text materials in corpus. Some of the authors even suggested and offered their other content which are not yet part of the LDC-IL corpus. Government publishers too expressed no objections regarding since LDC-IL itself is an initiative of Govt. of India.Private publishers also gave permission considering that LDC-IL is only using a part of a text, and it will not harm their business anyway. LDC-IL thanks all of them for the co-operation.

For some of the content where we have not yet got the explicit consent of the copyright holders, we have sent them the letters asking for the same. If any of the copyright holders disagree to consent, they may write so to us and their respective text will be removed from the sampling corpus and the same will be intimated to all the license holders of the respective dataset and they will have to abide by it.

2 BENGALI RAW TEXT CORPUS

Sonali Sutradhar, Rajesha N., Manasa G., Narayan Choudhary, L. Ramamoorthy

2.1 INTRODUCTION

Bengali also known by its endonym Bangla belongs to the Indo-Aryan language family spoken in South Asia. It is one of the scheduled languages of India. Bengali is the official language of the states of West Bengal, Tripura. This is widely spoken in the Kachar district of Assam. Bengali is usually counted as the seventh most spoken native language in the world by population. It is a matter of pride that Bengali is the only language for which a whole movement happened in the name of "Bhasha Andolan" on 21st February, 1952. The movement reached its climax when police killed student demonstrators on that day. The deaths provoked widespread civil unrest. In 1999, UNESCO declared 21 February as International Mother Language Day, in tribute to the Language Movement.

Bengali Script with the modern Bengali alphabet has undergone a long evolution cycle. A large number of ancient epigraphic records and manuscripts have been discovered from different parts of Bengal, which have supplied important information to reconstruct the historical origin and development of the modern Bangla alphabet. Bengali script is historically derived from the ancient Indian Brahmi. The modern Bengali alphabet was derived from the Northern class of Brahmi script. But in course of time the Northern class of Brahmi had turned into two separate branches, the Eastern variety and the Western variety. It is partly syllabic and partly alphabetic. It has close similarity to the Assamese script except two alphabets. The alphabet in Bengali script follows the same pattern of arrangement found in Devanagari script. It is written from left to right manner. Movements of the strokes comprising different symbols are also mostly from left to right. There are no capitals, and the punctuation system is almost wholly taken from English. The only difference is to mark the end of a sentence the symbol danda '।' is used instead of a dot '.'.

Bengali text corpus is collected from various libraries in West Bengal mostly from Kolkata. The greater part of the text has been taken from CIIL library and National Library. LDC-IL tried to cover the entire category in its standard list. Some categories like novel, short stories have huge amount of books but some categories like physics, chemistry, economics have very less amount of books. Literary texts are easily available in Bengali but getting scientific text is very difficult. Some categories like epigraphy, finance, oceanology text are too rare in Bengali.

2.2 PECULIARITIES OF BENGALI TEXT

The Corpus of Bengali 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 that are narrative and it contains elements of fiction.

Novels, short stories, plays are examples of literary text. Non-literary texts are texts 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.

2.2.1 khanda ta 'ৎ' of Bengali

Khanda ta is a letterform used in Bengali for a consonant 'ta' without vowel. It has the same phonological value as ta-hasanta, though usage conventions for these differ from each other. Early in the 20th century ta-hasanta was preferred for indigenous Bengali words (those derived from Prakrit) in contexts, in which conjunct forms would occur for loans from Sanskrit, Persian or other languages. Khanda ta originated, apparently, as an alternate way to write ta-hasanta in such contexts.

In the earlier versions of Unicode Khanda ta did not have a separate value and were represented with the combination of Zero Width Joiner (ZWJ)

BENGALI KHANDA ta 'ta' – 'ৎ'	ta (ত) + hasant/halant (্) + ZWJ

Zero Width Joiner (ZWJ) character has no value of its own. Thus, this issue could lead us to have more character count than the actual character count. But as Khanda 'ta' did not have separate value it was necessary for the data to keep it like ta (ত) + hasant/halant (্) + ZWJ followed by the next consonant.

To counter this inconsistency Unicode allotted separate code for Khanda 'ta'. LDC-IL Bengali text data is on par with the current Unicode standards of Khanda 'ta'. LDC-IL Bengali text data contains standard Khanda 'ta'.

2.2.2 YA-FALLA in Bengali

For borrowed words like 'অ্যাক্ট' (act), 'অ্যাকশন' (action), 'এ্যান্ড' (and) after vowels YA-FALLA has to be added. But in the earlier version of Unicode it was not there, so there was always an issue of typing this kind of words. But this was introduced in the later version.

Though, YA-FALLA is perferctly attaching after vowels, but still this issue is unresolved for particularly one consonent which is 'র'(ra). If we try to put YA-FALLA after 'র'(ra), it will come as র্য or REPH + য (ya). Hence, inputting words like Rank, Ragging, Racket in corpus is still an unreloved issue for Bengali.

2.3 DATA SAMPLING NOTES

2.3.1 Principles of Data Sampling

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

2.3.2 Field Works Undertaken

Bengali text corpus is collected from various libraries from West Bengal, mostly from Kolkata. The text materials were collected by conducting three field works undertaken in the period from 2007 to 2012. The greater part of the text has been taken from CIIL Library, ERLC Library, National Library, Asiatic Society and Sahitya Akademi Library, Kolkata. Some corpus was collected from some local libraries.

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

- CIIL Library, Central Institute of Indian Languages, Mysore.
- ERLC Library, Eastern Regional Language Center, Bhubaneshwar.
- National Library, Kolkata.
- Asiatic Society, Kolkata.
- Sahitya Akademi Library, Kolkata.
- Mrinalini Dutta Mahavidyapith Library.

Collected text materials have been published at various places within West Bengal and other states of India such as Tripura, Delhi as well as other countries such as Bangladesh etc.

An attempt has been made to cover the entire category in its standard list. Some categories like novel, short stories have huge amount of books but some categories like physics, chemistry, economics has very less amount of books. Literary texts are easily available in Bengali but getting scientific text in Bengali is very difficult. Some categories like forestry, criminology, botany text are too rare in Bengali.

Collecting text data from the field is a difficult job. Most of the libraries do not allow to take huge amount of text from their shelves at a time because it is against their rules and principles. Specially in National Library nobody can have direct access to the books. For a particular period, they issue a maximum three or four books. Even if the librarian allowed to take many books at a time, the photocopy kiosk had issues as there was a long queue.

Some time photocopy attendant refused to photocopy randomly selected pages because of the long queue waiting and it takes up more time for them to turn the pages compared to continuous page photocopying they are accustomed to. It was another issue that the field worker/linguist had to carry a huge list of photocopy bundles with them which was many a times cumbersome to travel with.

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

2.3.3 Data Inputting

All the text has been typed in Unicode using the InScript Keyboard directly onto the XML files. The data has been inputted by Ms. Papiya Das, Ms. Tumpa Dutta Gupta and Ms. Rina Sarkar, who are the native speakers of Bengali.

2.3.4 Validation and Normalization Workshops

A 5-day workshop was conducted at Linguistic Data Consortium from 23-August-2010 to 03-September-2010 with Dr. Niladri Sekhar Dash (ISI-Kolkata), Mr Atanu Saha (New Delhi), Ms. Baidehi Sengupta (Kolkata) and Mr. Tanmay Bir (Kolkata) as experts. All the experts suggested that the Bengali text corpus should remain true to the text.

2.3.5 Proofreading

Bengali text data has been proofread by internal resource persons and as well as by workshop resource person. The text has always been kept true to the printed material and typos, if any, occurring at the time of typing have only been corrected. In the process of being true to the printed material some issue always arise, which is spelling variations. When the spelling is misprinted then it is corrected at the time of inputting. But when the lexicon has already spelling variation in the language, then we kept both of the variation for the purpose of differnet usage of text corpus. The printed materials collected for the corpus is contemporary , mainly published after 1990 such that we can capture the contemporary essence of the language. We tried to cover almost all the Bengali authors who was actively giving their masterpieces to the language.

2.4 TRANSLITERATIONS IN LDC-IL BENGALI TEXT CORPUS

For easy reference and uniformity of metadata, some entries in the metadata file, namely *'Title', 'Headline', 'Author', 'Editor', 'Translator'* are transliterated from Bengali to Roman letters. Numeric characters were transliterated from Bengali to Hindu-Arabic system.

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

LDC-IL Transliteration Schema
Bengali characters to Roman and Bengali 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	র	ল	শ	স	ষ	হ	ড়	ঢ়	য়	ৎ
ya	ra	la	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

2.5 COPYRIGHT CONSENTS

The Bengali text corpus have been collected from various sources and the copyright for the same stays with different sources. However, for the purposes of this corpus, consents have been sought from all the stakeholders. Most of the copyrights (around 93%) belong to private parties with only 7% belonging to the government agencies, either state or the central.

2.6 OVERVIEW OF REPRESENTED DOMAINS

LDC-IL Bengali Text Corpus size is: 42,37,440 Words drawn from 1,460 different titles. Bengali Corpus character size is 2,54,81,407. The following table gives a summary of the typed and cleaned text of the Bengali Raw Text Corpus. The representation of the three major domains covered has been shown in the table below:

Domain	Domain Word Count	Percentage
Aesthetics	4037854	95.29%
Science and Technology	76231	1.80%
Social Sciences	123355	2.91%
Total	4,237,440	100

Table 2-1 Representation of the Domains in Bengali Text Corpus

As each domain has several sub-domains, the following table shows the representation of the several domains, both within the domain and across all the domains.

2.6.1 Aesthetics

The aesthetics category of Bengali text corpus covers 24 sub-categories bearing a total of 40,37,854 words along with the overall percentage of 95.29%. The representational details are given in the table below.

#	Sub Domain	Word Count	Percentage within Subdomain	Overall Percentage
1	Autobiographies	116683	2.89%	2.75%
2	Biographies	79141	1.96%	1.87%
3	Culture	2184	0.05%	0.05%
4	Fine Arts-Drawing	308	0.01%	0.01%
5	Fine Arts-Music	9738	0.24%	0.23%
6	Fine Arts-Sculpture	1387	0.03%	0.03%
7	Folk Tales	2969	0.07%	0.07%
8	Folklore	2158	0.05%	0.05%
9	Handicrafts	664	0.02%	0.02%
10	Humour	27637	0.68%	0.65%
11	Literary Texts	83657	2.07%	1.97%
12	Literature-Children's Literature	17709	0.44%	0.42%
13	Literature-Criticism	239115	5.92%	5.64%
14	Literature-Diaries	4986	0.12%	0.12%
15	Literature-Essays	172729	4.28%	4.08%
16	Literature-Letters	2590	0.06%	0.06%
17	Literature-Novels	2222825	55.05%	52.46%
18	Literature-Plays	51704	1.28%	1.22%
19	Literature-Poetry	336	0.01%	0.01%
20	Literature-Science Fiction	2436	0.06%	0.06%
21	Literature-Short Stories	857850	21.25%	20.24%
22	Literature-Text Books (School)	16665	0.41%	0.39%
23	Literature-Travelogues	119329	2.96%	2.82%
24	Mythology	3054	0.08%	0.07%
	Total	4037854	100%	95.29%

Table 2-2 : Aesthetics Category Representation

2.6.2 Sceience and Technology

The Science And Technology category of Bengali text corpus covers 11 sub-categories bearing a total of 76,231 words along with the overall percentage of 1.80%. The representational details are given in the table below.

#	Sub Domain	Word Count	Percentage within Subdomain	Overall Percentage
1	Astronomy	6200	8.13%	0.15%
2	Ayurveda	5586	7.33%	0.13%
3	Botany	2759	3.62%	0.07%
4	Chemistry	2931	3.84%	0.07%
5	Criminology	720	0.94%	0.02%
6	Film Technology	26649	34.96%	0.63%
7	Forestry	3321	4.36%	0.08%
8	Medicine	2603	3.41%	0.06%
9	Naturopathy	5593	7.34%	0.13%
10	Physics	5689	7.46%	0.13%
11	Psychology	14180	18.60%	0.33%
	Total	76231	100%	1.80%

Table 2-3: Science and Technology Category Representation

2.6.3 Social Sciences

The Social Sciences category of Bengali text corpus covers 9 sub-categories bearing a total of 1,23,355 words along with the overall percentage of 2.91. The representational details are given in the table below.

#	Sub Domain	Word Count	Percentage within Subdomain	Overall Percentage
1	Anthropology	2474	2.01%	0.06%
2	Economics	4372	3.54%	0.10%
3	Health And Family Welfare	59357	48.12%	1.40%
4	History	21608	17.52%	0.51%
5	Home Science	4114	3.34%	0.10%
6	Philosophy	14431	11.70%	0.34%
7	Political Science	12807	10.38%	0.30%
8	Religion/Spiritual	2365	1.92%	0.06%
9	Sociology	1827	1.48%	0.04%
	Total	123355	100	2.91%

Table 2-4: Social Sciences Category Representation

3 BODO RAW TEXT CORPUS

Mansoor Khan, Farson Daimary, Bridul Basuamtary, Rajesha N, Manasa G, Narayan Choudhary, L. Ramamoorthy

3.1 INTRODUCTION

Bodo is a language that belongs to the branch of Barish section under Baric division of the Tibeto-Burman language and spoken by the Bodo people of North-Eastern India and some parts of Nepal. The language is one of the official languages of the Indian state of Assam, and is one of the 22 scheduled languages that are given a special constitutional status in the year 2003. The language is closely related to the Dimasa, Tiwa, Rabha languages of Assam, the Garo language of Meghalaya and the Kokborok language spoken in Tripura. The Bodo is the second major language of Assam and official language in the Bodo dominated areas. Many rivers like Dihing, Dibru, Dihong, Dikrai etc. in the North-East region were named after some Bodo words which reveals the spatial distribution pattern of related ethno-cultural groups.

The Bodos are one of the ethnic and linguistic communities and early settlers of Assam in North-East India. The word BODO means both the language as well as the community. The Bodos belongs to a larger group of ethnicity called the Bodo-Kachari. Mythologically, according to Dr. Suniti Kumar Chatterji, a well-known historian, they are "The offsprings of son of the Lord Vishnu and mother earth" who were called as 'Kiratas' during the epic period. Though they are Mongolian people, the Bodos come to North-East India in 2000 BC.

In the consequence of socio-political developing and movement launched by the Bodo organizations since 1913, the language was introduced as the medium of education (1963) in the primary schools in Bodo dominated areas. The Bodo language serves as a medium of education up to the secondary level and an associated official language in the state of Assam. The language has achieved a position of pride with the opening of the post-graduate course in Bodo language and literature in the University of Guwahati in 1996. At present, the post-graduate course in Bodo is introduced in the Bodoland University, Dibrugarh University and Cotton University of Assam. The Bodo language has to its credit large number of books of poetry, drama, short stories, novels, biography, travelogues, children's literature and literary criticism. Though the spoken language has been exaggerated by other communities, especially the Assamese, in and around Kokrajhar, it is still to be heard in its pure form, in and around Udalguri district.

In 1970, the Bodo Sahitya Sabha decided to approve roman script for the language in its 11th annual conference. The demand was raised before the Government of Assam till 1974, but was snubbed by the

government. The Bodo Sahitya Sabha then launched democratic movement from 12 September 1974. The movement saw the contribution by millions of general people and Bodo students. But unfortunately, the local Government of Assam conquered with strong hand resulting 16 peoples to death and many of the people to serious and minor injury. The movement was then called off on 13 February 1975, and Devanagari script was imposed on Bodos.

Bodo text corpus is collected from various libraries in Assam mostly from Kokrajhar, Chirang, Baksa, Udalguri and Guwahati. The greater part of the text has been taken from Kokrajhar, Chirang, Udalguri, Bodo Sahitya Sabha Library of Guwahati, Departmental library of the Department of Bodo, Guwahati University, and from some personal libraries. LDC-IL tried to cover the entire category in its standard list. Some categories like novel, short stories have huge amount of books but some categories like physics, chemistry, economics have very less amount of books. Literary texts are easily available in Bodo but getting scientific text is very difficult. Some categories like epigraphy, finance, oceanology text are too rare in Bodo.

3.2 PECULIARITIES OF BODO TEXT

The Corpus of Bodo 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 that are narrative and it contains elements of fiction. Novels, short stories, plays are examples of literary text. Non-literary texts are texts 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.

3.3 DATA SAMPLING NOTES

3.3.1 Principles of Data Sampling

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

3.3.2 Fieldworks Undertaken

Bodo text corpus is collected from various libraries in Assam, mostly from Bodoland Territorial Area District (BTAD) and the other parts of Assam. The text materials were collected by conducting four fieldworks undertaken in the period from 2010 to 2012. The greater part of the text has been taken from Kokrajhar Library, Udalguri Library, Bodo Sahitya Sabha Library and Guwahati University.

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

1. Kokrajhar, Assam.
2. Udalguri, Assam.
3. Guwahati University Library, Guwahati.
4. Bodo Sahitya Sabha Library, Guwahati
5. Personal Libraries from Kokrajhar, Chirang, Udalguri and Guwahati, Assam.

Collected text materials have been published at various places within Assam.

An attempt has been made to cover the entire category in its standard list. Some categories like novel, short stories have huge amount of books but some categories like physics, chemistry, economics have very less amount of books. Literary texts are easily available in Bodo but getting scientific text is very difficult. Some categories like epigraphy, finance, oceanology text are too rare in Bodo.

Collecting text data from the field is a difficult job. Most of the libraries do not allow to take 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 to take 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 takes up more time for them to turn the pages compared to continuous page photocopying they are accustomed to. It was another issue that the field worker/linguist had to carry a huge list of photocopy bundles with them which was many times cumbersome to travel with.

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

3.3.3 Data Inputting

All the text has been typed in Unicode using the InScript Keyboard directly onto the XML files. The data has been inputted by Ms. Mamatha and Ms. Bidya, who are the native speakers of Kannada.

3.3.4 Proofreading

Bodo text data has been proofread by internal resource persons and other resource persons who have been called by LDC-IL for short term program for 45 working days. The text has always been kept true to the printed material and typos, if any, occurring at the time of typing have only been corrected. The printed materials collected for the corpus is contemporary, mainly published after 1990.

3.4 TRANSLITERATIONS IN LDC-IL BODO TEXT CORPUS

For easy reference and uniformity of metadata, some entries in the metadata file, namely *'Title', 'Headline', 'Author', 'Editor', 'Translator'* are transliterated from Bodo to Roman letters. Numeric characters were transliterated from Bodo to Roman system.

The LDC-IL transliteration scheme of Devanagari to Roman and Numerals to Hindu-Arabic given below.

Vowels											
Vowel	अ	आ	इ	ई	उ	ऊ	ऋ	ए	ऐ	ओ	औ
Matra		ा	ि	ी	ु	ू	ृ	े	ै	ो	ौ
Key	a	A	i	I	u	U	x	E	Ai	O	au

Consonant					
Consonant	क	ख	ग	घ	ङ
Key	K	kh	g	gh	ng'

Consonant	च	छ	ज	झ	ञ
Key	C	ch	j	jh	nj'

Consonant	ट	ठ	ड	ढ	ण	ड़	ढ़
Key	T	Th	D	Dh	N	D'	Dh'

Consonant	त	थ	द	ध	न
Key	T	th	d	dh	n

Consonant	प	फ	ब	भ	म
Key	P	ph	b	bh	m

Consonant	य	र	ल	व	श	ष	स	ह
Key	Y	r	l	v	Sh	S	s	h

Consonant	श	ष	स	ह	
Key	Sh	S	s	h	

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

3.5 OVERVIEW OF REPRESENTED DOMAINS/CATEGORIES

LDC-IL Bodo Text Corpus size is: 29,15,544 Words and character count is 2,13,44,814 drawn from 78 different titles and 2 titles including the extracts from newspapers. The data can be categorized into two classes of typed+cleaned and crawled. The crawled data has been crawled mainly from news websites and archived using the standard processing of LDC-IL text corpus preparation.

The representation of the five major domains covered has been shown in the table below:

Domain	Word Count	Percentage
Aesthetics	474960	16.29%
Commerce	25064	0.86%
Mass Media	1679511	57.61%
Science and Technology	172151	5.90%
Social Sciences	563858	19.34%
Total	**29,15,544**	**100**

Table 3-1: Representation of the Domains in Bodo Text Corpus

As each domain has several sub-domains/sub-categories, the following table shows the representation of the several domains, both within the domains and across all the domains.

3.5.1 Aesthetics

The Social Science category of Bodo text corpus covers 13 subdomains bearing a total of 4, 74,960 words along with the overall percentage of 16.29%. The representational details are given in the table below.

Subdomain	Word Count	% (within Subdomain)	Overall Percentage
Biographies	2169	0.46%	0.07%
Cinema	72596	15.28%	2.49%
Culture	6441	1.36%	0.22%
Folklore	5836	1.23%	0.20%
Literary Texts	5102	1.07%	0.17%
Literature-Criticism	208362	43.87%	7.15%
Literature-Essays	21405	4.51%	0.73%
Literature-Letters	350	0.07%	0.01%
Literature-Novels	22992	4.84%	0.79%
Literature-Plays	18335	3.86%	0.63%
Literature-Short Stories	72813	15.33%	2.50%
Literature-Speeches	281	0.06%	0.01%
Literature-Text Books (School)	38278	8.06%	1.31%
Total	**474960**	**100.00%**	**16.29%**

Table 3-2: Aesthetics Category Representation

3.5.2 Commerce

The Commerce category of Bodo text corpus covers a subdomain bearing a total of 25,064 words along with the overall percentage of 16.29%. The representational details are given in the table below.

Subdomain	Word Count	% (within Subdomain)	Overall Percentage
Business	25064	100.00%	0.86%

Table 3-3: Commerce Category Representation

3.5.3 Mass Media

The Mass Media category of Bodo text corpus covers 10 subdomains bearing a total of 16,79,511 words along with the overall percentage of 57.61%. The representational details are given in the table below.

Subdomain	Word Count	% (within Subdomain)	Overall Percentage
Classifieds	761	0.05%	0.03%
Discussions	288409	17.17%	9.89%
Editorial	65020	3.87%	2.23%
General News	1232689	73.40%	42.28%
Health	2261	0.13%	0.08%
Religious/Spiritual News	2022	0.12%	0.07%
Social	43104	2.57%	1.48%
Sports News	43219	2.57%	1.48%
Cinema News	507	0.03%	0.02%
Weather	1519	0.09%	0.05%
Total	**1679511**	**100.00%**	**57.61%**

Table 3-4: Mass Media Category Representation

3.5.4 Science and Technology

The Science and Technology category of Bodo text corpus covers 5 subdomains bearing a total of 1,72,151 words along with the overall percentage of 5.90%. The representational details are given in the table below.

Subdomain	Word Count	% (within Subdomain)	Overall Percentage
Agriculture	239	0.14%	0.01%
Astrology	6060	3.52%	0.21%
Engineering-Mechanical	1508	0.88%	0.05%
Environmental Science	1039	0.60%	0.04%
Text Book (Science)	163305	94.86%	5.60%
Total	**172151**	**100.00%**	**5.90%**

Table 3-5: Science and Technology Category Representation

3.5.5 Social Sciences

The Social Sciences category of Bodo text corpus covers 13 subdomains bearing a total of 5,63,858 words along with the overall percentage of 5.90%. The representational details are given in the table below.

Subdomain	Word Count	% (within Subdomain)	Overall Percentage
Economics	24774	4.39%	0.85%
Education	1326	0.24%	0.05%
Food and Wellness	13622	2.42%	0.47%
Health and Family Welfare	99942	17.72%	3.43%
History	66343	11.77%	2.28%
Law	902	0.16%	0.03%
Linguistics	2106	0.37%	0.07%
Personality Development	600	0.11%	0.02%
Political Science	11589	2.06%	0.40%
Public Administration	3590	0.64%	0.12%
Religion/Spiritual	3751	0.67%	0.13%
Sports	199423	35.37%	6.84%
Text Book (Social Science)	135890	24.10%	4.66%
Total	**563858**	**100.00%**	**19.34%**

Table 3-6: Social Science Category Representation

3.6 COPYRIGHT CONSENTS

The Bodo text corpus has been collected from various sources and the copyright for the same stays with different sources. However, for the purposes of this corpus, consent has been sought from all the stakeholders. Most of the copyrights (around 85%) belong to private parties with only 15% belonging to the government agencies, either state or the central.

4 DOGRI RAW TEXT CORPUS

Shahnawaz Alam, Sunil Kumar, Rajesha N, Manasa G, Narayan Choudhary, L. Ramamoorthy

4.1 INTRODUCTION

Dogri, is an Indo-Aryan Language spoken by about five million people in India and Pakistan, particularly in the Jammu region of Jammu and Kashmir and Himachal Pradesh, also in northern Punjab, other parts of Jammu and Kashmir. Dogri was originally written using the Dogri script which is very close to the Takri script.The language is now more commonly written in Devanagari in India, and in the Nastaʿliq form of Perso-Arabic in Pakistan and Pakistani-administered Kashmir.

Dogri has several varieties, all with greater than 80% lexical similarity (within Jammu and Kashmir). Before gaining language status, per the Census of India, Dogri was classified as one of the many varieties of Punjabi, such as Majhi or Doabi.

Western Pahari languages, Punjabi and Punjabi dialects are frequently tonal, which is very unusual for Indo-European languages (although Swedish and Norwegian are tonal also). This tonality makes it difficult for speakers of other Indo-Aryan languages to gain facility in Dogri, though native Punjabi speakers (especially speakers of Northern dialects such as Hindko and Mirpuri) may find it easier to make the transition.

Official recognition of the language has been gradual, but progressive. On 2 August 1969, the General Council of the Sahitya Academy, Delhi recognized Dogri as an "independent modern literary language" of India, based on the unanimous recommendation of a panel of linguists. (Indian Express, New Delhi, 3 August 1969).

In 2005, a collection of over 100 works of prose and poetry in Dogri published over the last 50 years was made accessible online at the **Central Institute of Indian Languages** (CIIL), Mysore. This included works of eminent writer Dhinu Bhai Panth, Professor Madan Mohan Sharma, B.P. Sathai and Ram Nath Shastri.

Dogri text corpus is collected from various libraries in Jammu and Kashmir, mostly from Jammu. The greater part of the text has been taken from library of Department of Dogri, Jammu University, Jammu University Library, J&K Academy of Arts, Culture and Languages and Dogri Sansatha-Jammu

LDC-IL tried to cover the entire category in its standard list. Some categories like novel, short stories have huge amount of books but some categories like physics, chemistry, economics have very less amount of books. Literary texts are easily available in Dogri but getting scientific text is very difficult. Some categories like epigraphy, finance, oceanology text are too rare in Dogri

4.2 PECULIARITIES OF DOGRI TEXT

The Corpus of Dogri 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 that are narrative and it contains elements of fiction. Novels, short stories, plays are examples of literary text. Non-literary texts are texts 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.

4.3 DATA SAMPLING NOTES

4.3.1 Principles of Data Sampling

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

4.3.2 Field Works Undertaken

Dogri text corpus is collected from various libraries in Jammu & Kshmir, mostly from Jammu. The text materials were collected by conducting one field work undertaken in the period from August – October 2010. The greater part of the text has been taken from Library of Department of Dogri, Jammu University, Jammu University Library and Dogri Sansatha-Jammu

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

- Library of PG Department of Dogri, University of Jammu, Jammu
- J&K Academy of Art, Culture and Languages, Jammu & Kashmir
- Dogri Sansatha-Jammu

Collected text materials have been published at various places within J&K and other states of India such as J&K, Himachal Pradesh, Delhi, Mumbai etc.

An attempt has been made to cover the entire category in its standard list. Some categories like novel, short stories have huge amount of books but some categories like physics, chemistry, economics have very less amount of books. Literary texts are easily available in Dogri but getting scientific text is very difficult. Some categories like epigraphy, finance, oceanology text are too rare in Dogri.

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

Some time Xerox attendents refused to photocopy randomly selected pages because of the long queue waiting and it takes up more time for them to turn the pages compared to continuous page photocopying they are accustomed to. It was another issue that the field worker/linguist had to carry a huge list of photocopy bundles with them which was many a times cumbersome to travel with.

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

4.3.3 Data Inputting

All the text has been typed in Unicode using the InScript Keyboard directly onto the XML files. The data has been inputted by Mrs. Rajeshwari.

4.3.4 Proofreading

Dogri text data has been proofread by internal resource persons. The text has always been kept true to the printed material and typos, if any, occurring at the time of typing have only been corrected.

The printed materials collected for the corpus is contemporary , mainly published after 1990. hence The text material available is with the reformed script which came into effect from 1969.

4.3.5 Validation and Normalization Workshops

A 45-day workshop was conducted at Linguistic Data Consortium from 19th Sept. to 31st Oct., 2013 with three resource persons from Jammu. The input data of Dogri text has been cleaned by these external resource persons as well as internal resource persons.

4.4 TRANSLITERATIONS IN LDC-IL DOGRI TEXT CORPUS

For easy reference and uniformity of metadata, some entries in the metadata file, namely *'Title', 'Headline', 'Author', 'Editor', 'Translator'* are transliterated from Dogri to Roman letters. Numeric characters are same as Roman.

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

LDC-IL Transliteration Scheme
Dogri characters to Roman

Vowels and Vowel Signs									
अ	आ	इ	ई	उ	ऊ	ए	ऐ	ओ	औ
	ा	ि	ी	ु	ू	े	ै	ो	ौ
a	A	i	I	u	U	e	ai	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		
य	र	ल	व	श	स	ह
ya	ra	la	va	sha	sa	ha

4.5 OVERVIEW OF REPRESENTED DOMAINS

LDC-IL Dogri Text Corpus size is: 8,01,771 Words and character count is 41,25,617 drawn from 183 different titles, including the extracts from newspapers.

The following table gives a summary of the typed and crawled text of the Dogri Raw Text Corpus.

Text Type	Word Count	KeyStroke/Character Count
Typed+Cleaned	8,01,771	4125617

Table 4-1 Representation of Typed and Cleaned Dogri Text Copus

The representation of the five major domains covered has been shown in the table below:

Domain	Word Count	Percentage
Mass Media	156,756	19.55%
Science & Technology	2,730	0.34%
Aesthetics	594,609	74.16%
Commerce	1,350	0.17%
Social Sciences	46,326	5.78%
Total	**8,01,771**	**100**

Table 4-2 Representation of the Domains in Dogri Text Corpus

As each domain has several sub-domains, the following table shows the representation of the several domains, both within the domain and across all the domains.

4.5.1 Mass Media

The Mass Media category of Dogri text corpus covers 5 sub-categories bearing a total of 156,756 words along with the overall percentage of 19.55%. The representational details are given in the table below.

Subdomain	Word Count	% (within Subdomain)	Overall Percentage
Discussions	947	0.604124%	0.12%
Editorial	74555	47.56118%	9.30%
General News	80828	51.56294%	10.08%
Letters	426	0.27176%	0.05%
Total	**156,756**	**100%**	**19.55%**

Table 4-3 Mass Media Category Representation

4.5.2 Science and Technology

The Science and Technology category of Dogri text corpus covers 1 sub-categories bearing a total of 2730 words along with the overall percentage of 0.34%. The representational details are given in the table below.

Subdomain	Word Count	% (within Subdomain)	Overall Percentage
Agriculture	2730	100%	0.34%

Table 4-4 Science and Technology Category Representation

4.5.3 Aesthetics

The Aesthetics category of Dogri text corpus covers 14 sub-categories bearing a total of 594,609 words along with the overall percentage of 74.16%. The representational details are given in the table below.

Subdomain	Word Count	% (within Subdomain)	Overall Percentage

Autobiographies	8758	1.472901%	1.09%
Biographies	34892	5.868058%	4.35%
Cinema	18740	3.151651%	2.34%
Culture	11972	2.013424%	1.49%
Fine Arts-Sculpture	2464	0.41439%	0.31%
Folklore	50178	8.438823%	6.26%
Humour	3536	0.594677%	0.44%
Literature-Criticism	32139	5.405065%	4.01%
Literature-Essays	121110	20.36801%	15.11%
Literature-Novels	85273	14.34102%	10.64%
Literature-Plays	77736	13.07347%	9.70%
Literature-Short Stories	138874	23.35552%	17.32%
Literature-Speeches	931	0.156573%	0.12%
Literature-Travelogues	8006	1.346431%	1.00%
Total	**594,609**	**100%**	**74.16%**

Table 4-5 Aeshthetics Category Representation

4.5.4 Commerce

The Commerce category of Dogri text corpus covers 1 sub-categories bearing a total of 1350 words along with the overall percentage of 0.17%. The representational details are given in the table below.

Subdomain	Word Count	% (within Subdomain)	Overall Percentage
Business	1350	100%	0.17%

Table 4-6 Commerce Category Representation

4.5.5 Socical Sciences

The Social Science category of Dogri text corpus covers 6 sub-categories bearing a total of 46,326 words along with the overall percentage of 3.99%. The representational details are given in the table below.

Subdomain	Word Count	% (within Subdomain)	Overall Percentage
Food and Wellness	582	1.256314%	0.07%
Health and Family Welfare	3846	8.302033%	0.48%
Linguistics	3673	7.928593%	0.46%
Religion/Spiritual	3610	7.7926%	0.45%
Sociology	2664	5.75055%	0.33%
Sports	31951	68.96991%	3.99%
Total	**46,326**	**100%**	**5.78%**

Table 4-7 Social Science Category Representation

4.6 COPYRIGHT CONSENTS

The Dogri text corpus have been collected from various sources and the copyright for the same stays with different sources. However, for the purposes of this corpus, consents have been sought from all the stakeholders. Most of the copyrights (around 49%) belong to private parties with only 51% belonging to the government agencies, either state or the central.

5 GUJARATI RAW TEXT CORPUS

Santosh Kumar Mohanty, Gadhavi Hirenkumar, Rajesha N, Manasa G, Narayan Choudhary, L. Ramamoorthy

5.1 INTRODUCTION

Gujarati is the principal and official language of Gujarat, union territories of Daman and Diu and Dadra and Nagar Haveli. It is recognized and taught as a minority language in the states of Rajasthan, Maharashtra, Madhya Pradesh, Tamilnadu and the union territory of New Delhi. Gujarati is one of the major languges of Indo-Aryan language family and it is written in Gujarati Script from left to right direction. This script is a variant of Devanagari script differentiated by the lack of horizontal line running top of the letters and by a number of modifications of some characters. LDC-IL Gujarati text corpus is collected in Gujarati script of contemporary usage.

Gujarati text corpus is collected from different libraries from Gujarat. The greater part of the text has been taken from Bhaikaka Library, Vidyanagar and Shrimati Hansa Mehta Library,Vodadara. LDC-IL tried to cover the entire domains/subdomains (categories/subcategories) in its standard list. Some subdomains like novel, short-story have huge amount of books but some subdomains like mythology, philosophy, cinema have very less amount of books. Literary texts are easily available in Gujarati but getting scientific/knowledge text is very difficult; even some subdomains like sports, homeopathy, epigraphy, finance, oceanology text are too rare in Gujarati.

5.2 PECULIARITIES OF GUJARATI TEXT

The Corpus of Gujarati text can be broadly classified into two types: 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 that are narrative and it contains elements of fiction. Novel, short-story, play are the examples of literary text. Non-literary texts are texts whose primary purpose is to convey knowledge/information. Example of non-literary texts are text about various scientific or technical subjects, articles/papers in academic journals. In literary text, language has creative elements, cultural information, dialectical variations and ambiguities etc. But technical or scientific terms, foreign words etc. have widely appeared in non-literary texts.

5.3 DATA SAMPLING NOTES

5.3.1 Principles of Data Sampling

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

5.3.2 Fieldworks Undertaken

Gujarati text corpus is collected from various libraries in Gujarat, mostly from Vodadara. The text materials were collected by conducting six fieldworks undertaken in the period from 2008 to 2012. The following resource persons were engaged in the fieldworks: Mona Parakh, Gadavi Hirenkumar and Purva S. Dholakia. The greater part of the text has been taken from the Bhaikaka Library, Sardar Patel University, Vidyanagar and Smt. Hansa Mehta Library, M.S. University of Baroda.

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

- Bhaikaka Library, Sardar Patel University, Vidyanagar
- Central Library, Gujarat Vidyapith, Ahmedabad
- Shrimati Hansa Mehta Library, M.S University of Baroda, Vodadara

Collected text materials have been published at various places within Gujarat and other states of India such as Karnataka, Tamilnadu, Maharashtra, Uttarakhand, Uttar Pradesh, New Delhi as well as other countries like UK and USA.

An attempt has been made to cover the entire domains and subdomains in its standard list. Some subdomains like novel, short-story have huge amount of books but some subdomains like cinema, weather, philosophy have very less amount of books. Literary texts are easily available in Gujarati but getting scientific/knowledge text is very difficult. Some subdomains like epigraphy, finance, oceanology text are too rare in Gujarati.

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

Some time Xerox attendants refused to photocopy randomly selected pages because of the long queue waiting and it takes up more time for them to turn the pages compared to continuous page photocopying they are accustomed to. It was another issue that the fieldworker had to carry a huge list of photocopy bundles with them which was many a times cumbersome to travel with.

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

5.3.3 Data Inputting

All the text has been typed in Unicode compatible font using the InScript Keyboard directly into the XML files. The data has been inputted by Harshith M.R., Jignesh Dave, Manisha, Monali, Purva S. Dholakia, Seethalakshmi M.L. and Varinder Singh. Instead of Varinder Singh all are the native speakers of Gujarati.

5.3.4 Validation and Normalization Workshops

Linguistic Data Consortium for Indian Languages (LDC-IL) conducted workshop for data Cleaning/validation and normalization. The experts unanimously suggested that the Gujarati text corpus should remain true to the text.

5.3.5 Proofreading

Gujarati text data has been proofread by both internal resource persons and the resource persons engaged in the programme for Corpus Cleaning/validation. The program was Text Corpus Cleaning Workshop: Gujarati from 23rd August 2010 to 31st August 2010.

It was so decided and followed across the languages that text manipulation be avoided thoroughly and only the typo errors committed during the input process have been corrected with reference to the source materials/hard copies. The source printed materials collected for the corpus are contemporary, mainly published after 1990.

The following resource persons attended in the above-mentioned workshop for Gujarati corpus. They are Gadhavi Hirenkumar, Purva Dholakia, Moti Prajapati, Mahesh Solanki, Sushila, Natwarlal D.Modha, Dr. Nilotpala Gandhi, Dr. Pinky Y. Pandya, Rameschandra V. Chauduri respectively.

5.4 TRANSLITERATION IN LDC-IL GUJARATI TEXT CORPUS

For easy reference and uniformity of metadata, some entries in the metadata file, namely *'Title', 'Headline', 'Author', 'Editor', 'Translator'* are 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 Characters 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

Ayogavaaha		
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

Table 5-1: LDC-IL Gujarati Transliteration Schema

5.5 COPYRIGHT CONSENTS

The Gujarati text corpus have been collected from various sources and the copyright for the same stays with different sources. However, for the purposes of this corpus, consents have been sought from all the stakeholders. Most of the copyrights (around 82%) belong to private parties with only 18% belonging to the government agencies, either state or the central.

5.6 OVERVIEW OF REPRESENTED DOMAINS

LDC-IL Gujarati Text Corpus size is 28,62,413 words and 1,71,69,357 characters drawn from 1364 different titles, including the extracts from newspapers. The representation of the six domains covered has been shown in the table below:

Text Type	Word Count	Keystroke/Character Count

Typed + Cleaned	17,45,808	1,03,10,911
Crawled	11,16,605	68,58,446
Total	**28,62,413**	1,7,169,357

Table 5-2: Representation of the Typed and Crawled Text in Gujarati Raw Text Corpus

The following table gives a summary of the typed+cleaned and crawled text of the Gujarati Raw Text Corpus. The representation of the six domains/categories covered has been shown in the table below:

#	Domain	Word Count	Percentage
1	Aesthetics	7,42,260	25.93%
2	Commerce	43,733	1.53%
3	Mass Media	10,70,099	37.38%
4	Official Document	29,599	1.03%
5	Science & Technology	6,43,737	22.49%
6	Social Sciences	3,32,985	11.63%
	Total	**28,62,413**	**100.00%**

Table 5-3: Representation of the Domains in Gujarati Raw Text Corpus

As each domain has several subdomains and total number of subdomains are 76, the following table shows the representation of the several domains, both within the domain and across all the domains.

5.6.1 Aesthetics

The Aesthetics domain/category of LDC-IL Gujarati text corpus covers 21 subdomains/subcategories bearing a total of 7,42,260 words along with the overall percentage of 25.93%. The representational details are given in the table below.

#	Subdomain	Word Count	Percentage (within Subdomain	Overall Percentage
1	Autobiographies	16,804	2.26%	0.59%
2	Biographies	2,57,668	34.71%	9.00%
3	Cinema	63,427	8.55%	2.22%
4	Culture	4,228	0.57%	0.15%
5	Fine Arts-Dance	5,867	0.79%	0.20%
6	Fine Arts-Music	3,611	0.49%	0.13%
7	Fine Arts-Musical Instruments	3,207	0.43%	0.11%
8	Folklore	5,299	0.71%	0.19%

9	Handicrafts	4,828	0.65%	0.17%
10	Literature-Criticism	10,239	1.38%	0.36%
11	Literature-Diaries	10,291	1.39%	0.36%
12	Literature-Essays	21,966	2.96%	0.77%
13	Literature-Letters	9,130	1.23%	0.32%
14	Literature-Novels	23,041	3.10%	0.80%
15	Literature-Plays	43,002	5.79%	1.50%
16	Literature-Science Fiction	12,105	1.63%	0.42%
17	Literature-Short Stories	1,09,950	14.81%	3.84%
18	Literature-Speeches	6,311	0.85%	0.22%
19	Literature-Text Books (School)	571	0.08%	0.02%
20	Literature-Travelogues	1,22,533	16.51%	4.28%
21	Mythology	8,182	1.10%	0.29%
	Total	**7,42,260**	**100.00%**	**25.93%**

Table 5-4: Representation of Aesthetics Domain

5.6.2 Commerce

The Commerce domain/category of Gujarati text corpus covers 5 subdomains/subcategories bearing a total of 43,733 words along with the overall percentage of 1.53%. The representational details are given in the table below.

#	Subdomain	Word Count	Percentage (within Subdomain)	Overall Percentage
1	Accountancy	5,056	11.56%	0.18%
2	Banking	25,985	59.42%	0.91%
3	Finance	5,277	12.07%	0.18%
4	Industry	4,001	9.15%	0.14%
5	Management	3,414	7.81%	0.12%
	Total	**43,733**	**100.00%**	**1.53%**

Table 5-5: Representation of Commerce Domain

5.6.3 Mass Media

The Mass Media domain/category of LDC-IL Gujarati text corpus covers 4 subdomains/subcategories bearing a total of 10,70,099 words along with the overall percentage of 37.38%. The representational details are given in the table below.

#	Subdomain	Word Count	Percentage (within Subdomain)	Overall Percentage
1	Article	3,32,566	31.08%	11.62%
2	Editorial	1,16,298	10.87%	4.06%
3	General News	3,41,211	31.89%	11.92%
4	Sports News	2,80,024	26.17%	9.78%
	Total	**10,70,099**	**100.00%**	**37.38%**

Table 5-6: Representation of Mass Media Domain

5.6.4 Official Document

The Official Document domain/category of Gujarati text corpus covers 2 subdomains/subcategories bearing a total of 29,599 words along with the overall percentage of 1.03%. The representational details are given in the table below.

#	Subdomain	Word Count	Percentage (within Subdomain)	Overall Percentage
1	Administration	6,685	22.59%	0.23%
2	Parliamentary/Assembly Debates	22,914	77.41%	0.80%
	Total	**29,599**	**100.00%**	**1.03%**

Table 5-7: Representation of Oficial Documents Domain

5.6.5 Sceience and Technology

The Science and Technology domain/category of Gujarati text corpus covers 26 subdomains/subcategories bearing a total of 6,43,737 words along with the overall percentage of 22.49%. The representational details are given in the table below.

#	Subdomain	Word Count	Percentage (within Subdomain)	Overall Percentage
1	Agriculture	85,298	13.25%	2.98%
2	Architecture	2,081	0.32%	0.07%
3	Astrology	2,574	0.40%	0.09%
4	Astronomy	16,161	2.51%	0.56%
5	Ayurveda	36,153	5.62%	1.26%
6	Bio Chemistry	10,625	1.65%	0.37%
7	Biology	20,857	3.24%	0.73%
8	Botany	48,548	7.54%	1.70%
9	Chemistry	27,553	4.28%	0.96%
10	Computer Sciences	6,086	0.95%	0.21%

11	Criminology	2,229	0.35%	0.08%
12	Engineering-Civil	16,721	2.60%	0.58%
13	Engineering-Electrical	7,247	1.13%	0.25%
14	Engineering-Electronics Communication	7,584	1.18%	0.26%
15	Engineering-Mechanical	27,606	4.29%	0.96%
16	Engineering-Others	23,222	3.61%	0.81%
17	Film Technology	25,529	3.97%	0.89%
18	Geology	13,656	2.12%	0.48%
19	Homeopathy	3,863	0.60%	0.13%
20	Mathematics	16,088	2.50%	0.56%
21	Medicine	46,104	7.16%	1.61%
22	Physics	51,699	8.03%	1.81%
23	Psychology	34,288	5.33%	1.20%
24	Text Book (Science)	18,629	2.89%	0.65%
25	Veterinary	59,404	9.23%	2.08%
26	Yoga	13,922	2.16%	0.49%
	Total	**6,43,737**	**100%**	**22.49%**

Table 5-8: Representation of Science and Technology Domain

5.6.6 Social Sciences

The Social Sciences domain/category of Gujarati text corpus covers 17 subdomains/subcategories bearing a total of 332985 words along with the overall percentage of 11.63%. The representational details are given in the table below.

#	Subdomain	Word Count	Percentage	Overall Percentage
1	Economics	44,789	13.45%	1.56%
2	Education	27,662	8.31%	0.97%
3	Fisheries	4,684	1.41%	0.16%
4	Geography	8,847	2.66%	0.31%
5	Health and Family Welfare	40,502	12.16%	1.41%
6	History	36,390	10.93%	1.27%
7	Home Science	24,354	7.31%	0.85%
8	Journalism	6,415	1.93%	0.22%
9	Law	25,844	7.76%	0.90%
10	Linguistics	5,944	1.79%	0.21%

11	Philosophy	1,465	0.44%	0.05%
12	Political Science	27,186	8.16%	0.95%
13	Public Administration	3,725	1.12%	0.13%
14	Religion/Spiritual	14,513	4.36%	0.51%
15	Sociology	51,695	15.52%	1.81%
16	Sports	1,749	0.53%	0.06%
17	Text Book (Social Science)	7,221	2.17%	0.25%
	Total	**3,32,985**	**100.00%**	**11.63%**

Table 5-9: Representation of Social Science Domain

6 HINDI RAW TEXT CORPUS

Satyendra Awasthi, Madhupriya Pathak, Rajesha N, Manasa G, Narayan Choudhary, L. Ramamoorthy

6.1 INTRODUCTION

Hindi is an Indo-Aryan language, a descendent of *Sanskrit*, which is spoken in the central and northern India, in the states of *Bihar, Chhattisgarh, Delhi, Haryana, Himachal Pradesh, Jharkhand, Madhya Pradesh, Rajasthan, Uttarakhand* and *Uttar Pradesh*. It is the official language of the Union of India and is also *lingua franca* across India. Being the most intelligible language of India, it is currently reported to be spoken as the first language by 528.35 million people in India (as per 2011 census of India) i.e. a total of 43.63% of the populace of India speaks Hindi as their primary language.

According to the constitution of India the official languages of the union of India are *Hindi* and *English*, where *Hindi* to be written in *Devanagari*. Based on the provisions mentioned in the Official Language Act, *Hindi* is used for official activities such as communications between the Central Government and State Government which recognizes *Hindi* as official language, judiciary and parliamentary proceedings.

Hindi is written in Devenagari script, a Left to Right script which is a descendent of *Brahmi* script. The script is also used to write several other languages of India and neighboring countries such as Nepali, Marathi, Maithili etc.

Hindi text corpus has been collected from various areas in India, mostly from Uttar Pradesh. The greater part of the corpus has been taken from *Kendriya Hindi Sansthaan* (Central Institute of Hindi), Delhi and Agra libraries and *Bhartiya Bhasha Sansthan* (Central Institute of Indian Languages), Mysore library. LDC-IL has tried to cover the entire category in its standard list. Some categories such as novel, short stories have huge greater proportion of content share than the other domains or sub-domains such as science, technology, economics etc.

6.2 PECULIARITIES OF HINDI TEXT

The Hindi text corpus can be broadly classified into two: literary text and non-literary text. These two explicitly show their differences in terms of the types and the source of the word usage and variety that it brings into corpus. Literary texts are mainly narratives, and more or less they contain the elements of fiction. Some instances of such texts are: novels, short stories and plays. On the other hand the non-literary texts convey information based on their primary purpose. These range from the articles in academic journals, legal documents to the texts of various scientific or technical subjects. The literary texts have emotional elements, cultural information, dialectical variations, ambiguity etc. whereas the non-literary texts show a tendency of usage of technical and scientific terms, foreign words etc.

6.3 DATA SAMPLING NOTES

6.3.1 Principles of Data Sampling

Hindi text data sampling strictly follows the LDC-IL text corpus collection generic guidelines, which are noted in the LDC-IL generic corpus documentation.

6.3.2 Fieldworks Undertaken

Hindi text corpus is collected from various libraries in India, mostly from Uttar Pradesh. The text materials were collected by conducting fieldworks undertaken during the period from 2005 to 2008. The greater part of the text has been taken from Kendriya Hindi Sansthaan, Delhi and Agra library and Central Institute of Indian Languages, Mysore library.

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

1. Allahabad Public Library, Allahabad
2. Banaras Hindu University, Varanasi
3. Central Institute of Indian Languages, Mysore
4. Kendriya Hindi Sansthaan, Agra, and
5. Kendriya Hindi Sansthaan, Delhi

Collected text materials have been published at various places in India. Such as Delhi, Uttar Pradesh, Rajsthan, Madhya Pradesh, Uttarakhand, Bihar, Himachal Pradesh, Hariyana, Jharkhand, Maharashtra, Kerala etc.

An attempt has been made to cover the entire category in its standard list. Some categories like novel, short stories have huge amount of books but some categories like physics, chemistry, economics have very less amount of books. Literary texts are easily available in Hindi but getting scientific text is very difficult. Some categories like epigraphy, finance, oceanology text are very less then Literary texts in Hindi.

Collecting text data from the field is a difficult job. Most of the libraries do not allow to take 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 to take many books at a time, the photocopy kiosk had issues as there was a long queue.

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

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

6.3.3 Data Inputting

All the text has been typed in Unicode using the InScript Keyboard directly onto the XML files. The data has been inputted by around 26 inputters.

6.3.4 Validation and Normalization Workshops

As per the validation and normalization workshop conducted by LDC-IL, it was recommended by the experts that the Hindi text corpus should remain true to the text.

6.3.5 Proofreading

Hindi text data has been proofread by internal resource persons and the resource persons engaged in the Short-term goal oriented projects (Text corpus cleaning Workshops). An account of such workshops is as below:

1. August 2010
2. 31st Dec. 12 – 1st March 13
3. 1st June – 31st July, 2015
4. 23rd May – 15th July, 2016
5. 2nd August – 28th September 2018

Text manipulation has been avoided thoroughly, and only the mistakes occurred during the input process has been corrected with a reference to the hard copies of the texts. Since poetry doesn't serve the purpose of general machine learning due to its anomalous scheme, grammar and construction which doesn't adhere to the day to day language needs, therefore instances of poetry have been removed from the running texts.

The printed material collected for the corpus is contemporary, mainly published after 1990.

6.3.6 Data Extracted from Websites

Hindi News corpus data is extracted from News websites of *"Ranchi Express" (*http://ranchiexpress.com*), "Dainik Bhaskar" (*https://www.bhaskar.com), *"Rajasthan Patrika" (https://www.patrika.com),* and *"Nav Bharat Times" (https://navbharattimes.indiatimes.com).* The news content was categorized based on the content of the text and archived. The period of selection of the news corpus ranges from 2005 to 2008.

6.4 TRANSLITERATIONS IN LDC-IL HINDI TEXT CORPUS

For easy reference and uniformity of metadata, some entries in the metadata file, namely *'Title', 'Headline', 'Author', 'Editor', 'Translator'* are transliterated from Hindi to Roman letters. Numeric characters were transliterated from Hindi to roman system.

For such purpose the LDC-IL transliteration scheme for Devenagari to Roman characters is given below:

LDC-IL Transliteration Schema

Devanagari characters to Roman and Hindi Numerals to Roman

Vowels and Vowel Signs															
Vowel	अ	आ	इ	ई	उ	ऊ	ऋ	ए	ऐ	ओ	औ	अं	अः	अँ	आँ
Matra		ा	ि	ी	ु	ू	ृ	े	ै	ो	ौ	ं	ः	ँ	ॉ
Key	a	A	i	I	u	U	x	E	ai	O	au	M	H	m'	ao

Consonant							
Consonant	क्	ख्	ग्	घ्	ङ्		
Key	k	kh	g	gh	ng'		
Consonant	च्	छ्	ज्	झ्	ञ्		
Key	c	ch	j	jh	nj'		
Consonant	ट्	ठ्	ड्	ढ्	ण्	ड़्	ढ़्
Key	T	Th	D	Dh	N	D'	Dh'
Consonant	त्	थ्	द्	ध्	न्		
Key	t	th	d	dh	n		
Consonant	प्	फ्	ब्	भ्	म्		
Key	p	ph	b	bh	m		
Consonant	य्	र्	ल्	व्			
Key	y	r	l	v			
Consonant	श्	ष्	स्	ह्			
Key	sh	S	s	h			

These are the borrowed sounds, however they are listed in the schema since they occur frequently in the literary text.

Borrowed					
Consonant	क़	ख़	ग़	ज़	फ़
Key	k'a	Kh'a	g'a	j'a	ph'a

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

6.5 Overview of Represented Domains

The size of LDC-IL Hindi Text Corpus is: 10317177 Words and 52569629 characters, gathered from 1223 different titles, including the extracts from newspapers. The data can be categorized into two classes namely ‘Typed and cleaned corpus’ and ‘Crawled corpus’. The crawled data has been crawled mainly from news websites and archived using the standard processing of LDC-IL text corpus preparation.

The following table gives a summary of the typed and crawled text of the Hindi Raw Text Corpus.

Text Type	Word Count	Keystroke/Character Count
Typed+Cleaned	5315913	26812263

Crawled	5001264	25757366
Total	10317177	52569629

Table 6-1: Representation of the typed and crawled Hindi Text Corpus

The representation of the four major domains covered has been shown in the table below:

Domain	Domain Word Count	Percentage
Aesthetics	3822697	37.05%
Mass Media	5012327	48.58%
Science & Technology	549143	5.32%
Social Sciences	933010	9.04%
Total	10317177	100.00%

Table 6-2: Representation of the Domains in Hindi Text Corpus

As each domain has several sub-domains, the following table shows the representation of the several domains, both within the domain and across all the domains.

6.5.1 Aesthetics

The Aesthetics category of Hindi text corpus covers 22 sub-categories bearing a total of 38,22,697 words along with the overall percentage of 37.05%. The representational details are given in the table below.

#	Sub Category	Word Count	Percentage within Subdomain	Overall Percentage
1	Autobiographies	57409	1.50%	0.56%
2	Biographies	221526	5.80%	2.15%
3	Culture	60574	1.58%	0.59%
4	Fine Arts-Dance	6156	0.16%	0.06%
5	Fine Arts-Music	31900	0.83%	0.31%
6	Fine Arts-Sculpture	5737	0.15%	0.06%
7	Folk Tales	5963	0.16%	0.06%
8	Folklore	6102	0.16%	0.06%
9	Humour	24849	0.65%	0.24%
10	Literary Texts	22098	0.58%	0.21%
11	Literature-Children's Literature	33749	0.88%	0.33%
12	Literature-Criticism	149527	3.91%	1.45%
13	Literature-Diaries	16585	0.43%	0.16%
14	Literature-Essays	60496	1.58%	0.59%
15	Literature-Letters	20421	0.53%	0.20%
16	Literature-Novels	1646693	43.08%	15.96%

17	Literature-Plays	163518	4.28%	1.58%
18	Literature-Science Fiction	3544	0.09%	0.03%
19	Literature-Short Stories	1235074	32.31%	11.97%
20	Literature-Speeches	5634	0.15%	0.05%
21	Literature-Travelogues	16151	0.42%	0.16%
22	Mythology	28991	0.76%	0.28%
	Total	**3822697**	**100%**	**37.05%**

Table 6-3: Aesthetics Category Representation

6.5.2 Mass Media

The Mass Media category of Hindi text corpus covers 5 sub-categories bearing a total of 50,12,327 words along with the overall percentage of 48.58%. The representational details are given in the table below.

#	Sub Category	Word Count	Percentage within Subdomain	Overall Percentage
1	Interviews	1823101	36.37%	17.67%
2	Political	32499	0.65%	0.31%
3	General News	11063	0.22%	0.11%
4	Editorial	2558326	51.04%	24.80%
5	Sports News	587338	11.72%	5.69%
	Total	**5012327**	**100%**	**48.58%**

Table 6-4: Mass Media Category Representation

6.5.3 Science and Technology

The Science and Technology category of Hindi text corpus covers 24 sub-categories bearing a total of 5,49,143 words along with the overall percentage of 5.32%. The representational details are given in the table below.

#	Sub Category	Word Count	Percentage within Subdomain	Overall Percentage
1	Oceanology	20175	3.67%	0.20%
2	Geology	2629	0.48%	0.03%
3	Botany	26371	4.80%	0.26%
4	Film Technology	36763	6.69%	0.36%
5	Astronomy	36276	6.61%	0.35%
6	Medicine	5962	1.09%	0.06%
7	Yoga	27692	5.04%	0.27%
8	Agriculture	22026	4.01%	0.21%
9	Mathematics	92869	16.91%	0.90%
10	Biology	5489	1.00%	0.05%
11	Ayurveda	36544	6.65%	0.35%
12	Homeopathy	21870	3.98%	0.21%
13	Psychology	15858	2.89%	0.15%

14	Computer Sciences	13325	2.43%	0.13%
15	Physics	5239	0.95%	0.05%
16	Environmental Science	472	0.09%	0.00%
17	Criminology	18812	3.43%	0.18%
18	Chemistry	42934	7.82%	0.42%
19	Zoology	6991	1.27%	0.07%
20	Astrology	55389	10.09%	0.54%
21	Textile Technology	20596	3.75%	0.20%
22	Architecture	5735	1.04%	0.06%
23	Forestry	9534	1.74%	0.09%
24	Horticulture	19592	3.57%	0.19%
	**Total**	**549143**	**100%**	**5.32%**

Table 6-5: Science and Technology Category Representation

6.5.4 Social Sciences

The Social Sciences category of Hindi text corpus covers 15 sub-categories bearing a total of 9,33,010 words along with the overall percentage of 9.04%. The representational details are given in the table below.

#	Sub Category	Word Count	Percentage within Subdomain	Overall Percentage
1	Archeology	9610	1.03%	0.09%
2	Economics	15756	1.69%	0.15%
3	Education	18831	2.02%	0.18%
4	Health and Family Welfare	67383	7.22%	0.65%
5	History	203476	21.81%	1.97%
6	Home Science	20537	2.20%	0.20%
7	Journalism	94379	10.12%	0.91%
8	Law	5192	0.56%	0.05%
9	Library Science	42094	4.51%	0.41%
10	Personality Development	5206	0.56%	0.05%
11	Philosophy	45591	4.89%	0.44%
12	Political Science	97128	10.41%	0.94%
13	Religion/Spiritual	260543	27.92%	2.53%
14	Sociology	29889	3.20%	0.29%
15	Sports	17395	1.86%	0.17%
	**Total**	**933010**	**100%**	**9.04%**

Table 6-6: Social Sciences Category Representation

6.6 COPYRIGHT CONSENTS

The Hindi text corpus has been collected from various sources therefore the copyright for the same stays with the different sources. However, for this purposes consents have been sought from all the stakeholders. Most of the copyrights belong to private parties with only a minor part belonging to the government agencies, either state or the central.

7 KANNADA RAW TEXT CORPUS

Vijayalaxmi F. Patil, Chetan Baji, Rajesha N, Manasa G, Narayan Choudhary, L. Ramamoorthy

7.1 INTRODUCTION

One of the most ancient languages of India and a prominent language among the Dravidian languages, the Kannada language is used in its various forms by about 45 million people as a spoken language. Kannada is widely spoken in the areas of Karnataka, Kerala, Maharashtra, Andhra Pradesh, Telangana, Goa, and Tamilnadu and also in some of the border areas of other adjacent states. Kannada is the administrative language of the state of Karnataka and recognised as one among the Classical Languages of India. The Kannada language is written in Kannada script which is evolved from Kadamba script of Bramhi script family. The written forms of Kannada have a history of around 1500 years. The classical Kannada literature received the highest royal patronage during the reigns of the Western Ganga Empire in the 6th Century A.D and the Rashtrakuta Empire during the 9th Century A.D. Also, Kannada has a literary tradition dating back to a 1,000 years.

The intensity and eagerness to develop and preserve the true essence of Kannada Language has increased significantly during the last century. The use of Kannada on the web has seen an upward graph due to the participation of more number of users consistently over the years. It has grown into a frontal language in the commercial sphere. Software development on the lines of Kannada language has become increasingly popular and is a trend in the market.
The script for writing has evolved from the Brahmi script and is more than 1500-1600 years old. Kannada had already developed sufficiently even at the time of the Halmidi epigraph in the 5th century A.D. According to Dravidian linguist Stanford Schwarz, the linguistic history of Kannada can be divided into three stages. (Kittel, F (1993) [1993]);
1. Halegannada - Old Age AD 450 to AD until 1200,
2. Nadugannada- 1200 AD Until 1700 and up
3. Hosagannada From 1700 to the present time.

There is a considerable difference between spoken and written forms. Spoken Kannada has a social and regional dialectal variation across the geography of Karnataka. Regional dialects are mainly four, Dharwad (Mumbai Karnataka), Mangalore (canara), Mysore and Gulbarga Kannada (Hyderabad Karnataka). The written form is more or less consistent across Karnataka. "About 20 social dialects" of Kannada were reported in the Ethnologue. The prominent among them are Kundagannada (spoken in Kundapur separately), Nadavar-Kannada (spoken by Nadavar), Haviviganna (mainly by the Havyaka Brahmins (Spoken by Madikeri and Sullia region of Dakshina Kannada), Malenadu Kannada (Sakleshpur, Kodagu, Shimoga, and Chikmagalur), Sholaga, Gulbarga Kannada, Dharwad Kannada etc. Vokkaligas speak in Kannada language in their native languages of Uttara Kannada, Shimoga and Dakshina Kannada districts.

7.2 KANNADA SCRIPT EVOLUTION AND TEXT

The Old Kannada Script (*Halagannada lipi*) evolved from Kadamba Script underwent modification and developed as Kannada and Telugu Scripts by 1500 C.E. The printing press technology brought by the Europeans standardized both the scripts distinctively. There were attempts to further modify the Kannada script in the past, for the administrative reasons like to have a standardized keyboard for Kannada typewriter. In the early 19th Century when Tamil script underwent changes, there were attempts to modify the Kannada script in similar lines. But unlike Tamil, Kannada retained much of its letters. One of the propositions was to structure the consonant clusters by writing halfletters (also known as *ardhaksharas*, since the consonant is always written with an inherent vowel) in sequence like in most north Indian scripts and Tamil, instead of stacking the consonant diactric markes known as *ottakshara*. It was not accepted by the majority of mass. Because of these various proposals, long drawn discussions and rejection by the mass for any modification, in 1956 Kannada keyboard layout was finally declared as standard which was designed in 1932 itself by K. Anantha Subbarao. In a way this typewriter shaped the Kannada script to a great extent.

The Kannada corpus at LDC-IL is taken both from literary and non-literary texts. It contains Novels, short stories, plays, humour, folklore etc. Non-literary texts are the text about various scientific or technical oriented for example horticulture, film industry, medicine etc. We can see foreign language words mainly English words written in both Kannada as well as in foreign script. Being a Language of Dravidian Language family, Kannada is one among the languages which is known for its agglutinativeness. Like in English and many other languages, Kannada also uses period as a sentence boundary maker and the question mark, and other punctuation markers seen in Kannada text are also similar.

There are continuous proposals to make further modifications in the Kannada script, to drop vowels like ‘ಐ’ to add obsolete letters back, to drop voiced consonant, to modify the way certain consonants are written in order to bring a uniformity etc. But none of the attempts were able to change the script being used. In 1990 only ‘ಱ’ was dropped from the Varnamala (alphabets) taught to children as it has no usage in Kannada.

The Kannada corpus at LDC-IL is taken both from literary and non-literary texts. It contains Novels, Short Stories, Plays, Humour, Folklore etc. Non-literary texts are the texts about various scientific or technical oriented for example Horticulture, Film Industry, Medicine etc. We can see foreign language words, mainly English words, written in both Kannada as well as in foreign script. Being a Language of Dravidian Language family Kannada is one among the languages which is known for its agglutinativenes. Kannada uses period as a sentence boundary maker just like in English and many other languages.

7.3 DATA SAMPLING NOTES

7.3.1 Principles of Data Sampling

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

7.3.2 Source Text Collection

Some part of the Kannada Text Corpus had taken from KHS (Kendriya Hindi Samiti) Corpora Project and some text material was collected by conducting field works.
Overall, As per the present information the following libraries served as the source of the Kannada text corpus:

Source	Year	Field Investigator
• SRLC-CIIL, Mysore • Main Library-CIIL, Mysore • GIA-CIIL, Mysore	2006	Deepti R.
• Karnataka University Library - Dharwad • Karnataka Arts College Library - Dharwad • Dr. B.D. Jatti Homeopathic College and Hospital- Dharwad	2012	Dr. Vijayalaxmi F. Patil
• Personal Library	2016	Rajesha N. & Chetan Baji

Table 7-1: Source of Kannada Text Data Collection

Collected text materials have been published at various places within Karnataka and other states of India such as Karnataka, Tamilnadu, Maharashtra, Delhi etc.
An attempt has been made to cover the entire domain in its standard list. Some domains like Novel, Short Stories have huge amount of books but some domains like Physics, Chemistry, Economics have very less amount of books. Literary Kannada texts are easily available but getting texts from science and commerce is really difficult. But the effort is made to collect from all domains.

Collecting text data from the field is a difficult job. It is difficult to get it photocopied within given span of time. 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 a maximum three or four books in library.

Sometime Xerox attendants refused to photocopy randomly selected pages because of the long queue waiting and it takes up more time for them to turn the pages compared to continuous page photocopying they are accustomed to. It was another issue that the Field Investigator had to carry a huge list of photocopy bundles with them which was many times cumbersome to travel with.

Despite of all the issues as mentioned in above lines, the Filed Invetigator/Linguists working on the data collection had to deal with and get going.

7.3.3 Data inputting

All the text has been typed in Unicode onto the XML files. The data has been inputted by L.Shashikala, M. Mamatha, K.N.Amruta, R.Rajeshwari, R.Sevanthi, J.Shobha, K.R.Veena, Kavitha Lenin, C.J.Anand, B.H.Kumaraswamy and P.Anitha, all being native speakers of Kannada.

7.3.4 Normalization Workshops[2]

Workshop on text normalization was conducted at Linguistic Data Consortium from June-July 2010, three other short term workshops were also conducted. Short term goal oriented project-Kannada for text corpus were conducted on these dates - 29th October -21st Dec 2012, 5th January 2015 – 6th March 2015, 1st June – 31st July 2015.

7.3.5 Proofreading

Kannada text data has been proofread by internal resource persons and external people. The text has always been kept true to the printed material and typos, if any, occurring at the time of typing have only been corrected. The printed materials collected for the corpus is contemporary, mainly published after 1990.

7.3.6 Data Extracted from Websites

Kannada News corpus data is extracted from News websites of *" http://epaper.udayavani.com/" and http://www.kannadaprabha.com/* . *The* news content was categorized based on the content of the text and archived. The period of selection of the news corpus ranges from 2005 to 2008.

7.4 TRANSLITERATION IN LDC-IL KANNADA TEXT CORPUS

For easy reference and uniformity of metadata, some entries in the metadata file, namely *'Title', 'Headline', 'Author', 'Editor', 'Translator'* are transliterated from Kannada to Roman letters. Numeric characters were transliterated from Kannada to Hindu-Arabic system.

2

The LDC-IL transliteration scheme of Kannada to Roman is given below. The greyed out characters are obsolete. They may rarely present in the current LDC-IL corpus.

LDC-IL Transliteration Schema														
Kannada characters to Roman and Kannada Numerals to Hindu-Arabic														
Vowels														
ಅ	ಆ	ಇ	ಈ	ಉ	ಊ	ಋ	ೠ	ಌ	ೡ	ಎ	ಏ	ಐ	ಒ	ಓ	ಔ
	◌ಾ	◌ಿ	◌ೀ	◌ು	◌ೂ	◌ೃ	◌ೄ	◌ೢ	◌ೣ	◌ೆ	◌ೇ	◌ೈ	◌ೊ	◌ೋ	◌ೌ
a	A	i	I	u	U	x	X	q	Q	e	E	ai	o	O	au
Consonants						**Symbols**									
ಕ	ಖ	ಗ	ಘ	ಙ		◌ಂ	◌ಁ	◌ಂ	◌ಃ	◌ಽ	ಂಂ				
ka	kha	ga	gha	ng'a		M'	m'	M	H	J	G				
ಚ	ಛ	ಜ	ಝ	ಞ											
ca	cha	ja	jha	nj'a											
ಟ	ಠ	ಡ	ಢ	ಣ											
Ta	Tha	Da	Dha	Na											
ತ	ಥ	ದ	ಧ	ನ											
ta	tha	da	dha	na											
ಪ	ಫ	ಬ	ಭ	ಮ											
pa	pha	ba	bha	ma											
ಯ	ರ	ಲ	ವ	ಶ	ಷ	ಸ	ಹ	ಳ	ೞ	ಱ					
ya	ra	la	va	sha	Sa	sa	ha	La	Za	Ra					
Numerals (Kannada to Hindu-Arabic)														
೦	೧	೨	೩	೪	೫	೬	೭	೮	೯						
0	1	2	3	4	5	6	7	8	9						

Table 7-2: Transliteration Scheme of Kannada to Roman

Note: The letters in gray cells are obsolete in usage or only used for Sanskrit language written in Kannada Script. These letters may rarely present in the text corpus.

7.5 OBSERVATIONS WORTH MENTIONING IN KANNADA CORPUS

7.5.1 Contoids in the Middle of the word

It is observed that Kannada is a vowel ending language. Consonants will occur only in the initial and medial positions. The consonant ending words found in Kannada are either named entities or borrowed/native words from English, Arabic, Persian and Indo-Aryan language influence. When a word with pure consonant is borrowed and needed to be affixed with a Kannada affixation, zero with non-joiner is used after the *suruli* diacritic symbol that marks the ardhakshara (‘್’) to avoid forming consonant clusters.
For eg.

Word (Gloss: in December)	**Split form**
ಡಿಸೆಂಬರ್‌ನಲ್ಲಿ (accepted form)	ಡ + ಿ + ಸ + ೆ + ಂ + ಬ + ರ + ್ + **ZWNJ** + ನ + ಲ + ್ + ಲ + ಿ
ಡಿಸೆಂಬರ್ನಲ್ಲಿ (non accepted form)	ಡ + ಿ + ಸ + ೆ + ಂ + ಬ + ರ + ್ + ನ + ಲ + ್ + ಲ + ಿ

Zero Width Non-Joiner (ZWNJ) character has no value of its own. It is supposed to be only font directives, directing a font to select from two or more semantically same renderings. When it comes to Kannada, ZWNJ becomes an alien language construct introduced to Kannada by Unicode to produce a Non-Joiner letters. Thus, it is possible to produce two semantically different words, which differ only by ZWNJ in their Unicode representation. Fortunately Kannada being vowel ending language, no such incidence is yet reported. LDC-IL Kannada Text Corpus uses ZWNJ wherever it needed. Normally search algorithm ignores ZWNJ because it should not care about the rendering of the word. So nothing was reported yet about ZWNJ usage in Kannada text has an impact the search results.

7.5.2 Repha Modifying into arkavattu

If there is a consonant cluster in Kannada, the first consonant will be fully rendered, and following consonants will be written with conjunct symbol. But in case of ‘*repha*’ being the first symbol, it can allow the next immediate consonant to write fully and modify itself into ‘*arkavattu’,* and follow a sequence of consonant cluster.

For Eg. ಸೂರ್ಯ (*Repha,* written fully and taking the conjunct symbol of following consonant ‘*ya*’)
ಸೂರ್ಯ (*Repha,* allowing the following consonant to write fully and modifying itself as ‘arkavattu’)

Both are acceptable forms in Kannada Script grammar, but the second one is most preferred in print media because it avoids stacking of *ottakshara* (conjunct consonant diacritic symbol). Since the second form is outnumbered in the available printed corpus, the logic is built to make the initial repha to modify itself as ‘arkavattu’ when it is followed by another consonant. But when the repha comes in the initial position it should not form ‘*arkavattu*’ which is unacceptable by the script grammar. To keep the repha in full form which is a conjunct letter, in the beginning of

a word, the knack was to add a ZWJ (zero width joiner) after the *suruli* diacritic (‘್’) of *ardhakshara*.

- ರ + ್ + *ZWJ* + ಯ = ರ್‍ಯ
- ರ + ್ + ಯ = ರ್ಯ

Word (Gloss: Rank)	Split form
ರಾಷ್ಟ್ರಕ್ (accepted form)	ರ + ್ + ZWJ + ಯ + ಾ + ಂ + ಕ + ್
ರ್ಯಾಂಕ್ (non-accepted form)	ರ + ್ + ಯ + ಾ + ಂ + ಕ + ್

Zero Width Joiner (ZWJ) character has no value of its own. It is supposed to be only font directives, directing a font to select from two or more semantically same renderings. When it comes to Kannada, ZWJ becomes an alien language construct introduced to Kannada by Unicode to produce a different form of ligature. Thus, it is possible to produce two semantically different words, which differ only by ZWNJ in their Unicode representation. Fortunately this is just a case *repha* in the initial position and happens only in borrowed foreign words, no such incidence is reported where it conflicts with semantically different word of Kannada. LDC-IL Kannada Text Corpus uses ZWJ in few places and does not uses in few places to keep the representation for both. Normally search algorithm ignores ZWJ because it should not care about the rendering of the word. So nothing was reported yet about ZWJ usage in Kannada text has an impact upon the search results, *arkavatthu* is being opposed by many scholars, if it is removed in the future, and fonts can get tuned properly rendering will be proper without ZWJ as it is not creating any issue.

7.5.3 Vowel with Consonant Conjunct

The Kannada script grammar does not permit the use of vowel to have a consonant conjunct symbol because the conjunct is always a combination of two or more pure consonants. Vowel being a full letter, cannot be considered as half character to attach to a consonant diacritic as conjunct. In the LDC-IL Kannada corpus we find ‘ಆ್ಯ’ which is a combination of “ಆ + ್ + ಯ”, which is known as ‘garbage writing’. This is being practiced in recent time to make a distinction in the sound of ಆ, as in ಆನೆ, and ಆ್ಯ as in ಆ್ಯಕ್ಟ್. To keep it true to text and LDC-IL Corpus has retained.

7.6 COPYRIGHT CONSENTS

The Kannada text corpus has been collected from various sources and the copyright for the same stays with different sources. However, for the purposes of this corpus, consent has been sought from all the stakeholders. Most of the copyrights (around 82%) belong to private parties with and 18% belonging to the government agencies, either state or the central. Copyright holders were contacted through telephonic conversation and the respective consents have been received via various sources such as email, letters and direct contact through field work.

7.7 OVERVIEW OF REPRESENTED DOMAINS

LDC-IL Kannada Text Corpus size is: 77,63,124 Words drawn from 1,772 different titles, including web news. The total Corpus character size is 6,49,09,781. The data can be categorized into two classes of typed+cleaned and crawled. The crawled data has been crawled mainly from news websites and archived using the standard processing of LDC-IL text corpus preparation.

The following table gives a summary of the typed and crawled text of the Kannada Raw Text Corpus.

Text Type	Word Count	Keystroke/Character Count
Typed+Cleaned	50,95,039	4,30,86,957
Crawled	26,68,085	2,18,22,824
Total	**77,63,124**	**6,49,09,781**

Table 7-3: Representation of the Typed and Crawled Text Corpus

The representation of the six major domains covered has been shown in the table below:

Domain	Domain Word Count	Percentage
Aesthetics	37,78,723	48.68%
Commerce	2,07,053	2.67%
Mass Media	26,81,611	34.54%
Official Document	5,357	0.07%
Science and Technology	2,43,166	3.13%
Social Sciences	8,47,214	10.91%
Total	**77,63,124**	**100.00%**

Table 7-4: Representation of the various domains in Kannada text corpus

As each domain has several sub-domains, the following table shows the representation of the several domains, both within the domain and across all the domains.

7.7.1 Aesthetics

The Aesthetic domain of Kannada text corpus covers 28 subdomains bearing a total of 37,78,723 words along with the overall percentage of 48.68%. The representational details are given in the table below.

#	Subdomain	Word Count	Percentage (within Subdomain)	Overall Percentage
1	Autobiographies	1,96,416	5.20%	2.53%
2	Biographies	3,43,728	9.10%	4.43%
3	Cinema	23,886	0.63%	0.31%
4	Culture	20,981	0.56%	0.27%
5	Fine Arts-Dance	25,956	0.69%	0.33%
6	Fine Arts-Drawing	16,855	0.45%	0.22%
7	Fine Arts-Hobbies	503	0.01%	0.01%
8	Fine Arts-Music	44,766	1.18%	0.58%
9	Fine Arts-Sculpture	21,075	0.56%	0.27%
10	Folk Tales	23,957	0.63%	0.31%
11	Folklore	1,05,377	2.79%	1.36%
12	Humour	956	0.03%	0.01%
13	Literary Texts	7,73,876	20.48%	9.97%
14	Literature-Children's Literature	5,781	0.15%	0.07%
15	Literature-Criticism	8,57,540	22.69%	11.05%
16	Literature-Diaries	4,466	0.12%	0.06%
17	Literature-Epics	16,499	0.44%	0.21%
18	Literature-Letters	4,578	0.12%	0.06%
19	Literature-Novels	4,49,300	11.89%	5.79%
20	Literature-Plays	2,06,951	5.48%	2.67%
21	Literature-Poetry	30,320	0.80%	0.39%
22	Literature-Science Fiction	22,502	0.60%	0.29%
23	Literature-Short Stories	3,61,969	9.58%	4.66%
24	Literature-Speeches	15,025	0.40%	0.19%
25	Literature-Text Books (School)	39,626	1.05%	0.51%
26	Literature-Travelogues	58,012	1.54%	0.75%
27	Mythology	1,04,976	2.78%	1.35%
28	Photography	2,846	0.08%	0.04%
	Total	**37,78,723**	**100%**	**48.68%**

Table 7-5: Representation of Aesthetics

7.7.2 Commerce

The Commerce domainof Kannada text corpus covers 7 subdomains bearing a total of 2,07,053 words along with the overall percentage of 2.67%. The representational details are given in the table below.

#	Sub domain	Word Count	Percentage (within Subdomain)	Overall Percentage
1	Accountancy	64,677	31.24%	0.83%
2	Banking	42,763	20.65%	0.55%
3	Business	16,333	7.89%	0.21%
4	Finance	59,136	28.56%	0.76%
5	Industry	11,898	5.75%	0.15%
6	Management	8,187	3.95%	0.11%
7	Share Market	4,059	1.96%	0.05%
	Total	**2,07,053**	**100%**	**2.67%**

Table 7-6: Representation of Commerce

7.7.3 Mass Media

The Mass Media domainof Kannada text corpus covers 7 subdomains bearing a total of 26,81,611 words along with the overall percentage of 34.54%. The representational details are given in the table below.

#	Sub domain	Word Count	Percentage (within Subdomain)	Overall Percentage
1	Editorial	9,95,339	37.12%	12.82%
2	General News	11,251	0.42%	0.14%
3	Interviews	2,693	0.10%	0.03%
4	Political	7,13,953	26.62%	9.20%
5	Social	367	0.01%	0.00%
6	Speeches	7,474	0.28%	0.10%
7	Sports News	9,50,534	35.45%	12.24%
	Total	**26,81,611**	**100%**	**34.54%**

Table 7-7: Representation of Mass Media

7.7.4 Official Document

The Official Document domainof Kannada text corpus covers a single sub-domainbearing a total of 5,357 words along with the overall percentage of 0.07%. The representational details are given in the table below.

Sub domain	Word Count	Percentage (within Subdomain)	Overall Percentage
Police Document	5,357	100.00%	0.07%

Table 7-8: Representation of Official Document

7.7.5 Science and Technology

The Science and Technology domainof Kannada text corpus cover 31 sub-domainbearing a total of 2,43,166 words along with the overall percentage of 3.13%. The representational details are given in the table below.

#	Sub domain	Word Count	Percentage (within Subdomain)	Overall Percentage
1	Agriculture	56,979	23.43%	0.73%
2	Architecture	6,768	2.78%	0.09%
3	Astrology	4,266	1.75%	0.05%
4	Astronomy	19,719	8.11%	0.25%
5	Ayurveda	12,337	5.07%	0.16%
6	Bio Chemistry	196	0.08%	0.00%
7	Biology	7,427	3.05%	0.10%
8	Biotechnology	205	0.08%	0.00%
9	Botany	4,020	1.65%	0.05%
10	Chemistry	591	0.24%	0.01%
11	Computer Sciences	8,987	3.70%	0.12%
12	Criminology	13,228	5.44%	0.17%
13	Educational Psychology	1,998	0.82%	0.03%
14	Engineering-Others	680	0.28%	0.01%
15	Environmental Science	5,142	2.11%	0.07%
16	Film Technology	1,939	0.80%	0.02%
17	Forestry	4,469	1.84%	0.06%
18	Geology	1,796	0.74%	0.02%
19	Homeopathy	9,314	3.83%	0.12%
20	Horticulture	1,415	0.58%	0.02%
21	Language Technology	4,395	1.81%	0.06%
22	Logic	698	0.29%	0.01%
23	Medicine	14,807	6.09%	0.19%
24	Naturopathy	1,721	0.71%	0.02%
25	Physics	9,132	3.76%	0.12%
26	Psychology	14,956	6.15%	0.19%
27	Sexology	1,280	0.53%	0.02%
28	Text Book (Science)	9,478	3.90%	0.12%
29	Veterinary	9,621	3.96%	0.12%
30	Yoga	15,390	6.33%	0.20%
31	Zoology	212	0.09%	0.00%
	Total	**2,43,166**	**100%**	**3.13%**

Table 7-9: Representation of Science and Technology

7.7.6 Social Science

The Social Science domainof Kannada text corpus cover 17 sub-domainbearing a total of 8,47,214 words along with the overall percentage of 10.91%. The representational details are given in the table below.

#	Sub domain	Word Count	Percentage (within Subdomain)	Overall Percentage
1	Geography	7,243	0.85%	0.09%
2	Health and Family Welfare	19,948	2.35%	0.26%
3	History	20,333	2.40%	0.26%
4	Home Science	26,227	3.10%	0.34%
5	Journalism	6,361	0.75%	0.08%
6	Law	2,640	0.31%	0.03%
7	Library Science	8,047	0.95%	0.10%
8	Linguistics	31,343	3.70%	0.40%
9	Personality Development	1,10,659	13.06%	1.43%
10	Philosophy	6,472	0.76%	0.08%
11	Physical Education	30,361	3.58%	0.39%
12	Political Science	56,538	6.67%	0.73%
13	Public Administration	25,828	3.05%	0.33%
14	Religion/Spiritual	38,005	4.49%	0.49%
15	Sociology	4,566	0.54%	0.06%
16	Sports	39,678	4.68%	0.51%
17	Text Book (Social Science)	2,064	0.24%	0.03%
	Total	**8,47,214**	**100%**	**10.91%**

Table 7-10: Representation of Social Science

Reference:

1. Kittel, F (1993) [1993]. *A Grammar of the Kannada Language Comprising the Three Dialects of the Language (Ancient, Medieval and Modern)*. New Delhi, Madras: Asian Educational Services
2. Kamath, Suryanath U. (2002) [2001]. *A Concise History of Karnataka from Pre-historic times to the Present*. Bangalore: Jupiter books
3. Buchanan, Francis Hamilton (1807). *A Journey from Madras through the Countries of Mysore, Canara, and Malabar. Volume 3*. London: Cadell.
4. https://kn.wikisource.org

8 KASHMIRI RAW TEXT CORPUS

Bi Bi Mariyam, Shahid Bhatt, Rajesha N, Manasa G, Narayan Choudhary, L. Ramamoorthy

8.1 INTRODUCTION

Kashmiri language belongs to Dardic sub-group of Indo-Aryan languages. The Kashmiri language is called "Koshur". The Kashmiri language is written in Sharda, Perso-Arabic and Devanagari scripts. It is primarily spoken in Kashmir Valley and Chenab valley of Jammu and Kashmir state of India. The language spoken in and around Srinagar is regarded as the standard variety. It is used in literature, mass media, and education. It is one of the 22 scheduled languages of India.

Kashmiri has two types of dialects: Regional dialects and Social dialects. Regional dialects are those dialects or variations which are spoken in the regions inside the valley and those spoken outside the valley of Kashmir. Kashmiri speaking area in the valley is ethno-semantically divided into three regions: Maraz (southern and south-eastern region), Kamraz (northern and north-western region) and Srinagar and its neighboring areas. Kashmiri spoken in the three regions is not only mutually intelligible but quite homogeneous. These dialectical variations can be termed as different styles of the same speech. Regional dialects, namely, Poguli and Kishtawari, are spoken outside Kashmir valley. Poguli is spoken in the Pogul and Paristan valleys bordered on the east by Rambani and Siraji, and on the west by mixed dialects of Lahanda and Pahari. Social dialects depend on the extent to which they were affected by either Sanskrit and Perso-Arabic influence.

Kashmiri text corpus is collected from various libraries in Kashmir mostly from Allama Iqbal Library, University of Kashmir.

LDC-IL tried to cover the entire category in its standard list. Some categories like a Novel, Short stories Criticism, and Literature textbooks have a huge amount of books, but some categories like Epic, Letters, Administration, Botany, Physics, Chemistry, Zoology, Legislature, etc have very less amount of books. Literary texts are easily available in Kashmiri but getting a text like Epigraphy, Finance, Share Market is very difficult. Some categories.

8.2 PECULIARITIES OF KASHMIRI TEXT

Linguistically, the Kashmiri language holds a peculiar position because it has some formal features, which show its Dardic characteristics and many other features which it shares with the Indo-Aryan languages. The Corpus of Kashmiri 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 the corpus. Literary texts are texts that are narrative and it contains elements of Fiction. Novels, Short stories, plays are examples of literary text. Non-literary texts are texts whose primary purpose is to convey information. Examples of non-literary texts are a text about various scientific or technical subjects, legal documents, articles in academic journals. In a 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.

8.2.1 The writing System of Kashmiri

The writing system of Kashmiri is based on various scripts. There are three orthographical systems used to write the Kashmiri language. The Sharada script, Devanagari script and Perso-Arabic script. The Kashmiri language is traditionally written in the Sharada script after the 8th Century A.D. Devanagari, also called Nagari, is an abugida used in India and Nepal. It is written from *left to right.* The Perso-Arabic script with additional diacritical marks now known as Kashmiri script has been recognized as the official script for Kashmiri. The Perso-Arabic script that is from *right to left* as in Urdu. Kashmiri is written in both Perso-Arabic and Devanagari scripts.

8.2.2 Fonts

The people usually use a special software "Inpage" for writing languages like Urdu, Persian, Arabic, Kashmiri and Pushto etc. It is a word processor and page layout software. Narqalam font (Naskh) was made to enable typing Kashmiri text. This font is Unicode based and the characters specific to Kashmiri has been added. The scheme LDC-IL used Narqalam font (Naskh) and Afan Koshur, which is nowadays commonly used for modern facilities like Micro soft office and Open office. Afan koshur was built in 2008, The first ever Linux and Windows are compatible with Kashmiri modified Perso-Arabic font.

8.3 DATA SAMPLING NOTES

8.3.1 Principles of Data Sampling

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

8.3.2 Fieldworks Undertaken

Fieldwork for speech data and text corpora collection on June and July 2010 was conducted by Linguistic Data Consortium for Indian Languages (LDC-IL) in Srinagar and fieldwork for text data collection was also conducted during 16th Aug to 6th Sept 2012 at Srinagar. Kashmiri text corpus is collected from various libraries in Srinagar.

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

1. Allama Iqbal Library University of Kashmir
2. Personal Library of Masroor Ahmad Mir
3. Govt Girls Middle School Dever Tral Pulwama Kashmir
4. GIA, CIIL Mysore

The collected text materials have been published at various places within Kashmir. An attempt has been made to cover the entire category in its standard list. The categories like Literature Criticism, Short Stories, Literature Textbooks, and Economics have a huge amount of books but some categories like Fine Arts-Dance, Letters, Botany, Health and Family Welfare, Legislature, Physics, Chemistry have very

less amount of books. Literary texts are easily available in Kashmiri but getting a scientific text is very difficult.

Collecting text data from the field is a difficult job. Most of the libraries do not allow taking a huge amount of text from their shelves at a time because it is against their rules and principles. For a particular period, they issue a maximum number of three or four books. Even if the librarian allowed to take many books at a time. There was an issue in getting photocopies of the text for selected pages. It was another issue that the field worker/linguist had to carry a huge list of photocopy bundles with them which was many a time cumbersome to travel.

The linguists working on the data collection had to deal with all the issues as mentioned above.

8.3.3 Data Inputting

All the texts have been typed in Unicode using the InScript Keyboard directly on to the XML files. The data has been inputted by Bi Bi Mariyam, a native speaker of Kannada.

8.3.4 Validation and Normalization Workshops

A Workshop and NLP Orientation-cum-Training Programme on Creation of Large Scale Annotated Data were conducted by Linguistic Data Consortium.

- NLP Orientation-cum-Training Programme on November 4-15, 2011 at University of Kashmir, Srinagar.
- Workshop on Creation of Large Scale Annotated Data on 20th Dec. 12 – 4th Jan. 13, University of Kashmir, Srinagar.
- 15-day workshop on Speech Corpus Annotation and Text Corpus Sanitation on 21st Oct. to 6th Nov. 2013. University of Kashmir, Srinagar

The experts suggested that the Kashmiri text corpus should remain true to the text.

A Workshop and NLP Orientation-cum-Training Programme on Creation of Large Scale Annotated Data conducted by Linguistic Data Consortium.

8.3.5 Proofreading

Kashmiri text data has been proofread by internal resource persons. The text has always been kept true to the printed material and typos, if any, occurring at the time of typing have only been corrected.

8.4 Overview of Represented Domains

LDC-IL Kashmiri Text Corpus size is 466,054 Words and character count is 26,46948 drawn from titles, including the extracts from Newspaper, Magazines etc. The representation of the 2 major domains covered has been shown in the table below:

Text Type	Word Count	Keystroke/Character Count
Typed+Cleaned	466054	2646948

The representation of the two major domains covered has been shown in the table below:

Domain	Word Count	Percentage
Aesthetics	400474	85.93%
Social Sciences	65580	14.07%
Total	466,054	100

Table 8-1: Representation of the Domains in Kashmiri Text Corpus

As each domain has several sub-domains, the following table shows the representation of the several domains, both within the domain and across all the domains.

Category	Sub Category	Word Count	Percentage (within Sub domain).	Overall Percentage
Aesthetics	Culture	5191	1.30%	1.11%
Aesthetics	Literature-Criticism	296315	73.99%	63.58%
Aesthetics	Literature-Essays	18252	4.56%	3.92%
Aesthetics	Literature-Novels	5897	1.47%	1.27%
Aesthetics	Literature-Plays	7186	1.79%	1.54%
Aesthetics	Literature-Short Stories	12359	3.09%	2.65%
Aesthetics	Literature-Text Books (School)	16333	4.08%	3.50%
Aesthetics	Literature-Travelogues	38941	9.72%	8.36%
Social Sciences	History	8644	13.18%	1.85%
Social Sciences	Linguistics	12735	19.42%	2.73%
Social Sciences	Personality Development	11255	17.16%	2.41%
Social Sciences	Religion/Spiritual	23806	36.30%	5.11%
Social Sciences	Sociology	9140	13.94%	1.96%

Table 8-2: Representation of Sub domains in Kashmiri Text Corpus

8.5 COPYRIGHT CONSENTS

The Kashmiri text corpus has been collected from various sources and the copyright for the same stays with different sources. However, for the purposes of this corpus, consent has been sought from all the stakeholders. Most of the copyrights belong to private parties with only and some belonging to the government agencies, either state or the central.

9 KONKANI RAW TEXT CORPUS

Saurabh Varik, Rajesha N, Manasa G, Narayan Choudhary, L. Ramamoorthy

9.1 INTRODUCTION

Konkani is the principal and administrative language of Goa. Konkani is an Indo-Aryan language belonging to the Indo-European family of languages and is spoken along the western coast of India. The Konkani language is spoken widely in the western coastal region of India known as Konkan. This consists of the Konkan division of Maharashtra, the state of Goa, and the Uttara Kannada (formerly North Canara), Udupi, and Dakshina Kannada (formerly South Canara) districts of Karnataka, together with many districts in Kerala (such as Kasargod, Kochi, Alappuzha, Trivandrum, and Kottayam). Each region has a different dialect, pronunciation style, vocabulary, tone and sometimes, significant differences in grammar. It is a minority language in Karnataka, Maharashtra and Kerala, Dadra and Nagar Haveli, and Daman and Diu. Konkani is a member of the southern Indo-Aryan language group. It retains elements of Proto-Dravidian structures and shows similarities with both western and eastern Indo-Aryan languages. There are many fractured Konkani dialects, most of which are not mutually intelligible with one another.

Konkani is written in five scripts: Devanagari, Roman, Kannada, Malayalam, and Perso-Arabic. Because Devanagari is the official script used to write Konkani in Goa and Maharashtra, most Konkanis (especially Hindus) in those two states write the language in Devanagari. Konkani occupies the southernmost position in the Indo-Aryan linguistic continuum on the Indian peninsula . Towards the North and the North-East it merges gradually with Marathi , its closest kin. Towards the South and the South-East it touches Kannada , a language belonging to the Dravidian family.

Konkani text corpus is collected from various libraries in Goa mostly from Goa University, Panaji, Goa. The greater part of the text has been taken from Goa University library and Konkani Bhasha Mandal Campus library. LDC-IL tried to cover the entire category in its standard list. Some categories like novel, short stories have huge amount of books but some categories like physics, chemistry, economics have very less amount of books. Literary texts are easily available in Konkani but getting scientific text is very difficult. Some categories like epigraphy, finance, oceanology text are too rare in Konkani.

9.2 PECULIARITIES OF KONKANI TEXT

The Corpus of Konkani 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 that are narrative and it contains elements of fiction. Novels, short stories, plays are examples of literary text. Non-literary texts are texts 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.

Average word length of Konkani text is the average among all the scheduled languages of India. Konkani is highly agglutinative and morphologically rich language; hence the saturation level of Konkani i.e. the

new words coming into corpus for a unit amount of input is much higher compared to other languages. One needs to have much larger text corpora for good coverage of words.

9.3 DATA SAMPLING NOTES

9.3.1 Principles of Data Sampling

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

9.3.2 Field Works Undertaken

Konkani text corpus is collected from various libraries in Goa and mostly from the CIIL Library. The text materials were collected by conducting two field works undertaken in the period from 2010 to 2011. The greater part of the text has been taken from Goa University library, Konkani Bhasha Mandal and Goa Konkani Academy library.

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

- Goa University Campus Library, Taleigao, Panaji, Goa
- Konkani Bhasha Mandal, Margao, Goa
- Goa Konkani Academy library, Panaji, Goa
- CIIL Library, Mysore, Karnataka
- Goa State Central Library, Panaji, Goa

Collected text materials have been published at various places within Goa and other states of India such as Karnataka, Kerala, Maharashtra, Delhi as well as other countries such as Portugal, USA etc.

An attempt has been made to cover the entire category in its standard list. Some categories like novel, short stories have huge amount of books but some categories like physics, chemistry, economics have very less amount of books. Literary texts are easily available in Konkani but getting scientific text is very difficult. Some categories like epigraphy, finance, oceanology text are too rare in Konkani.

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

Some time Xerox attendents refused to photocopy randomly selected pages because of the long queue waiting and it takes up more time for them to turn the pages compared to continuous page photocopying they are accustomed to. It was another issue that the field worker/linguist had to carry a huge list of photocopy bundles with them which was many a times cumbersome to travel with.

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

9.3.3 Data Inputting

All the text has been typed in Unicode using the InScript Keyboard directly onto the XML files. The data has been inputted by Ms. M.Vidya, Harshith M.R., Shital (a native speaker of Konkani), Syeda Aliya Usmani, T.Renuka and Veena K R.

9.3.4 Validation and Normalization Workshops

A 5-day workshop was conducted at Linguistic Data Consortium in the year July 2010 with Dr. Madhavi Sardesai, Mrs. Priyadarshani Tadkodkar and Dr. Gunaji Desai from Department of Konkani, Goa University, Taleigao, Panaji, Goa as experts. The experts suggested that the Konkani text corpus should remain true to the text.

9.3.5 Proofreading

Konkani text data has been proofread by internal resource persons. The text has always been kept true to the printed material and typos, if any, occurring at the time of typing have only been corrected.

The printed materials collected for the corpus is contemporary , mainly published after 1990. Konkani alphabet refers to the five different scripts (Devanagari, Roman, Kannada, Malayalam and Perso-Arabic scripts) currently used to write the Konkani language. As of 1987, the "Goan Antruz dialect" in the Devanagari script has been declared Standard Konkani and promulgated as an official language in the Indian state of Goa [1][2]. As Konkani in the Roman script is not mandated as an official script by law, however, an ordinance passed by the government of Goa allows the use of Roman script for official communication. This ordinance has been put into effect by various ministries in varying degrees. For example, the Goa Panchayat Rules, 1996 stipulates that the various forms used in the election process must be in both the Roman and Devanagari script.

The rules for writing Konkani in the Devanagari script are elucidated in a book released by the Goa Konkani Academy titled kōṅkaṇī śuddhalēkhanācē nēm, while the rules for writing Konkani in the Roman script are elucidated in a book titled thomas sṭīvans koṅkaṇi kēndr Romi Lipi by writer Pratap Naik, released by Konkani singer Ullās Buyẽv at Dalgado Konkani Academy.

9.3.6 Data Extracted from Web Sites

Konkani News cropus data is extracted from News websites of *"Sunaparant" (*https://www.goacom.com*)* The news content was categorized based on the content of the text and archived. The period of selection of the news corpus ranges from 2009 to 2014.

9.4 TRANSLITERATIONS IN LDC-IL KONKANI TEXT CORPUS

For easy reference and uniformity of metadata, some entries in the metadata file, namely *'Title', 'Headline', 'Author', 'Editor', 'Translator'* are transliterated from Komkani to Roman letters. Numeric characters were transliterated from Konkani to Hindu-Arabic system.

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

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

Vowels and Vowel Signs															
Vowel	अ	आ	इ	ई	उ	ऊ	ऋ	ए	ऐ	ओ	औ	अं	अः	अँ	ऑ
Matra		ा	ि	ी	ु	ू	ृ	े	ै	ो	ौ	ं	ः	ँ	ॉ
Key	**a**	**A**	**i**	**I**	**u**	**U**	**x**	**E**	**ai**	**O**	**au**	**M**	**H**	**m'**	**ao**

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						
य	र	ऱ	ल	व	श	ष	स	ह	ळ	ऴ
ya	ra	Ra	la	va	sha	Sa	sa	Ha	La	Za

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

9.5 OVERVIEW OF REPRESENTED DOMAINS

LDC-IL Konkani Text Corpus size is: 39,95,611 words drawn from 282 different titles, including the extracts from newspapers. Konkani at present does not have any crawled text. We hope add some text by crawling/manually collecting in near future.

Text Type	Word Count	KeyStroke/Character Count
Typed+Cleaned	39,95,611	26,531,127
Crawled	Nil	Nil
Total	39,95,611	26,531,127

Table 9-1: Representation of the Typed and Crawled Konkani Text Corpus

The representation of the four major domains covered has been shown in the table below:

Domain	Domain Word Count	Percentage
Aesthetics	1770477	44.31%
Mass Media	2016151	50.46%
Science and Technology	104471	2.61%
Social Sciences	104512	2.62%
Total	3,995,611	100

Table 9-2: Representation of the Domains in Konkani Text Corpus

As each domain has several sub-domains, the following table shows the representation of the several domains, both within the domain and across all the domains.

9.5.1 AESTHETICS

The Aesthetics category of Konkani text corpus covers 22 sub-categories bearing a total of 17,70,477 words along with the overall percentage of 44.31%. The representational details are given in the table below.

#	Sub Category	Word Count	Percentage (within Subdomain).	Overall Percentage
1	Autobiographies	57996	3.28%	1.45%
2	Biographies	53833	3.04%	1.35%
3	Cinema	8590	0.49%	0.21%
4	Culture	2025	0.11%	0.05%
5	Fine Arts-Dance	7032	0.40%	0.18%
6	Fine Arts-Music	16431	0.93%	0.41%
7	Fine-Arts-Musical Instruments	3627	0.20%	0.09%

8	Fine Arts-Sculpture	2021	0.11%	0.05%
9	Folk Tales	43546	2.46%	1.09%
10	Folklore	54201	3.06%	1.36%
11	Handicrafts	1407	0.08%	0.04%
12	Literature-Children's Literature	34669	1.96%	0.87%
13	Literature-Criticism	92977	5.25%	2.33%
14	Literature-Diaries	38033	2.15%	0.95%
15	Literature-Epics	8006	0.45%	0.20%
16	Literature-Essays	177656	10.03%	4.45%
17	Literature-Letters	4438	0.25%	0.11%
18	Literature-Novels	163379	9.23%	4.09%
19	Literature-Plays	116601	6.59%	2.92%
20	Literature-Poetry	356	0.02%	0.01%
21	Literature-Science Fiction	20900	1.18%	0.52%
22	Literature-Short Stories	667172	37.68%	16.70%
23	Literature-Speeches	133760	7.56%	3.35%
24	Literature-Text Books (School)	20645	1.17%	0.52%
25	Literature-Travelogues	22480	1.27%	0.56%
26	Mythology	18696	1.06%	0.47%
	Total	**1770477**	100	**44.31%**

Table 9-3: Aesthetics Category Representation

Mass Media

The Mass Media category of Konkani text corpus covers 16 sub-categories bearing a total of 20,16,151 words along with the overall percentage of 50.46%. The representational details are given in the table below.

#	Sub Category	Word Count	Percentage (within Subdomain).	Overall Percentage
1	Article	3421	0.17%	0.09%
2	Cinema News	5379	0.27%	0.13%
3	Classifieds	4296	0.21%	0.11%
4	Discussions	31936	1.58%	0.80%
5	Editorial	243658	12.09%	6.10%
6	General News	556183	27.59%	13.92%
7	Interviews	22868	1.13%	0.57%
8	Letters	23269	1.15%	0.58%
9	Obituary	5625	0.28%	0.14%
10	Political	109878	5.45%	2.75%
11	Religous/Spiritual News	10712	0.53%	0.27%
12	SMS	1710	0.08%	0.04%
13	Social	423823	21.02%	10.61%
14	Speeches	12274	0.61%	0.31%
15	Sports News	559305	27.74%	14.00%

16	Weather	1814	0.09%	0.05%
	Total	2016151	100	50.46%

Table 9-4: Mass Media Category Representation

Science and Technology

The Science and Technology category of Konkani text corpus covers 09 sub-categories bearing a total of 10,44,71 words along with the overall percentage of 2.61%. The representational details are given in the table below.

#	Sub Category	Word Count	Percentage (within Subdomain).	Overall Percentage
1	Agriculture	782	0.75%	0.02%
2	Botany	9209	8.81%	0.23%
3	Environmental Science	21426	20.51%	0.54%
4	Geology	974	0.93%	0.02%
5	Homeopathy	2178	2.08%	0.05%
6	Medicine	45661	43.71%	1.14%
7	Psychology	1970	1.89%	0.05%
8	Sexology	15587	14.92%	0.39%
9	Yoga	6684	6.40%	0.17%
	Total	**104471**	**100**	2.61%

Table 9-5: Science and Technology Category Representation

Social Sciences

The Social Sciences category of Konkani text corpus covers 14 sub-categories bearing a total of 10,45,12 words along with the overall percentage of 2.62%. The representational details are given in the table below.

#	Sub Category	Word Count	Percentage (within Subdomain).	Overall Percentage
1	Archeology	866	0.83%	0.02%
2	Education	6620	6.33%	0.17%
3	Epigraphy	748	0.72%	0.02%
4	Fisheries	251	0.24%	0.01%
5	Food and Wellness	5158	4.94%	0.13%
6	Geography	4765	4.56%	0.12%
7	Health and Family Welfare	3875	3.71%	0.10%
8	History	21194	20.28%	0.53%
9	Journalism	3615	3.46%	0.09%
10	Linguistics	26178	25.05%	0.66%
11	Political Science	5953	5.70%	0.15%
12	Religion/Spiritual	11814	11.30%	0.30%
13	Sociology	3598	3.44%	0.09%
14	Sports	9877	9.45%	0.25%

	Total	**104512**	**100**	2.62%

Table 9-6: Social Sciences Category Representation

9.6 COPYRIGHT CONSENTS

The Konkani text corpus has been collected from various sources and the copyright for the same stays with different sources. However, for the purposes of this corpus, consents have been sought from all the stakeholders. Most of the copyrights (around 83%) belong to private parties with only 17% belonging to the government agencies, either state or the central.

10 MAITHILI RAW TEXT CORPUS

Dinesh Mishra, Rajesha N, Manasa G, Narayan Choudhary, L. Ramamoorthy

10.1 INTRODUCTION

Maithili is an Indio-Aryan language, a direct descendent of Sanskrit, which is spoken in the states of Bihar, Jharkhand and Nepal. It is one of the scheduled languages of India. This is the most intelligible language in India, as according to the 2011 census of India, five million people listed it as their mother tongue.

The name Maithili is derived from the word Mithila, an ancient Kingdom of which King Janaka was the ruler (see Ramayana). Maithili is also of the name of Sita, the wife of King Rama and daughter of King Janaka. Scholars in Mithila used Sanskrit for their literary work and Maithili was the language of the common folk (Abahatta).

It can be observed that Mithila region has been rich in cultural heritage which has produced a distinct cultural landscape over the years. Its evolution has been influenced by geographical isolation surrounded by the three big rivers and lofty mountains. The region remained secluded peaceful and least influenced tract.

Vidyapati, also known by the sobriquet of Maithil Kavi Kokil (the poet cuckoo of Maithili), was a Maithili poet and a Sanskrit writer.

Vidyapati's poetry was widely influential in centuries to come, in the Hindustani as well as Bengali, Maithili, Newari, less actively in Nepali language and other Eastern literary traditions. The language at the time of Vidyapati, the Prakrit-derived late Abahatta, had just begun to transition into early versions of the Eastern languages Maithili, Nepali, Bengali, Oriya, etc. Thus, Vidyapati's influence on making these languages has been described as "analogous to that of Dante in Italy and Chaucer in England"

The influence of the lyrics of Vidyapati on the love of Radha and Krishna on the Bengali poets of the medieval period was so overwhelming that they largely imitated it. As a result, an artificial literary language, known as Brajabuli was developed in the sixteenth century. Brajabuli is basically Maithili (as prevalent during the medieval period) but its forms are modified to look like Bengali. The medieval Bengali poets, Gobindadas Kabiraj, Jnandas, Balaramdas and Narottamdas composed their padas (poems) in this language. Rabindranath Tagore composed his Bhanusingha Thakurer Padabali (1884) in a mix of Western Hindi (Braj Bhasha) and archaic Bengali and named the language Brajabuli as an imitation of Vidyapati (he initially promoted these lyrics as those of a newly discovered poet, Bhanusingha). Other 19th-century figures in the Bengal Renaissance like Bankim Chandra Chatterjee have also written in Brajabuli.

Vidyapati is known as one of the main pillars of 'Bhakti-Parampara', along with 'Shringar-Parampara' of Indian literature and as the paramount poet of Maithili. The appearance of medieval Maithili language can be seen in their works. They have been accepted as Vaishnava, Shaiva and Shakta Bhakti bridges. By making the people of Mithila the “Desil Biyana Sabh Jan Mittha”, they have made a great effort to revive the public awareness of the language of the North-Bihar.

In the songs used in Mithilanchal's folklore, the compositions of Vidyapati are still alive and the Pagi compositions in devotion and ritual are alive. Kirtipataka and kirtilata are their immortal creations.

Maithili dates back to the 14th century. The “Varna Ratnakara” is the earliest known prose text, preserved from 1507, and is written in “Mithilikshar” script. Maithili was traditionally written in their own script which is known as Mithilakshar or Tirhuta. This script is similar to Bengali-Assamese script. Devanagari script started being used most commonly used since the start of the second half of the 20th century. It was also written in the local variant of Kaithi script. The Tirhuta (Mithilakshar) and Kaithi scripts are both currently included in Unicode.

In 2003, Maithili was included in the Schedule of the Indian Constitution as a recognized Indian language, which allows it to be used in education, government, and other official contexts in India. The Maithili language is included as an optional paper in the UPSC Exam.

Mithila was a kingdom in ancient India. It is believed that this is the lowland area of northern Bihar and Nepal which was known as Mithila. Mithila's public life has been running for many centuries, which has been known outside of India for its intellectual tradition. The main language of this area is Maithili. In Hindu religious texts, it is first mentioned in the Sathpath Brahmin and the explicit mention is found in the Valmiki Ramayana. Mithila is mentioned in Mahabharata, Ramayana, Purana and Jain and Buddhist texts.

After the Magadha in Mahabharata (north) the status of Mithila has been deemed to be describing Shri Krishna, Arjun and Bhima of Magadha Yatra, first to cross the Saryu river and to cross the eastern Koshal Pradesh and then Mahashon, Gandakki and Sadanira. It seems obvious that at that time the status of Mithila in the north of Magadha has been assumed. Vajjee Pradesh (Vaishali state) was included under Mithila. In this mention of the Mahabharata, the area of Mithila's border area is mentioned elsewhere, i.e., Gandkhi in the west and the areas of the Gangas in the south are just as indicated.

Explaining the boundary of Mithila (Chauhaddi) in Vrudavishnupuran, it has been dictated as-

Devanagari:	IPA:
कौशिकीन्तु समारभ्य गण्डकीमधिगम्यवै। योजनानि चतुर्विंश व्यायामः परिकीर्त्तितः॥ गङ्गा प्रवाहमारभ्य यावद्धैमवतम्वनम् । विस्तारः षोडशप्रोक्तो देशस्य कुलनन्दन॥	kɔʃiki:nt̪u səmarəbʱjə gənɖəki:məɖ̪ʱigəmjəvɛ। joɟənani cət̪urviŋʃə vjajaməh pəriki:rt̪t̪it̪əh॥ gəŋgɑ prəvaɦəmarəbʱjə jɑvəɖ̪ɖ̪ʱɛməvət̪əmvənəm। vist̪arəh ʂodəʃəprokt̪o ɖ̪eʃəsjə kulənənɖənə॥

Roman Transliteration: kaushikIntu samArabhya gaNDakImadhigamyavai. yOjanAni caturviMsha vyAyAmaH parikIrttitaH.. gang'gA pravAhamArabhya yAvaddhaimavatamvanam . vistAraH SODashaprOktO dEshasya kulanandana..	Gloss: That is, from the beginning of Kosi in the east, 24 planes to Gandki in the west and from the river Ganga in the south, to extension of 16 plans till the Himalaya forest (Tarai region) in the north is Mithila.

And, Mahakavi Chanda Jha refers to the above mentioned verse as Maithili, describing the boundary of Mithila as-

Devanagari: गंगा बहथि जनिक दक्षिण दिशि पूब कौशिकी धारा। पश्चिम बहथि गंडकी उत्तर हिमवत वन विस्तारा॥	IPA: gəŋgɑ bəɦəʈ̪ʰi ɟənikə ɖ̪əkʂiɳə ɖ̪iʃi puːbə kɔʃiki: ɖ̪ʱɑrɑ. pəʃcimə bəɦəʈ̪ʰi gəŋɖ̪əki: uʈ̪ʈ̪ərə ɦiməvəʈ̪ə vənə visʈ̪ɑrɑ..
Roman Transliteration: gaMgA bahathi janika dakSiNa dishi pUba kaushikI dhArA. pashcima bahathi gaMDakI uttara himavata vana vistArA..	Gloss: Ganga flow at the south, and Kaushiki at the east. At the west it is the Gandaki river, and to the North the forest of Himalayas.

In India, Maithili is Spoken mainly in Bihar and Jharkhand in the districts of Darbhanga, Madhubani, Samastipur, Muzaffarpur, Sitamarhi, Begusarai, Khagaria, Purnia, Katihar, Kishanganj, Sheohar, Bhagalpur, Madhepura, Araria, Supaul, Vaishali, Saharsa (Bihar) Ranchi, Bokaro, Jamshedpur, Dhanbad, and Deoghar (Jharkhand). The geographic region comprising of these districts is also called as Mithilanchal Region. Darbhanga and Madhubani constitute cultural and linguistic centers. Native speakers also reside in Patna, Delhi, Kolkata, Mumbai and Bengaluru.

In 1965, Maithili was officially accepted by Sahitya Academy, an organization dedicated to the promotion of Indian literature. In March 2018, Maithili received the second official language status in the Indian state of Jharkhand.

In the 19th century, linguistic scholars considered Maithili as a dialect of Bihari languages and grouped it with other languages spoken in Bihar. Hoernle compared it with Gaudian languages and recognized that it

shows more similarities with Bengali languages than with Hindi. Grierson recognized it as a distinct language and published the first grammar in 1881.

Presently Maithili language is predominately written in the Devanagari. Mithilakshar Script is also in practice. Both the Scripts are Left to Right scripts which are descendent of Brahmi script. The Devanagari script is also used to write several other languages of India and neighbouring countries such as Nepal. The dataset prepared for LDC-IL Maithili Text Data is in Devanagari script.

Maithili is written in Devenagari script, a Left to Right script which is a descendent of Brahmi script. The script is also used to write Maithili, Nepali, Bhojpuri, Rajasthani, Chhattisgarhi, Santali, Kashmiri, Konkani, Sindhi, Dogri, Bodo, Newar, Awadhi, Magahi, Haryanvi, Bhili, Mundari, Pali and Sanskrit as their sole script or one of the scripts.

Maithili text corpus is collected from various libraries in Darbhanga, Madhubani, Patna, Saharsa etc. The greater part of the text has been taken from L.NM.U. Central Library Darbhanga, Local Library, Local Author and Publisher. LDC-IL tried to cover the entire category in its standard list. Some categories like novel, short stories have huge amount of books but some categories like physics, chemistry, economics, agriculture have very less amount of books. Literary texts are easily available in Maithili but getting scientific text is very difficult. Some categories like epigraphy, finance, oceanology text are too rare in Maithili.

10.2 PECULIARITIES OF MAITHILI TEXT

The Corpus of Maithili text can be broadly classified into two: literary text and non-literary text. These two explicitly show their differences in terms of the types and the source of the word usage and variety that it brings into corpus. Literary texts are mainly narratives, and more or less they contain the elements of fiction. Some instances of such texts are: novels, short stories and plays. On the other hand the non-literary texts convey information based on their primary purpose. These range from the articles in academic journals, legal documents to the texts of various scientific or technical subjects. The literary texts have emotional elements, cultural information, dialectical variations, ambiguity etc. whereas the non-literary texts show a tendency of usage of technical and scientific terms, foreign words etc.

Maithili is highly agglutinative and morphologically rich language; hence the saturation level of Maithili i.e. the new words coming into corpus for a unit amount of input is much higher compared to other languages. One needs to have much larger text corpora for good coverage of words.

10.2.1 Orthographic variation in devnagari/MAiTHiLi

A glyph has no intrinsic meaning and it conveys distinctions in form. In information technology, a glyph is a graphic symbol that provides the appearance or form for a character. Time to time the user or developers made small variation in Devanagari script and the same changes come into in Maithili. These were अ, झ, ण, श. It was not unique feature of Maithili, but it made small changes in use of Maithili orthography system. Besides that, Maithili has its typical orthography, which is called ‘Shaja’ [A publisher, which is from Darbhanga, Madhubani Bihar,, i.e., /dz/( ) jha.

10.3 DATA SAMPLING NOTES

10.3.1 Principles of Data Sampling

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

10.3.2 Field Works Undertaken

1. Maithili text corpus is collected from various libraries in Darbhanga, Madhubani, Patna and Saharsa. The text materials were collected by conducting three field works undertaken in the period from 2011 to 2012. The greater part of the text has been taken from L.N.M.U. Central library Darbhanga, Patna university library, Patna, Chetna Samiti, Vidayapati Bhawan, Patna(Bihar) and various public library and Author.

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

2. L.N.M.U.Central library Darbhanga (Bihar)
3. CM College library, Darbhanga (Bihar)
4. R.K.College, library, Madhubabi (Bihar)
5. ChetnaSamiti, Vidayapati Bhawan, Patna (Bihar)
6. Patna university library, Patna (Bihar)
7. Local Author and Publisher (Bihar)
8. Public library (Bihar)
9. Central Institute of Indian Language Library, Mysore

Collected text materials have been published at various places within Bihar and other states of India such as Uttar Pradesh, Delhi, Calcutta, as well as other countries such as Nepal etc.

An attempt has been made to cover the entire category in its standard list. Some categories like novel, short stories have huge amount of books but some categories like physics, chemistry, economics have very less amount of books. Literary texts are easily available in Maithili but getting scientific text is very difficult. Some categories like epigraphy, finance, oceanology text are too rare in Maithili.

Collecting text data from the field is a difficult job. Most of the libraries do not allow to take 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 to take many books at a time, the photocopy kiosk had issues as there was a long queue.

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

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

10.3.3 Data Inputting

All the text has been typed in Unicode using the InScript Keyboard directly onto the XML files. The data has been inputted by Ms. Mamatha M, Ms. Radhika.M, Ms. Veena K.R , Ms. M.Vidya, Ms. H. S. Rupa, native speakers of Kannada.

10.3.4 Validation and Normalization Workshops

A Short term goal oriented project – Text corpus cleaning from 29th July to 30th to August, 2013, LDC-IL, CIIL, Mysore had been organized for cleaning Maithili raw texts.

10.3.5 Proofreading

Maithili text data has been proofread by internal resource persons. The text has always been kept true to the printed material and typos, if any, occurring at the time of typing have only been corrected.
The printed materials collected for the corpus are contemporary, mainly published after 1990.

10.4 TRANSLITERATIONS IN LDC-IL MAITHILI TEXT CORPUS

For easy reference and uniformity of metadata, some entries in the metadata file, namely *'Title', 'Headline', 'Author', 'Editor', 'Translator'* are transliterated from Maithili to Roman letters.

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

LDC-IL Transliteration Schema
Devanagari characters to Roman and Maithili Numerals to Roman

Vowels and Vowel Signs														
अ	आ	इ	ई	उ	ऊ	ऋ	ए	ऐ	ऑ	ओ	औ			
	ा	ि	ी	ु	ू	ृ	े	ै	ॉ	ो	ौ	ं	ः	ँ
a	A	i	I	u	U	x	e	ai	ao	o	au	M	H	m'

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

य	र	ल	व	श	ष	स	ह	ड़	ढ़
ya	Ra	la	va	sha	Sa	sa	ha	D	Dh

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

10.5 OVERVIEW OF REPRESENTED DOMAINS

LDC-IL Maithili Text Corpus size is: 5,316,552 Words and character count is 29,658,942drawn from 499 different titles, including the extracts from Magazine and newspapers.

The representation of the five major domains covered has been shown in the table below:

Domain	Word Count	Percentage
Aesthetics	3,897,264	73.30%
Commerce	50,975	0.96%
Mass Media	1,253,090	23.57%
Science and Technology	3,136	0.06%
Social Sciences	112,087	2.11%
Total	**5,316,552**	**100%**

Table 10-1 Representation of the Domains in Maithili Text Corpus

As each domain has several sub-domains, the following tables show the representations of each subdomain where the number of subcategories that fall under the same domain along with their total word count, percentage within the subdomain as well as the overall percentage is provided.

10.5.1 Aesthetics

The Aesthetics category of Maithili text corpus covers 18 sub-categories bearing a total of 38, 97,264 words along with the overall percentage of 73.30%. The representational details are given in the table below.

Sub Category	Word Count	Percentage (within Subdomain).	Overall Percentage
Autobiographies	80559	2.07%	1.52%
Biographies	263068	6.75%	4.95%
Cinema	14697	0.38%	0.28%
Culture	10096	0.26%	0.19%
Folk Tales	4214	0.11%	0.08%
Folklore	54453	1.40%	1.02%
Humor	81893	2.10%	1.54%
Literature-Criticism	1227895	31.51%	23.10%
Literature-Diaries	12615	0.32%	0.24%

Literature-Epics	9570	0.25%	0.18%
Literature-Essays	567462	14.56%	10.67%
Literature-Letters	18449	0.47%	0.35%
Literature-Novels	310239	7.96%	5.84%
Literature-Plays	133898	3.44%	2.52%
Literature-Science Fiction	6660	0.17%	0.13%
Literature-Short Stories	1036313	26.59%	19.49%
Literature-Speeches	13203	0.34%	0.25%
Literature-Travelogues	51980	1.33%	0.98%
Total	**3897264**	**100%**	**73.30%**

Table 10-2 Aesthetics Category Representation

10.5.2 Commerce

The Commerce category of Maithili text corpus covers 06 sub-categories bearing a total of 50,975 words along with the overall percentage of 0.96%. The representational details are given in the table below.

Sub Category	Word Count	Percentage (within Subdomain).	Overall Percentage
Banking	165	0.32%	0.00%
Business	44788	87.86%	0.84%
Career and Employment	1322	2.59%	0.02%
Finance	1122	2.20%	0.02%
Management	507	0.99%	0.01%
Share Market	3071	6.02%	0.06%
Total	**50975**	**100%**	**0.96%**

Table 10-3 Commerce Category Representation

10.5.3 Mass Media

The Mass Media category of Maithili text corpus covers 08 sub-categories bearing a total of 12, 53,090 words along with the overall percentage of 23.57%. The representational details are given in the table below.

Sub Category	Word Count	Percentage (within Subdomain).	Overall Percentage
Discussions	2020	0.16%	0.04%
Editorial	241777	19.29%	4.55%
General News	961554	76.73%	18.08%
Interviews	3285	0.26%	0.06%
Political	37646	3.00%	0.71%
Social	2459	0.20%	0.05%
Sports News	4349	0.35%	0.08%
Total	**1253090**	**100%**	**23.57%**

Table 10-4 Mass Media Category Representation

10.5.4 Science and Technology

The Science and Technology category of Maithili text corpus covers 03 sub-categories bearing a total of 3,136 words along with the overall percentage of 0.06%. The representational details are given in the table below.

Sub Category	Word Count	Percentage (within Subdomain).	Overall Percentage
Astronomy	1241	39.57%	0.02%
Homeopathy	302	9.63%	0.01%
Medicine	1593	50.80%	0.03%
Total	**3136**	**100%**	**0.06%**

Table 10-5 Science and Technology Category Representation

10.5.5 Social Sciences

The Social Sciences category of Maithili text corpus covers 08 sub-categories bearing a total of 1, 12,087 words along with the overall percentage of 2.11%. The representational details are given in the table below.

Sub Category	Word Count	Percentage (within Subdomain).	Overall Percentage
Education	1079	0.96%	0.02%
Health and Family Welfare	1681	1.50%	0.03%
History	10808	9.64%	0.20%
Linguistics	10341	9.23%	0.19%
Philosophy	12612	11.25%	0.24%
Religion/Spiritual	22111	19.73%	0.42%
Sociology	6990	6.24%	0.13%
Sports	46465	41.45%	0.87%
Total	**112087**	**100%**	**2.11%**

Table 10-6 Social Sciences Category Representation

10.6 COPYRIGHT CONSENTS

The Maithili text corpus has been collected from various sources and the copyright for the same stays with different sources. However, for the purposes of this corpus, consent has been sought from all the stakeholders. Most of the copyrights (around 51.24%) belong to private parties with only48.76% belonging to the government agencies, either state or the central.

10.7 REFERENCE

Brass, P. R. (2005). *Language, Religion and Politics in North India*. Lincoln: iUniverse. ISBN 0-595-34394-5.

Baleshwar Thakur, George Pomeroy, Chris Cusack, Sudhir K Thakur. *City Society and Planning*. Volume (II), pp-429, ISBN-10:81-8069-460-7

Chaudhary, Radhakrishna. (2010). *A Survey of Maithili Literature*. (pp-12) ISBN: 978-93-80538-36-5;

Grierson, George Abraham. *Seven Grammars of the Dialects and Sub-dialects of the Bihari Language.* (1883–87). ISBN 81-7835-451-9

11 MALAYALAM RAW TEXT CORPUS

Rejitha K.S., Saritha S.L., Sajila S., Rajesha N., Manasa G., Narayan Choudhary, L. Ramamoorthy

11.1 INTRODUCTION

Malayalam is the principal and administrative language of Kerala, Union Territory of Lakshadweep and Mahé district, one of the four districts of the Union Territory of Puducherry. Malayalam is written in Malayalam Script. It is a unicase script, meaning it does not have a case distinction. It is written from left to right direction. The modern Malayalam script has evolved from the Grantha alphabet which was also used to write Sanskrit and Tulu. With the objective to simplify the script for print and typewriting technology of that time, the Government of Kerala reformed the orthography of Malayalam by a government order to the education department by reducing the number of glyphs required. The reformed script came into effect in 1971 thereby reducing the number of glyphs required. Print media almost entirely uses reformed orthography. Primary education introduces the Malayalam writing to the pupils in reformed script only and the books are printed accordingly. The script is also used to write several minority languages such as Paniya, Betta Kurumba, Ravula etc.

Malayalam is written in other scripts as well. 'Arabi Malayalam' is a variant of 'Arabic' script. 'Syriac Malayalam' is a variant form of 'Syriac script'. LDC-IL Malayalam text corpus is collected in Malayalam script of contemporary usage.

Malayalam text corpus is collected from various libraries in Kerala mostly from Thiruvananthapuram. The greater part of the text has been taken from University of Kerala library and University of Kerala Campus library. LDC-IL tried to cover the entire category in its standard list. Some categories like novel, short stories have huge amount of books but some categories like physics, chemistry, economics have very less amount of books. Literary texts are easily available in Malayalam but getting scientific text is very difficult. Some categories like epigraphy, finance, oceanology text are too rare in Malayalam.

11.2 PECULIARITIES OF MALAYALAM TEXT

The Corpus of Malayalam 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 that are narrative and it contains elements of fiction. Novels, short stories, plays are examples of literary text. Non-literary texts are texts 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.

Average word length of Malayalam text is the highest among all the scheduled languages of India. Malayalam is highly agglutinative and morphologically rich language; hence the saturation level of Malayalam i.e. the new words coming into corpus for a unit amount of input is much higher compared to other languages. One needs to have much larger text corpora for good coverage of words.

11.2.1 Chillu Letters of Malayalam

A chillu or *chillaksharam* represents pure independent consonants, without help of a '*chandrakala*' unlike a consonant represented by an ordinary consonant letter; these consonants are never followed by an inherent vowel.

In the earlier versions of Unicode Chillu letters did not have a separate value and were represented with the combination of Zero Width Joiner (ZWJ)

MALAYALAM LETTER CHILLU 'n' – ൻ	ന (na) + ് (Chandrakala) + ZWJ
MALAYALAM LETTER CHILLU 'ṇ' – ൺ	ണ (ṇa) + ് (Chandrakala) + ZWJ
MALAYALAM LETTER CHILLU 'r' – ർ	ര (ra) + ് (Chandrakala) + ZWJ
MALAYALAM LETTER CHILLU 'l' – ൽ	ല (la) + ് (Chandrakala) + ZWJ
MALAYALAM LETTER CHILLU 'ḷ' – ൾ	ള (ḷa) + ് (Chandrakala) + ZWJ

Table 11-1: Two Variations of representation of Chillu

Zero Width Joiner (ZWJ) and Zero Width Non-Joiner (ZWNJ) characters have no value of their own. They are supposed to be only font directives, directing a font to select from two or more semantically same renderings. When it comes to Malayalam, ZWJ becomes an alien language construct introduced to Malayalam by Unicode to produce Chillu letters. Thus, it is possible to produce two semantically different words, which differ only by ZWJ in their Unicode representation.

In the following examples, words differ only by ZWJ.

The word 'അവന്' is with visible *Chandrakala* after 'n' and pronounced as 'avanə'. This word means 'for him'.

The word 'അവൻ' is with Chillu 'n' and pronounced as 'avan'. This word means 'he'

Search algorithms used to fail to filter out when a word was searched in Unicode text, as normally search algorithms ignore ZWJ and ZWNJ because it should not care about the rendering of the word. As a fix if the search algorithm could match joiners, only in the case of Malayalam. Then the algorithm will not match those words that are semantically same but rendered differently by using or omitting a joiner (ZWJ or ZWNJ). For example, search for the Malayalam word 'മത്സരം' (maṯsaram) will not match 'മത്സരം' (maṯsaram), because later is written using ZWNJ. Semantically both words are same with a spell variation in orthographic representation. This inconsistency may lead to problems in developing language tools like morphological analyzers, grammar checkers etc.

To counter this inconsistency Unicode allotted separate chillu letters. LDC-IL Malayalam text data is on par with the current Unicode standards of chillu letters. LDC-IL Malayalam text data contains standard chillu letters.

11.3 DATA SAMPLING NOTES

11.3.1 Principles of Data Sampling

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

11.3.2 Fieldworks Undertaken

Malayalam text corpus is collected from various libraries in Kerala, mostly from Thiruvananthapuram. The text materials were collected by conducting four fieldworks undertaken in the period from 2009 to 2012. The greater part of the text has been taken from Kerala University library and Kerala University Campus library.

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

1. Kerala University Campus Library, Kariavattom
2. Kerala University Library, Thiruvananthapuram
3. Agricultural College Library Science and Technology, Thiruvananthapuram
4. International Centre for Kerala Studies, Kariavattom
5. Dept. of Islamic Studies, University of Kerala, Kariavattom
6. Govt. Ayurveda College Library, Thiruvananthapuram
7. Institute of Distance Education Library, Kariavattom
8. Linguistics Department Library, University of Kerala, Kariavattom
9. Southern Regional Language Centre Library, Mysore

Collected text materials have been published at various places within Kerala and other states of India such as Karnataka, Tamilnadu, Maharashtra, Delhi as well as other countries such as Bahrain, USA etc.

An attempt has been made to cover the entire category in its standard list. Some categories like novel, short stories have huge amount of books but some categories like physics, chemistry, economics have very less amount of books. Literary texts are easily available in Malayalam but getting scientific text is very difficult. Some categories like epigraphy, finance, oceanology text are too rare in Malayalam.

Collecting text data from the field is a difficult job. Most of the libraries do not allow to 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 to take many books at a time, the photocopy kiosk had issues as there was a long queue.

Sometime Xerox attendants refused to photocopy randomly selected pages because of the long queue waiting and it takes up more time for them to turn the pages compared to continuous page photocopying they are accustomed to. It was another issue that the fieldworker/linguist had to carry a huge list of photocopy bundles with them which was many times cumbersome to travel with.

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

11.3.3 Data Inputting

All the text has been typed in Unicode using the InScript Keyboard directly onto the XML files. The data has been inputted by Ms. Ramya K., a native speaker of Malayalam.

11.3.4 Validation and Normalization Workshops

A 5-day workshop was conducted at Linguistic Data Consortium from 6-June-2011 to 10-June-2011 with Dr. A. Rose Mary and Dr. S. A. Shanavas from Department of Linguistics, Kerala University, Thiruvananthapuram as experts. The experts suggested that the Malayalam text corpus should remain true to the text.

11.3.5 Proofreading

Malayalam text data has been proofread by internal resource persons. The text has always been kept true to the printed material and typos, if any, occurring at the time of typing have only been corrected.

The printed materials collected for the corpus is contemporary, mainly published after 1990. Hence the text material available is with the reformed script which came into effect from 15th April 1971 (Vishu Day) by the Government of Kerala order no. G. O. (P) 37/71/Edn. Dated 23rd March 1971. The Government order is published with the booklet titled "Lipiparishkaranam" that gives guidelines on how the script has been modified and how the spelling variations are to be taken care of. The government order and the booklet is available online at https://unicode.org/L2/L2008/08039-kerala-order.pdf

11.3.6 Data Extracted from Websites

Malayalam news corpus data is extracted from news websites of "Malayala Manorama" (https://www.manoramaonline.com) , "Mangalam" (www.mangalam.com/), "Mathrubhumi" (https://www.mathrubhumi.com/), "Metro Vartha" (http://www.metrovaartha.com) and "Vyga News" (http://www.vyganews.com/) . The news content was categorized based on the content of the text and archived. The period of selection of the news corpus ranges from 2008 to 2012.

11.4 TRANSLITERATIONS IN LDC-IL MALAYALAM TEXT CORPUS

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

LDC-IL Transliteration Schema

Malayalam characters to Roman and Malayalam Numerals to Hindu-Arabic

Vowels															
അ	ആ	ഇ	ഈ	ഉ	ഊ	ഋ	ൠ	ഌ	ൡ	എ	ഏ	ഐ	ഒ	ഓ	ഔ
	ാ	ി	ീ	ു	ൂ	ൃ	ൄ	ൢ	ൣ	െ	േ	ൈ	ൊ	ോ	ൌ
A	A	i	I	u	U	x	X	q	Q	e	E	ai	o	O	au

Consonants				
ക	ഖ	ഗ	ഘ	ങ

Symbols			
◌ൺ	◌ൻ	ം	ഃ

Ka	kha	ga	gha	ng'a
ച	ഛ	ജ	ഝ	ഞ
Ca	cha	ja	jha	nj'a
ട	ഠ	ഡ	ഢ	ണ
Ta	Tha	Da	Dha	Na
ത	ഥ	ദ	ധ	ന
Ta	tha	da	dha	na
പ	ഫ	ബ	ഭ	മ
Pa	pha	ba	bha	ma

m`	m'	M	H

യ	ര	ല	വ	ശ	ഷ	സ	ഹ	ള	ഴ	റ	ഩ
Ya	ra	la	va	sha	Sa	sa	ha	La	Za	Ra	TTTa

Chillu Letters								
ൺ	ൻ	ർ	ൽ	ൾ	ൿ	ഺ	ൟ	഻
N'	n'	R'	l'	L'	k'	M'	y'	Z'

Numerals (Malayalam to Hindu-Arabic)									
൦	൧	൨	൩	൪	൫	൬	൭	൮	൯
0	1	2	3	4	5	6	7	8	9

The greyed out characters are obsolete. They may rarely present in the current LDC-IL corpus.

Table 11-2: LDC-IL Transliteration Schema of Malayalam to Roman

11.5 OVERVIEW OF REPRESENTED DOMAINS

LDC-IL Malayalam Text Corpus size is: 63,70,954 words drawn from 1,119 different titles, including the extracts from newspapers. The representation of the six major domains covered has been shown in the table below:

Domain	Word Count	Percentage
Aesthetics	25,77,090	40.45%
Commerce	3,13,135	4.92%
Official Documents	7,733	0.12%
Social Sciences	8,75,568	13.74%
Mass Media	21,35,621	33.52%
Science & Technology	4,61,807	7.25%
Total	63,70,954	100

Table 11-3: Representation of the Domains in Malayalam Text Corpus

As each domain has several sub-domains, the following table shows the representation of the several domains, both within the domain and across all the domains.

11.5.1 Aesthetics

The Aesthetics Category of LDC-IL Malayalam text corpus covers 25 subdomains. The details of the representation of subdomains are given below:

Subdomain	Word Count	% (within Subdomain)	Overall Percentage
Autobiographies	1,44,572	5.61%	2.27%
Biographies	1,18,292	4.59%	1.86%
Cinema	2,93,878	11.40%	4.61%
Culture	51,150	1.98%	0.80%
Fine Arts-Dance	20,484	0.79%	0.32%
Fine Arts-Drawing	3,053	0.12%	0.05%
Fine Arts-Hobbies	143	0.01%	0.00%
Fine Arts-Music	22,440	0.87%	0.35%
Fine Arts-Musical Instruments	309	0.01%	0.00%
Folklore	13,297	0.52%	0.21%
Humour	20,732	0.80%	0.33%
Mythology	29,559	1.15%	0.46%
Literary Texts	7,090	0.28%	0.11%
Literature-Children's Literature	4,843	0.19%	0.08%
Literature-Criticism	90,465	3.51%	1.42%
Literature-Epics	2,852	0.11%	0.04%
Literature-Essays	4,85,024	18.82%	7.61%
Literature-Letters	2,754	0.11%	0.04%
Literature-Novels	6,59,531	25.59%	10.35%
Literature-Plays	48,123	1.87%	0.76%
Literature-Poetry	3,397	0.13%	0.05%
Literature-Short Stories	4,42,473	17.17%	6.95%
Literature-Speeches	3,357	0.13%	0.05%
Literature-Travelogues	1,06,258	4.12%	1.67%
Photography	3,014	0.12%	0.05%

Table 11-4 Aesthetics Category Representation

11.5.2 Commerce

The Commerce Category of LDC-IL Malayalam text corpus covers 4 subdomains. The details of the representation of subdomains are given below:

Subdomain	Word Count	% (within Subdomain)	Overall Percentage
Business	2,89,061	92.31%	4.54%
Management	5,677	1.81%	0.09%
Share Market	1,438	0.46%	0.02%
Tourism	16,959	5.42%	0.27%

Table 11-5 Commerce Category Representation

11.5.3 Official Documents

The Official Documents Category of LDC-IL Malayalam text corpus covers 2 subdomains. The details of the representation of subdomains are given below:

Subdomain	Word Count	% (within Subdomain)	Overall Percentage
Administration	4,668	60.36%	0.07%
Police Documents	3,065	39.64%	0.05%

Table 11-6 Official Documents Category Representation

11.5.4 Social Sciences

The Social Sciences Category of LDC-IL Malayalam text corpus covers 20 subdomains. The details of the representation of subdomains are given below:

Subdomain	Word Count	% (within Subdomain)	Overall Percentage
Anthropology	14,409	1.65%	0.23%
Economics	43,703	4.99%	0.69%
Education	30,926	3.53%	0.49%
Epigraphy	3,587	0.41%	0.06%
Fisheries	9,237	1.05%	0.14%
Food and Wellness	6,157	0.70%	0.10%
Geography	27,488	3.14%	0.43%
Health and Family Welfare	94,813	10.83%	1.49%
History	1,77,724	20.30%	2.79%
Home Science	4,141	0.47%	0.06%
Journalism	63,679	7.27%	1.00%
Law	18,980	2.17%	0.30%
Library Science	11,640	1.33%	0.18%
Linguistics	47,616	5.44%	0.75%
Philosophy	53,122	6.07%	0.83%
Physical Education	2,680	0.31%	0.04%
Political Science	1,26,328	14.43%	1.98%
Public Administration	17,796	2.03%	0.28%
Religion/Spiritual	88,118	10.06%	1.38%
Sociology	18,959	2.17%	0.30%
Sports	14,465	1.65%	0.23%

Table 11-7 Social Sciences Category Representation

11.5.5 Mass Media

The Mass Media Category of LDC-IL Malayalam text corpus covers 9 subdomains. The details of the representation of subdomains are given below:

Subdomain	Word Count	%	Overall

		(within Subdomain)	Percentage
Article	11,525	0.54%	0.18%
Editorial	79,474	3.72%	1.25%
General News	6,48,920	30.39%	10.19%
Letters	2,716	0.13%	0.04%
Obituary	15,252	0.71%	0.24%
Political	9,82,615	46.01%	15.42%
Social	20,774	0.97%	0.33%
Sports News	3,73,863	17.51%	5.87%
Weather	482	0.02%	0.01%

Table 11-8 Mass Media Category Representation

11.5.6 Science and Technology

The Science and Technology Category of LDC-IL Malayalam text corpus covers 25 subdomains. The details of the representation of subdomains are given below:

Subdomain	Word Count	% (within Subdomain)	Overall Percentage
Architecture	11,012	2.38%	0.17%
Astrology	5,808	1.26%	0.09%
Astronomy	8,294	1.80%	0.13%
Ayurveda	54,786	11.86%	0.86%
Biotechnology	552	0.12%	0.01%
Botany	15,544	3.37%	0.24%
Chemistry	3,762	0.81%	0.06%
Computer Sciences	20,978	4.54%	0.33%
Criminology	7,678	1.66%	0.12%
Educational Psychology	376	0.08%	0.01%
Engineering-Civil	2,125	0.46%	0.03%
Environmental Science	18,675	4.04%	0.29%
Film Technology	66,571	14.42%	1.04%
Forestry	3,485	0.75%	0.05%
Mathematics	4,597	1.00%	0.07%
Medicine	37,804	8.19%	0.59%
Micro Biology	1,942	0.42%	0.03%
Naturopathy	18,868	4.09%	0.30%
Oceanology	3,735	0.81%	0.06%
Physics	21,824	4.73%	0.34%
Psychology	14,410	3.12%	0.23%
Sexology	5,075	1.10%	0.08%
Veterinary	11,207	2.43%	0.18%
Yoga	10,630	2.30%	0.17%

Zoology	49,021	10.62%	0.77%

Table 11-9: Science and Technology Category Representation

11.6 COPYRIGHT CONSENTS

The Malayalam text corpus has been collected from various sources and the copyright for the same stays with different sources. However, for the purposes of this corpus, consents have been sought from all the stakeholders. Most of the copyrights (around 83%) belong to private parties with only 17% belonging to the government agencies, either state or the central.

12 MANIPURI RAW TEXT CORPUS

Amom Nandaraj Meetei, Yumnam Premila Chanu, Rajesha N, Manasa G, Narayan Choudhary, L. Ramamoorthy

12.1 INTRODUCTION

Manipuri (also locally known Meeteilon by the native speakers) is the lingua franca amongst the various communities in the state of Manipur. It is the official language in government offices. Manipuri is mainly spoken in the northeast Indian states of Assam and Tripura, and also in Bangladesh and Burma (now Myanmar). As the lingua franca of the state, the language is spoken and understood by almost all the speakers of other communities in the state. In fact, it has occupied a prominent place in inter-tribal communications and is understood by several additional speakers as the second language in the bordering countries like Bangladesh and Burma and, states like Assam and Tripura. Because of its cosmopolitan vocabulary and absence of inflections, Manipuri is easily accessible to the different tribal people in eastern India, especially in the states of Assam, Tripura, Meghalaya, Nagaland, etc.

Since the 29th August 1992 Manipuri has been recognized as the first Tibeto-Burman language amongst the languages included in the Eighth Schedule of the constitution of India. Manipuri is the medium of instruction in all levels of education while English is another language, which is used in the courts, offices, etc. and also as the medium of instruction. The Manipur University has recognized Manipuri as an Honours subject in B.A. Standard. The Manipur University, more and above, has Manipuri Department under which the courses in M.A., M.Phil and Ph.D. degree can be studied. Manipuri had been the state language of Manipur ever since the time immemorable. During the British regime (1891-1947) its state language status continued. Deliberations of the Darbar were held; judgements of law courts were delivered in this language. Its state language status finds special recognition in the relevant provision of the Manipur State Constitution Act, 1947. In short, it is an official, court, administrative, lingua-franca, and chief language of Manipur State.

Manipuri has its own indigenous scripts called MEETEI-MAYEK, utilizing consonant letters, non-syllabic (not fully released) final consonant letters, independent vowels, secondary vowel signs, punctuation, numerals and ancient sign. LDC-IL Manipuri text corpus in Bengali script of contemporary usage has been collected. The Bengali script was adopted during the reign of King Pamheiba (1709-1748), the ruler of the Manipur Kingdom in the 18th century. The earliest use of Meetei Mayek is generally dated between 11th and 12th centuries. To cite the case in point, a stone inscription found at Khoibu in Tengnoupal district contains royal edicts of Kiyamba- this was the beginning of Cheitharol Kumbaba, the Royal Chronicle of Manipur. Pamheiba, the ruler of the Manipur Kingdom, introduced Hinduism, banned the use of the Meetei Script and adopted the Bengali Script. At present, in schools and colleges the Bengali script is gradually being replaced by the Meetei Script. The Government of Manipur has assured that Bengali Script

would be completely replaced in a phase-wise manner. LDC-IL Manipuri text corpus can be available in both Bengali Script and Meetei Mayek Script.

Manipuri text corpus is collected from various sources such as Manipur University Library, Manipur State Assembly Library, individual collections, etc. in Manipur. The majority of the text has been collected from the Manipur University library. LDC-IL tried to cover the entire categories in its standard list. Some categories like novel, short stories, cinema that fall under aesthetics, business (commerce), general news, political, administration and police documents (official documents) have huge amount of books and materials while some categories like astrology, computer science, physics, chemistry, zoology have very less amount of books. The overview of the represented domains (see 5 below) shows that literary texts are easily available in Manipuri but getting scientific text is found difficult. Some categories like educational psychology, biotechnology, music & musical instruments, and weather text are too rare in Manipuri.

12.2 PECULIARITIES OF MANIPURI TEXT

The Corpus of Manipuri 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 that are narrative and it contains elements of fiction. Novels, short stories, plays are examples of literary text. Non-literary texts are texts 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.

Average word length of Manipuri text is at par in comparison to other morphologically agglutinative languages included in the scheduled languages of India. Linguistically, Manipuri is agglutinative by morphology, non-inflectional by syntax and missing subjects can be traced back from the previous discourse as found in most of the text. Many Aryan and Iranian words are found in the Manipuri vocabulary. Nevertheless, such loan words started vanishing and many of these elements have being substituted by the newly coined terms/words as found in some contemporary texts. One needs to have much larger text corpora for the good coverage of words.

12.2.1 The Scripts used in writing the language

Manipuri is the only scheduled Tibeto-Burman language of India. It is also one of the Tibeto-Burman languages, which has its own script. However, as mentioned earlier, Manipuri has been written with the Manipuri alphabet, i.e., Meetei Mayek, or with the Bengali alphabets. Manipuri was, in full swing, written with the Bengali alphabet between 1709 and the middle of the 20th century. The alphabets used in the teaching of Manipuri are basically Assamese-Bengali Script. The Assamese/Bengali scripts found in the School textbooks along with Meetei-Mayek scripts are given in section 4.

The children have been taught as having 41 consonants and 14 vowels according to the textbook. However such use of the many symbols led to different writings/spellings of the same word having the same sound as shown below:

সন/শন/ ষন/ ষণ/সন্/শন্/ষণ্	pronounced as /sən/ ‘cow’
সাতনি/ছাটনি/ছাটণি/শাতনি/সাতীন/শাতীন	pronounced as /satin/ ‘umbrella’
সাতন্/ছাটন্/ছাটণ্/শাতন্/সাতীন্/শাতীন্	
গারি/গাড়ি/গাড়ী	pronounced as /gari/ ‘cart/van’
রতু/রীতু/ঋতু	pronounced as /ritu/ ‘season’

In Manipuri, there is no use for several Bengali letters, some of which its speakers are unable to pronounce correctly. However, such problems of inconsistency in spellings can be solved if a native speaker happens to make use of his/her own Meetei Mayek script because it doesn’t have any other redundant alphabet to represent the same sound. All the sounds relating to the alveolar fricative /s/ of the above examples can be integrated into a single letter ꯁ in Manipuri. All the non-syllabic final consonant letters such as ন্ and ণ্ in the above example will be represented as ꯨ, which is phonologically an unreleased stop phoneme. Hence the varied spellings of the same word ‘*cow*’ will be have the single graphical representation as ꯁꯨ /sən/. In a similar way, we will have the following representational forms for the items shown above.

সন/শন/ ষন/ ষণ/সন্/শন্/ষণ্	ꯁꯨ /sən/ ‘cow’
সাতনি/ছাটনি/ছাটণি/শাতনি/সাতীন/শাতীন	ꯁ꯭ꯇꯤꯅꯨ /satin/ ‘umbrella’
সাতন্/ছাটন্/ছাটণ্/শাতন্/সাতীন্/শাতীন্	
গারি/গাড়ি/গাড়ী	ꯒꯥ꯭ꯔꯤ/gari/ ‘cart/van’
রতু/রীতু/ঋতু	ꯔꯤꯇꯨ/ritu/ ‘season’

The LDC-IL transliteration tool has carried out maximum transliterating work in mapping the Bengali scripts to Meetei Mayek scripts by incorporating certain mapping rules and algorithm as briefly explained in the above examples.

12.3 DATA SAMPLING NOTES

12.3.1 Principles of Data Sampling

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

12.3.2 Field Works Undertaken

Manipuri text corpus is collected from various sources such as Manipur University Library, Manipur State Assembly Library, individual collections, etc. in Manipur. The text materials were

collected by conducting two field works undertaken in the period from 2010 to 2012. The greater part of the text has been collected from the Manipur University library, Canchipur, and Manipur Legislative Assembly Library, Imphal and CIIL Library, NERLC Library, Guwahati, Centre for Information on Language Sciences, Central Institute of Indian Languages, Mysore. LDC-IL tried to cover the entire categories in its standard list. Some corpus was collected from some homes or libraries.

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

- CIIL Library, Central Institute of Indian Languages, Mysore.
- NERLC Library, Guwahati
- Manipur University Library, Canchipur.
- Manipur Legislative Assembly Library, Imphal
- Saratchand Thiyam and Bimabati Thiyam Ongbi Home Library, Sagolband Lukram Leirak, Imphal.
- Sahitya Thoupanglup Library, Imphal.

Collected text materials have been published at various places within Manipur and other states of India such as Tripura, Assam, Delhi as well as other countries such as Bangladesh, etc.

An attempt has been made to cover the entire category in its standard list. As mentioned earlier, some categories like novel, short stories, general news, political, administration and police documents have huge amount of books and materials while some categories like astrology, computer science, physics, chemistry, zoology have very less amount of books. This shows that literary texts are easily available in Manipuri but getting scientific text is found difficult. On the other hand, some categories like educational psychology, biotechnology, music & musical instruments, and weather text are too rare in Manipuri.

It is not an easy task that text data are collected from the field. In general, most of the libraries do not allow us to take huge amount of text from their shelves at a time because it is against their rules and principles. It took certain formalities for taking permission to take Xerox of more than ten books since the issuing of maximum three to four books is the only standard limit of the library concerned. Even if the librarian happened to allow us taking many books at a time, the photocopy kiosk had issues thereupon as there was a long queue.

Some time photocopy attendant refused to photocopy randomly selected pages because of the long queue waiting and it takes up more time for them to turn the pages compared to the continuous page photocopying they are accustomed to. It was another issue, too, that the field worker/linguist had to carry a huge list of photocopy bundles with them which was many a time cumbersome to travel with.

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

12.3.3 Data Inputting

All the texts have been typed in Unicode using the InScript Keyboard directly onto the XML files. The data has been inputted by Ms. Oinam Bina Devi, Khundrakpam Ibema Devi, and L. Urmila, all native speakers of Manipuri.

12.3.4 Validation and Normalization Workshops

Linguistic Data Consortium for Indian Languages (LDC-IL) conducted workshops for data validation and normalization. The experts unanimously suggested that the Manipuri text corpus should remain true to the text.

12.3.5 Proofreading

Manipuri text data has been proofread by both internal resource persons and the resource persons engaged in the programmes for Corpus Normalization and the Short-Term Goal Oriented Projects (Text Corpus Cleaning Workshops). The following account shows the workshop facet:

6. Manipuri Corpus Normalization, 2010
7. Short Term Goal Oriented Project- Manipuri Language Text Cleaning Workshop: 7th Oct. – 11th Dec. 2014
8. Short Term Goal Oriented Project- Manipuri Language Text Cleaning Workshop: 3rd Aug.. – 11th Sept. 2015

It was so decided and followed across the languages that text manipulation be avoided thoroughly and only the typo errors committed during the input process have been corrected with reference to the hard copies.

As a native speaker of the language concerned, the cleaners become oriented towards the elimination and modification of idiosyncratic spelling and punctuations. There still invites a thought on the term "Standard" when the cleaners are asked to do the work of printed languages which are normally subjected to examination and correction. To instantiate the idea, in the category like Drama (in dialogue form) or narratives, there contain full of pseudo-spoken, rather than edited written language, and many a 'house style' eye dialects. The cleaners always find themselves intended to change these 'orthodox spellings' into the normal natural forms. Verses found in prose have been excluded or eliminated since it syntax and its lexicon are so different from those of prose, escaping from special problems they would present. In addition to the issues mentioned above, there are other problems the cleaners find when they are to check the details of the sampled files with respect to the years of publication (whether there is any positive evidence on which it was written more than one or two year(s) before it was printed), editions (1st, 2nd, 3rd so on.), translation (sometimes author's name missing or translator's), and publisher's name etc. To add some more, there are a number of few quotations from archaic or old Manipuri language which introduce older syntax (verbs in initial or even in medial positions and in the root form also in SOV language), lexicon which is no longer used in the present day, old writing style of

graphics etc. In order to avoid such complexities, we centre on the contemporary texts and collected so.

However, there must be certain limitation that the cleaners have to follow in the perspectives of 'trust the text' (Sinclair, 2004). Since the sources have been found in the printed form, they should be printed as such. Hence the spelling, punctuation, capitalization, word division or boundary, the use of boldface and italics have to be coded for the computer as found. The cleaner should not correct even the typographical errors seeing the editorial alteration of the text. In short, all errors and inconsistencies which basically stem from the original sources are allowed to stand independently. This task leads to the understanding level of expression by different writers regardless of such unwanted phenomena. The following resource persons attended in the above-mentioned projects did major works in retaining the corpus intact. They are Prof. Ch. Yashawanta Singh, Prof. N. Aruna Devi, Dr. T. Tampha Devi, Amom Nandaraj Meetei, Yumnam Premila Chanu, Longjam Anand Singh, Dr. M. Bidyaranani Devi, Dr. N. Brojen Singh, Nameirakpam Amit Singh, Takhellambam Debachand, Irengbam Spark, Oinam Nanao Devi, Taurangbam Anuradha Devi, Oinam Basanta Singh, Takhellambam Bijaya Devi, Naorem Brindebala Devi, Thingom Tarunkumar Singh, Moirangthem Rajesh Singh, Soibam Langlen Chanu, Tongbram Narmada Devi respectively.

The text materials collected are of contemporary kind as truly as possible that it is limited to materials published in the calendar year 1960 to 2012 as per the METADATA information of the LDC-IL Manipuri Corpora Monolingual Written Texts. Since the maximum categories as appeared in the standard list were made cover and there is no quotations from older or archaic Manipuri in terms of older syntax and lexicon and also that the corpus size in relation the whole list is significant, it is now safe to say that this Manipuri Raw Text Corpus is truly representative of the contemporary Manipuri.

12.4 TRANSLITERATIONS IN LDC-IL MANIPURI TEXT CORPUS

For easy reference and uniformity of metadata, some entries in the metadata file, namely *'Title', 'Headline', 'Author', 'Editor', 'Translator'* are transliterated from Bengali and Meetei-Mayek to Roman letters. Numeric characters were transliterated from Bengali and Meetei-Mayek to Hindu-Arabic system.

The LDC-IL transliteration scheme of Manipuri (both in Bengali and Meetei-Mayek scripts) to Roman is given below.

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

Scripts	**Vowels and Vowel Signs in Bengali and Meetei Mayek Scripts**												
Bengali	অ	আ	ই	ঈ	উ	ঊ	ঋ	এ	ঐ	ও	ঔ		
Meetei Mayek	ꯑ		ꯏ		ꯎ								
Bengali	া		ি	ী	ু	ূ	ৃ	ে	ৈ	োা	ৌ	ং	ঃ
Meetei Mayek		ꯔ	ꯤ		ꯨ		ꯩ	ꯣ	ꯥ	ꯣꯔ	ꯦ	ꯟ	
Roman	a	A	i	I	U	U	x	E	ai	O	au	M	H

Consonants						
Bengali	ক	খ	গ	ঘ	ঙ	ক
Meetei Mayek	ꯀ	ꯈ	ꯒ	ꯘ	ꯉ	ꯀ
Roman	ka	kha	ga	gha	ng'a	ka
Bengali	চ	ছ	জ	ঝ	ঞ	চ
Meetei Mayek	ꯆ		ꯖ			ꯆ
Roman	ca	cha	Ja	jha	nj'a	ca
Bengali	ট	ঠ	ড	ঢ	ণ	ট
Roman	Ta	Tha	Da	Dha	Na	Ta
Bengali	ত	থ	দ	ধ	ন	ত
Meetei Mayek	ꯇ	ꯊ	ꯗ	ꯙ	ꯅ	ꯇ
Roman	ta	tha	Da	dha	na	ta
Bengali	প	ফ	ব	ভ	ম	প
Meetei Mayek	ꯄ	ꯐ	ꯚ	ꯜ	ꯃ	ꯄ
Roman	pa	pha	Ba	bha	ma	pa
Bengali	য	র	ল	শ	স	য	হ	ড়	ঢ়	য়	ৎ
Meetei Mayek		ꯔ	ꯂ		ꯁ		ꯍ			ꯝ	
Roman	ya	ra	la	sha	Sa	sa	ha	D'a	Dh'a	Ya	t

Unreleased Consonants		
ক্		ঙ্
ꯀ꯭		ꯉ꯭
k		ng'
ম্		প্
ꯃ꯭		ꯄ꯭
m		p
ন্	ত্	ল্
ꯅ꯭	ꯇ꯭	ꯂ꯭
n	t	l

Numerals (Bengali and Meetei Mayek to Hindu-Arabic)										
Bengali	০	১	২	৩	৪	৫	৬	৭	৮	৯
Meetei Mayek	꯰	꯱	꯲	꯳	꯴	꯵	꯶	꯷	꯸	꯹
Hindu-Arabic	0	1	2	3	4	5	6	7	8	9

12.5 OVERVIEW OF REPRESENTED DOMAINS

LDC-IL Manipuri Text Corpus size is 6148220 words and 43127842 characters drawn from 1202 different titles, including the extracts from newspapers.
The data is of typed+cleaned one.

The representation of the six major domains covered has been shown in the table below:

Domain	Domain Word Count	Percentage
Aesthetics	3772994	61.40%
Commerce	18450	0.30%
Mass Media	775261	12.62%
Official Document	442950	7.21%
Science and Technology	304545	4.96%
Social Sciences	831078	13.52%
Total	6145278	**100**

Table 12-1 Representation of the Domains in Manipuri Text Corpus

As each domain has several sub-domains, the following tables show the representations of each subdomain where the number of subcategories that fall under the same domain along with their total word count, percentage within the subdomain as well as the overall percentage is provided.

12.5.1 Aesthetics

The Aesthetics category of Manipuri text corpus covers 28 sub-categories bearing a total of 3,77,29,94 words along with the overall percentage of 61.40%. The representational details are given in the table below.

#	Subcategory	Word count	Percentage within Sub domain	Overall Percentage
01	Autobiography	45778	1.21%	0.74%
02	Biography	198381	5.26%	3.23%
03	Cinema	17473	0.46%	0.28%
04	Culture	321048	8.51%	5.22%
05	Fine Arts-Dance	14850	0.39%	0.24%
06	Fine Arts-Drawing	4658	0.12%	0.08%
07	Fine Arts-Music	33339	0.88%	0.54%
08	Fine Arts-Musical Instruments	22414	0.59%	0.36%
09	Fine Arts-Sculpture	3288	0.09%	0.05%
10	Folk Tales	37585	1.00%	0.61%
11	Folklore	63410	1.68%	1.03%
12	Handicrafts	2293	0.06%	0.04%
13	Literary Texts	9186	0.24%	0.15%
14	Literature-Children's Literature	2502	0.07%	0.04%
15	Literature-Criticism	809311	21.45%	13.17%
16	Literature-Diaries	36784	0.97%	0.60%
17	Literature-Epics	1398	0.04%	0.02%
18	Literature-Essays	319538	8.47%	5.20%
19	Literature-Letters	1559	0.04%	0.03%
20	Literature-Novels	354659	9.40%	5.77%
21	Literature-Plays	319728	8.47%	5.20%
22	Literature-Poetry	7668	0.20%	0.12%
23	Literature-Science Fiction	1061	0.03%	0.02%
24	Literature-Short Stories	852839	22.60%	13.88%
25	Literature-Speeches	58061	1.54%	0.94%
26	Literature-Text Books (School)	38644	1.02%	0.63%
27	Literature-Travelogues	173173	4.59%	2.82%
28	Mythology	22366	0.59%	0.36%
	Total	3772994	100%	61.40%

Table 12-2 Aesthetics Category Representation

12.5.2 commerce

The commerce category of Manipuri text corpus covers 7 sub-categories bearing a total of 1,84,50 words along with the overall percentage of 0.30%. The representational details are given in the table below.

#	Subcategory	Word count	Percentage within Sub domain	Overall Percentage
01	Banking	1468	7.96%	0.02%
02	Business	7259	39.34%	0.12%
03	Career and Employment	3224	17.47%	0.05%
04	Industry	3132	16.98%	0.05%
05	Management	2104	11.40%	0.03%
06	Share Market	682	3.70%	0.01%
07	Tourism	581	3.15%	0.01%
	Total	18450	100%	0.30%

Table 12-3 Commerce Category Representation

12.5.3 Mass Media

The Mass Media category of Manipuri text corpus covers 13 sub-categories bearing a total of 77,52,61 words along with the overall percentage of 12.62%. The representational details are given in the table below.

#	Subcategory	Word count	Percentage within Sub domain	Overall Percentage
01	Article	721	0.09%	0.01%
02	Classifieds	4020	0.52%	0.07%
03	Discussions	56527	7.29%	0.92%
04	Editorial	284062	36.64%	4.62%
05	General News	31352	4.04%	0.51%
06	Interviews	6016	0.78%	0.10%
07	Letters	92855	11.98%	1.51%
08	Obituary	584	0.08%	0.01%
09	Political	12456	1.61%	0.20%
10	SMS	4108	0.53%	0.07%
11	Social	47331	6.11%	0.77%
12	Speeches	666	0.09%	0.01%
13	Sports News	234563	30.26%	3.82%
	Total	775261	100%	12.62%

Table 12-4 Mass Media Category Representation

12.5.4 Official Document

The Official Document category of Manipuri text corpus covers 3 sub-categories bearing a total of 44,29,50 words along with the overall percentage of 7.21%. The representational details are given in the table below.

#	Subcategory	Word count	Percentage within Sub domain	Overall Percentage
01	Legislature	958	0.22%	0.02%
02	Parliamentary/Assembly Debates	441648	99.71%	7.19%
03	Police Documents	344	0.08%	0.01%
	Total	442950	100%	7.21%

Table 12-5 Official Document Category Representation

12.5.5 Science and Technology

The Science and Technology category of Manipuri text corpus covers 30 sub-categories bearing a total of 30,45,45 words along with the overall percentage of 4.96%. The representational details are given in the table below.

#	Subcategory	Word count	Percentage within Sub domain	Overall Percentage
01	Agriculture	29845	9.80%	0.49%
02	Architecture	1377	0.45%	0.02%
03	Astrology	38062	12.50%	0.62%
04	Astronomy	18959	6.23%	0.31%
05	Ayurveda	964	0.32%	0.02%
06	Bio Chemistry	2848	0.94%	0.05%
07	Biology	1368	0.45%	0.02%
08	Biotechnology	534	0.18%	0.01%
09	Botany	10292	3.38%	0.17%
10	Chemistry	14470	4.75%	0.24%
11	Computer Sciences	5308	1.74%	0.09%
12	Criminology	14586	4.79%	0.24%
13	Educational Psychology	927	0.30%	0.02%
14	Engineering-Mechanical	490	0.16%	0.01%
15	Engineering-Others	598	0.20%	0.01%
16	Environmental Science	13550	4.45%	0.22%
17	Film Technology	58139	19.09%	0.95%
18	Forestry	2764	0.91%	0.04%
19	Geology	1758	0.58%	0.03%
20	Horticulture	755	0.25%	0.01%
21	Medicine	28144	9.24%	0.46%

22	Naturopathy	16352	5.37%	0.27%
23	Oceanology	452	0.15%	0.01%
24	Physics	7530	2.47%	0.12%
25	Psychology	6009	1.97%	0.10%
26	Sexology	2542	0.83%	0.04%
27	Statistics	622	0.20%	0.01%
28	Text Book (Science)	1236	0.41%	0.02%
29	Veterinary	12567	4.13%	0.20%
30	Zoology	11497	3.78%	0.19%
	Total	304545	100%	4.96%

Table 12-6 Science and Technology Representation

12.5.6 Social Sciences

The Social Sciences category of Manipuri text corpus covers 25 sub-categories bearing a total of 82,43,99 words along with the overall percentage of 13.52%. The representational details are given in the table below.

#	Subcategory	Word count	Percentage within Sub domain	Overall Percentage
01	Anthropology	961	0.12%	0.02%
02	Archeology	6291	0.76%	0.10%
03	Demography	1193	0.14%	0.02%
04	Economics	20338	2.45%	0.33%
05	Education	44988	5.41%	0.73%
06	Epigraphy	1714	0.21%	0.03%
07	Fisheries	605	0.07%	0.01%
08	Food and Wellness	18491	2.22%	0.30%
09	Geography	2414	0.29%	0.04%
10	Health and Family Welfare	63170	7.60%	1.03%
11	History	127482	15.34%	2.07%
12	Home Science	14565	1.75%	0.24%
13	Journalism	37158	4.47%	0.60%
14	Law	56350	6.78%	0.92%
15	Library Science	3879	0.47%	0.06%
16	Linguistics	19592	2.36%	0.32%
17	Personality Development	1380	0.17%	0.02%
18	Philosophy	24407	2.94%	0.40%
19	Physical Education	11033	1.33%	0.18%
20	Political Science	89238	10.74%	1.45%
21	Public Administration	27543	3.31%	0.45%
22	Religion/Spiritual	125619	15.12%	2.04%

23	Sociology	43065	5.18%	0.70%
24	Sports	87554	10.53%	1.42%
25	Text Book (Social Science)	2048	0.25%	0.03%
	Total	831078	100%	13.52%

Table 12-7 Social Sciences Representation

12.6 COPYRIGHT CONSENTS

The Manipuri text corpus has been collected from various sources and the copyright for the same stays with different sources. However, for the purposes of this corpus, consent has been sought from all the stakeholders. Most of the copyrights (around 92%) belong to private parties with only 8% belonging to the government agencies, either state or the central.

12.7 REFERENCES

Amom, M. N. 2010. Corpus, Stylistics and Cognitive Aspects of Language Use. Paper presented in *International Conference on language Development and Computing Methods*. Karunya University, Coimbatore.

Francis, W. N. 1980. A Tagged Corpus: Problems and Prospects. In *Studies in English Linguistics for Randolph Quirk*, edited by Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. Longman: London and New York.

Manihar, Ch. 1996. *A History of Manipuri Literature*. Sahitya Akademi, Delhi.

Sinclair, J. 2004. *Trust the text*. London: Routledge.

13 NEPALI RAW TEXT CORPUS

Umesh Chamling, Rupesh Rai, Rajesha N., Manasa G., Dr. Narayan Choudhary, Dr. L. Ramamoorthy

13.1 INTRODUCTION

Nepali is one of the official language of West Bengal and Sikkim state. It is one of the 22 schedual languages of India. It is spoken in most of North-Eastern states of India and also other states, similarly Delhi, Uttranchal, Uttar Pradesh, Bihar, Jharkhand etc. Nepali is also an official language of Nepal. About a quarter of the population in Bhutan speaks Nepali. Nepali is written in Devanagari Script. It is written from left to right direction. It also called Nagari. Nagari script has roots in the ancient Brāhmī script family, It has long been used traditionally by religiously educated people in South Asia. The Devanagari script is used for over 120 languages, and those are Nepali, Hindi, Marathi, Bhojpuri, Maithili etc. It is closely related to the Nandinagari script commonly found in numerous ancient manuscripts of South India. The script is also used to write several minority languages of Nepali community such as Gurung, Magar, Bhujel, Thami etc.

Nepali text corpus is collected from various libraries of Darjeeling, Sikkim, Assam, Uttaranchal. Mostly from Kurseong, Mirik, Kalimpong, Silgadhi, Gangtok, Guwahati, Almora and Mussoorie. The greater part of the text has been taken from Desbandhu District Library Darjeeling, Sonada Library, Sarbajanik Sammelan Rural Library Mirik, Sub Divisional Library, Kalimpong, NERLC(North-East Regional Language Centre, Guwahati) Library. LDC-IL tried to cover the entire category in its standard list. Some categories like novel, short stories have huge amount of books but some categories like Physics, Chemistry, Economics, Agriculture, Photography have very less amount of books. Literary texts are easily available in Nepali but getting scientific, Technical text is very difficult. Some categories like Epigraphy, Finance, Oceanology text are too rare in Nepali.

13.2 PECULIARITIES OF NEPALI TEXT

The Corpus of Nepali 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 that are narrative and it contains elements of fiction. Novels, short stories, plays are examples of literary text. Non-literary texts are texts 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.

13.2.1 Orthographic variation and eyelash ‘ra’ in Nepali

A glyph has no intrinsic meaning, it conveys distinctions in form. Time to time the user or developers made small variation in devanagari script and same changes come into Nepali. These were in अ, झ, ण, श. It was not unique feature of Nepali, but it made small changes in use of nepali orthography system. We faced problem while inputing data from many text.

Besides that, Nepali has its typical orthography, which is called ‘Shaja’ [A publisher from Lalitpur Kathmandu, Nepal] /ʤ/ ‘jha’, no other Devanagari script users having the same.

Nepali has *eyelash* /r/ 'ra'. This is 'ra' with *halanta* (्), or half 'ra' (‾). It has its single Unicode value. There are more than three ways to type eyelash 'ra'.

13.2.2 Transliterations in LDC-IL Nepali text corpus

For easy reference and uniformity of metadata, some entries in the metadata file, namely *'Title', 'Headline', 'Author', 'Editor', 'Translator'* are transliterated from Nepali to Roman letters.

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

LDC-IL Transliteration Schema
Nepali characters to Roman and Nepali

Vowels and Vowel Signs												
अ	आ	इ	ई	उ	ऊ	ऋ	ए	ऐ	ओ	औ	अं	अः
	ा	ि	ी	ु	ू	ृ	े	ै	ो	ौ	ं	ः
a	A	i	I	u	U	x	e	ai	o	au	M	H

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

य	र		ल	व	श	ष	स	ह	ड़	ढ़	ऱ
ya	Ra		la	va	sha	Sa	sa	ha	La	Za	TTTa

Eyelash ra
‾

Numerals (Malayalam to Hindu-Arabic)									
൦	൧	൨	൩	൪	൫	൬	൭	൮	൯
0	1	2	3	4	5	6	7	8	9

13.3 DATA SAMPLING NOTES

13.3.1 Principles of Data Sampling

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

13.3.2 Field Works Undertaken

Nepali text corpus is collected from various libraries of Darjeeling, Sikkim, Assam and Uttaranchal. The text materials were collected by conducting five field works undertaken in the period from 2009 to 2012. The greater part of the text has been taken from Khappandas Memorial Library, Soureni Busty Mirik, Sub Divisional Library Kalimpong, North Bengal University Library Darjeeling, and various public library.

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

- Mirik Sarbajanik Sammelan Rural Library, Mirik
- Garidhura Public Library, Kurseong
- Sub Divisional Library, Kalimpong
- Nava Yowak Sangha Rural Library, Rungbull
- Gorkha Jana Pustakalaya, Kurseong
- Khappandas Memorial Library, Soureni Busty, Mirik
- Pankhabari Public Library, Pankhabari
- North Bengal University Library, Darjeeling
- Kurseong College Library, Kurseong
- Desbandhu District Library, Darjeeling
- Devkota Sangh Pustakalaya, Silgadhi
- NERLC Library, Assam
- Central Institute of Indian Language Library, Mysore
- Personal Collections.

Collected text materials have been published at various places within Darjeeling and other states of India such as Sikkim, Assam, Manipur, Meghalaya, Arunachal Pradesh, Nagaland, Uttra Pradesh, Uttranchal, Himachal Pradesh, Delhi, Bihar, Andra Pradesh, Karnataka as well as other countries such as Nepal, Russia, Denmark etc.

An attempt has been made to cover the entire category in its standard list. Some categories like novel, short stories have huge amount of books but some categories like physics, chemistry, economics have very less amount of books. Literary texts are easily available in Nepali but getting scientific text is very difficult. Some categories like epigraphy, finance, oceanology text are too rare in Nepali.

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

Some time Xerox attendents refused to photocopy randomly selected pages because of the long queue waiting and it takes up more time for them to turn the pages compared to continuous page photocopying they are accustomed to. It was another issue that the field worker/linguist had to carry a huge list of photocopy bundles with them which was many a times cumbersome to travel with.

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

13.3.3 Data Inputting

All the text has been typed in Unicode using the InScript Keyboard directly onto the XML files. The data has been inputted by Ms. Srilakshmi M P, Sithalakshmi M L, Vanamala B H, Rajeshwari R, Vidhyashree M, Padmashree H R, Radhika M, Mamatha, all native speakers of Kannada and Tamil but familiar enough with the scripts of Devanagari.

13.3.4 Proofreading

Nepali text data has been proofread by internal and external resource persons. We conducted corpus normalization workshop with external resource persons on 4thJune to 15th July 2010, 3rd January to 28th February 2013, 5th August to 4th October 2013, 10th November to 7th January 2015. The text has always been kept true to the printed material and typos, if any, occurring at the time of typing have only been corrected.

The printed materials collected for the corpus are contemporary , mainly published after 1990.

13.3.5 Data Extracted from Web Sites

Nepali News cropus data is extracted from News websites of *"Nepalsamachar Patra" (https://www. http://pknewspapers.com/) , " Gorkhapatra" (www. http://gorkhapatraonline.com/*). The news content was categorized based on the content of the text and archived. The period of selection of the news corpus ranges from 30,Jan 2009 to 11 Sep 2009.

13.4 Overview of Represented Domains

LDC-IL Nepali Text Corpus size is: 70,57,524 Words with character count at 46879154 drawn from 1,347 different titles, including the extracts from newspapers. The data can be categorized into two classes of typed+cleaned and crawled. The crawled data has been crawled mainly from news websites and archived using the standard processing of LDC-IL text corpus preparation.

The following table gives a summary of the typed and crawled text of the Nepali Raw Text Corpus.

Text Type	Word Count	KeyStroke/Character Count
Typed+Cleaned	6787918	45104255

Crawled	269606	1774899
Total	7057524	46879154

Table 13-1 Representation of the Domains in Nepali Text Corpus

The representation of the six major domains covered has been shown in the table below:

Domain	Word Count	Percentage
Aesthetics	4072977	57.71%
Commerce	30354	0.43%
Mass Media	2271064	32.18%
Official Documents	2426	0.03%
Science & Technology	80306	1.14%
Social Sciences	600397	8.51%
Total	70,57,524	100.00%

Table 13-2: Representation of the Domains in Nepali Text Corpus

As each domain has several sub-domains, the following table shows the representation of the several domains, both within the domain and across all the domains.

Domain	Subdomain	Word Count	% (within Subdomain)	Overall Percentage
Aesthetics	Autobiographies	24754	0.61%	0.35%
Aesthetics	Biographies	307829	7.56%	4.36%
Aesthetics	Cinema	3258	0.08%	0.05%
Aesthetics	Culture	96596	2.37%	1.37%
Aesthetics	Fine Arts-Dance	11002	0.27%	0.16%
Aesthetics	Fine Arts-Drawing	740	0.02%	0.01%
Aesthetics	Fine Arts-Music	10070	0.25%	0.14%
Aesthetics	Fine Arts-Musical Instruments	6620	0.16%	0.09%
Aesthetics	Fine Arts-Sculpture	10525	0.26%	0.15%
Aesthetics	Folk Tales	621	0.02%	0.01%
Aesthetics	Folklore	27720	0.68%	0.39%
Aesthetics	Humour	35026	0.86%	0.50%

Aesthetics	Literature-Children's Literature	10479	0.26%	0.15%
Aesthetics	Literature-Criticism	863007	21.19%	12.23%
Aesthetics	Literature-Diaries	307052	7.54%	4.35%
Aesthetics	Literature-Epics	200	0.00%	0.00%
Aesthetics	Literature-Essays	425981	10.46%	6.04%
Aesthetics	Literature-Letters	4835	0.12%	0.07%
Aesthetics	Literature-Novels	629468	15.45%	8.92%
Aesthetics	Literature-Plays	233675	5.74%	3.31%
Aesthetics	Literature-Science Fiction	7178	0.18%	0.10%
Aesthetics	Literature-Short Stories	788433	19.36%	11.17%
Aesthetics	Literature-Speeches	39681	0.97%	0.56%
Aesthetics	Literature-Text Books (School)	103956	2.55%	1.47%
Aesthetics	Literature-Travelogues	92892	2.28%	1.32%
Aesthetics	Mythology	27922	0.69%	0.40%
Aesthetics	Photography	3457	0.08%	0.05%
Commerce	Banking	9416	31.02%	0.13%
Commerce	Business	8391	27.64%	0.12%
Commerce	Finance	6957	22.92%	0.10%
Commerce	Industry	782	2.58%	0.01%
Commerce	Tourism	4808	15.84%	0.07%
Mass Media	Article	109118	4.80%	1.55%
Mass Media	Classifieds	454	0.02%	0.01%
Mass Media	Discussions	99652	4.39%	1.41%

Mass Media	Editorial	98814	4.35%	1.40%
Mass Media	General News	1144971	50.42%	16.22%
Mass Media	Interviews	36816	1.62%	0.52%
Mass Media	Letters	26657	1.17%	0.38%
Mass Media	Obituary	28248	1.24%	0.40%
Mass Media	Political	473831	20.86%	6.71%
Mass Media	Social	3834	0.17%	0.05%
Mass Media	Sports News	248669	10.95%	3.52%
Official Document	Police Documents	2426	100.00%	0.03%
Science and Technology	Agriculture	11910	14.83%	0.17%
Science and Technology	Architecture	282	0.35%	0.00%
Science and Technology	Astronomy	239	0.30%	0.00%
Science and Technology	Ayurveda	4477	5.57%	0.06%
Science and Technology	Biology	3242	4.04%	0.05%
Science and Technology	Botany	3653	4.55%	0.05%
Science and Technology	Chemistry	550	0.68%	0.01%
Science and Technology	Computer Sciences	242	0.30%	0.00%
Science and Technology	Criminology	4118	5.13%	0.06%
Science and Technology	Engineering-Electrical	300	0.37%	0.00%

Science and Technology	Environmental Science	1722	2.14%	0.02%
Science and Technology	Forestry	3822	4.76%	0.05%
Science and Technology	Geology	5194	6.47%	0.07%
Science and Technology	Homeopathy	3396	4.23%	0.05%
Science and Technology	Horticulture	217	0.27%	0.00%
Science and Technology	Logic	520	0.65%	0.01%
Science and Technology	Medicine	9924	12.36%	0.14%
Science and Technology	Psychology	20786	25.88%	0.29%
Science and Technology	Text Book (Science)	4934	6.14%	0.07%
Science and Technology	Yoga	511	0.64%	0.01%
Science and Technology	Zoology	267	0.33%	0.00%
Social Sciences	Anthropology	13856	2.31%	0.20%
Social Sciences	Archeology	696	0.12%	0.01%
Social Sciences	Economics	7890	1.31%	0.11%
Social Sciences	Education	61967	10.32%	0.88%
Social Sciences	Fisheries	305	0.05%	0.00%
Social Sciences	Geography	1475	0.25%	0.02%
Social Sciences	Health and Family Welfare	16799	2.80%	0.24%

Social Sciences	History	228123	38.00%	3.23%
Social Sciences	Home Science	612	0.10%	0.01%
Social Sciences	Journalism	12733	2.12%	0.18%
Social Sciences	Law	7407	1.23%	0.10%
Social Sciences	Linguistics	49593	8.26%	0.70%
Social Sciences	Philosophy	12488	2.08%	0.18%
Social Sciences	Physical Education	549	0.09%	0.01%
Social Sciences	Political Science	68929	11.48%	0.98%
Social Sciences	Public Administration	467	0.08%	0.01%
Social Sciences	Religion/Spiritual	66272	11.04%	0.94%
Social Sciences	Sociology	31702	5.28%	0.45%
Social Sciences	Sports	12135	2.02%	0.17%
Social Sciences	Text Book (Social Science)	6399	1.07%	0.09%

Table 13-3: Representation of Subdomains in Nepali Text Corpus

13.5 COPYRIGHT CONSENTS

The Nepali text corpus have been collected from various sources and the copyright for the same stays with different sources. However, for the purposes of this corpus, consents have been sought from all the stakeholders. Most of the copyrights (around 90%) belong to private parties with only 10% belonging to the government agencies, either state or the central.

14 MARATHI RAW TEXT CORPUS

Saurabh Varik, Bhageshree Khandale, Rajesha N, Manasa G, Narayan Choudhary, L. Ramamoorthy

14.1 INTRODUCTION

Marathi is an Indo-Aryan language spoken predominantly by around 83 million Marathi people of Maharashtra, India. It is the official language and co-official language in the Maharashtra and Goa states of Western India, respectively, and is one of the 22 scheduled languages of India. There were 83 million speakers in 2011; Marathi ranks 19th in the list of most spoken languages in the world. Marathi has the third largest number of native speakers in India, after Hindi and Bengali.Marathi has some of the oldest literature of all modern Indian languages, dating from about 900 AD. The major dialects of Marathi are Standard Marathi and the Varhadi dialect. Koli, Malvani Konkani has been heavily influenced by Marathi varieties.

Marathi distinguishes inclusive and exclusive forms of 'we' and possesses a three-way gender system that features the neuter in addition to the masculine and the feminine. In its phonology it contrasts apico-alveolar with alveopalatal affricates and, in common with Gujarati, alveolar with retroflex laterals ([l] and [ɭ], Marathi letters ल and ळ respectively).

Marathi text corpus is collected from various libraries in Maharastra mostly from Pune University, Marthwada University and WRLC, Pune. The greater part of the text has been taken from WRLC library, Pune. LDC-IL tried to cover the entire category in its standard list. Some categories like novel, short stories have huge amount of books but some categories like physics, chemistry, economics have very less amount of books. Literary texts are easily available in Marathi but getting scientific text is very difficult. Some categories like epigraphy, finance, oceanology text are too rare in Marathi.

14.2 PECULIARITIES OF MARATHI TEXT

The Corpus of Marathi 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 that are narrative and it contains elements of fiction. Novels, short stories, plays are examples of literary text. Non-literary texts are texts 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.

14.3 DATA SAMPLING NOTES

14.3.1 Principles of Data Sampling

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

14.3.2 Field Works Undertaken

Marathi text corpus is collected from various libraries in Maharastra, mostly from Pune University, Marthwada University and WRLC, Pune. The text materials were collected by conducting four field works undertaken in the period from 2009 to 2010. The greater part of the text has been taken from WRLC library and Pune University Campus library.

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

1. Pune University Campus Library, Pune
2. Marathwada University Library, Aurangabad
3. WRLC, Pune

Collected text materials have been published at various places within Maharastra and other states of India such as Karnataka, Goa, Delhi.

An attempt has been made to cover the entire category in its standard list. Some categories like novel, short stories have huge amount of books but some categories like physics, chemistry, economics have very less amount of books. Literary texts are easily available in Marathi but getting scientific text is very difficult. Some categories like epigraphy, finance, oceanology text are too rare in Marathi.

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

Some time Xerox attendents refused to photocopy randomly selected pages because of the long queue waiting and it takes up more time for them to turn the pages compared to continuous page photocopying they are accustomed to. It was another issue that the field worker/linguist had to carry a huge list of photocopy bundles with them which was many a times cumbersome to travel with.

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

14.3.3 Data Inputting

All the text has been typed in Unicode using the InScript Keyboard directly onto the XML files. The data has been inputted by H. S. Rupa, J. Shobha, K. R. Veena, Mamtha, Radhika M., Rajeshwari R. a native speaker of Kannada.

14.3.4 Validation and Normalization Workshops

A 5-day workshop was conducted at Linguistic Data Consortium from August 23-27, 2010 with Dr. Geeta Manjrekar (Expert), Mr. Vinay Bapat (RP), Mr. Vinayak Durge (RP), Ms. Anita Kore (RP) and Ms. Sneha Bhambre (RP) from the different part of Maharastra as experts. The experts suggested that the Marathi text corpus should remain true to the text.

``

14.3.5 Proofreading

Marathi text data has been proofread by internal resource persons. The text has always been kept true to the printed material and typos, if any, occurring at the time of typing have only been corrected.

The printed materials collected for the corpus is contemporary , mainly published after 1990.

14.3.6 Data Extracted from Web Sites

Marathi News cropus data is extracted from News websites of Marathi Newsaper i.e. http://www.esakal.com, http://maharashtratimes.indiatimes.com . The news content was categorized based on the content of the text and archived. The period of selection of the news corpus ranges from 2009 to 2010.

14.3.7 Transliterations in LDC-IL Marathi text corpus

For easy reference and uniformity of metadata, some entries in the metadata file, namely *'Title', 'Headline', 'Author', 'Editor', 'Translator'* are transliterated from Marathi to Roman letters. Numeric characters were transliterated from Marathi to Hindu-Arabic system.

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

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

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

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**

य	र	ऱ	ल	व	श	ष	स	ह	ळ	ऴ
ya	**ra**	**Ra**	**La**	**va**	**sha**	**Sa**	**sa**	**ha**	**La**	**Za**

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

14.4 OVERVIEW OF REPRESENTED DOMAINS

LDC-IL Marathi Text Corpus size is: 2157078 words drawn from 678 different titles, including the extracts from newspapers. The data can be categorized into two classes of typed+cleaned and crawled. The crawled data has been crawled mainly from news websites and archived using the standard processing of LDC-IL text corpus preparation.

The following table gives a summary of the typed and crawled text of the Marathii Raw Text Corpus.

Text Type	Word Count	KeyStroke/Character Count
Typed+Cleaned	**1865045**	**12417692**
Crawled	**292033**	**2070169**
Total	**2157078**	**14487861**

Table 14-1: Representation of the typed and crawled Marathi Text Corpus

The representation of the five major domains covered has been shown in the table below:

Domain	Domain Word Count	Percentage
Aesthetics	**1515006**	**70.23%**
Commerce	**20795**	**0.96%**
Mass Media	**363122**	**16.83%**
Science and Technology	**55902**	**2.59%**
Social Sciences	**202253**	**9.38%**
Total	**2,157,078**	**100**

Table 14-2: Representation of the Domains in Marathi Text Corpus

As each domain has several sub-domains, the following table shows the representation of the several domains, both within the domain and across all the domains.

14.4.1 Aesthetics

The Aesthetics category of Marathi text corpus covers 23 sub-categories bearing a total of 1515006 words along with the overall percentage of 70.23%. The representational details are given in the table below.

#	Sub Category	Word Count	Percentage (within Subdomain).	Overall Percentage
1	Autobiographies	10,548	0.70%	0.49%
2	Biographies	51,473	3.40%	2.39%
3	Cinema	23,644	1.56%	1.10%
4	Culture	14,222	0.94%	0.66%
5	Fine Arts-Drawing	2,664	0.18%	0.12%
6	Fine Arts-Music	13,225	0.87%	0.61%
7	Folk Tales	247	0.02%	0.01%
8	Folklore	9,980	0.66%	0.46%
9	Humour	19968	1.32%	0.93%
10	Literary Texts	287	0.02%	0.01%
11	Literature-Children's Literature	6,059	0.40%	0.28%
12	Literature-Criticism	173,690	11.46%	8.05%
13	Literature-Diaries	10,969	0.72%	0.51%
14	Literature-Essays	42,994	2.84%	1.99%
15	Literature-Letters	7,922	0.52%	0.37%
16	Literature-Novels	751,321	49.59%	34.83%
17	Literature-Plays	12,694	0.84%	0.59%
18	Literature-Science Fiction	10,472	0.69%	0.49%
19	Literature-Short Stories	249,589	16.47%	11.57%
20	Literature-Speeches	12,403	0.82%	0.57%
21	Literature-Text Books (School)	3,967	0.26%	0.18%
22	Literature-Travelogues	84,906	5.60%	3.94%
23	Mythology	1,762	0.12%	0.08%

	Total	**1515006**	**100**	**70.23%**

Table 14-3: Aesthetics Category Representation

14.4.2 Commerce

The Mass Media category of Marathi text corpus covers 3 sub-categories bearing a total of 20795 words along with the overall percentage of 0.96%. The representational details are given in the table below.

#	Sub Category	Word Count	Percentage (within Subdomain).	Overall Percentage
1	**Finance**	**2,924**	**14.06%**	**0.14%**
2	**Industry**	**6,055**	**29.12%**	**0.28%**
3	**Management**	**11,816**	**56.82%**	**0.55%**
	Total	**20795**	**100**	**0.96%**

Table 14-4: Commerce Category Representation

14.4.3 Mass Media

The Mass Media category of Marathi text corpus covers 8 sub-categories bearing a total of 363122 words along with the overall percentage of 16.83%. The representational details are given in the table below.

#	Sub Category	Word Count	Percentage (within Subdomain).	Overall Percentage
1	Article	55,818	15.37%	2.59%
2	Discussions	80,316	22.12%	3.72%
3	Editorial	1,03,914	28.62%	4.82%
4	General News	56,266	15.50%	2.61%
5	Lead news	15,082	4.15%	0.70%
6	Political	6,769	1.86%	0.31%
7	Social	2,519	0.69%	0.12%
8	Sports News	42,438	11.69%	1.97%
	Total	363122	100	16.83%

Table 14-5: Mass Media Category Representation

14.4.4 Science and Technology

The Science and Technology category of Marathi text corpus covers 17 sub-categories bearing a total of 55902 words along with the overall percentage of 2.59%. The representational details are given in the table below.

#	Sub Category	Word Count	Percentage (within Subdomain).	Overall Percentage
1	Agriculture	7,877	14.09%	0.37%
2	Architecture	2,503	4.48%	0.12%
3	Astronomy	1,590	2.84%	0.07%
4	Ayurveda	2,774	4.96%	0.13%

5	Biology	872	1.56%	0.04%
6	Chemistry	2,098	3.75%	0.10%
7	Educational Psychology	8,014	14.34%	0.37%
8	Engineering-Electrical	804	1.44%	0.04%
9	Engineering-Electronics Communication	2,506	4.48%	0.12%
10	Mathematics	1,002	1.79%	0.05%
11	Medicine	2,324	4.16%	0.11%
12	Psychology	14,296	25.57%	0.66%
13	Statistics	2,131	3.81%	0.10%
14	Textile Technology	3,273	5.85%	0.15%
15	Veterinary	1,335	2.39%	0.06%
16	Yoga	430	0.77%	0.02%
17	Zoology	2,073	3.71%	0.10%
	Total	55902	100	2.59%

Table 14-6: Science and Technology Category Representation

14.4.5 Social Sciences

The Social Sciences category of Marathi text corpus covers 15 sub-categories bearing a total of 202253 words along with the overall percentage of 9.38%. The representational details are given in the table below.

#	Sub Category	Word Count	Percentage (within Subdomain).	Overall Percentage
1	Economics	9,951	4.92%	0.46%
2	Education	18,448	9.12%	0.86%
3	Geography	3,171	1.57%	0.15%
4	Health and Family Welfare	8,312	4.11%	0.39%
5	History	35,203	17.41%	1.63%
6	Journalism	28,516	14.10%	1.32%
7	Law	2,660	1.32%	0.12%
8	Library Science	8,193	4.05%	0.38%
9	Linguistics	17,047	8.43%	0.79%
10	Philosophy	11552	5.71%	0.54%
11	Physical Education	2,145	1.06%	0.10%

12	Political Science	27,012	13.36%	1.25%
13	Public Administration	6,196	3.06%	0.29%
14	Religion/Spiritual	3,764	1.86%	0.17%
15	Sociology	20,083	9.93%	0.93%
	Total	202253	100	9.38%

Table 14-7: Social Sciences Category Representation

14.5 COPYRIGHT CONSENTS

The Marathi text corpus have been collected from various sources and the copyright for the same stays with different sources. However, for the purposes of this corpus, consents have been sought from all the stakeholders. Most of the copyrights (around 83%) belong to private parties with only 17% belonging to the government agencies, either state or the central.

15 ODIA RAW TEXT CORPUS

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

15.1 INTRODUCTION

Odia (formerly Oriya) is the principal and official language of Odisha (formerly Orissa) and second official language of Jharkhand. It is the sixth Classical Status Language as designated by the Govt. of India. Odia is one of the major languages of Indo-Aryan language family. It's written in Odia Script which is developed from Brahmi Script and written from left to right direction. Odia script has also been used as a regional writing system for Sanskrit as well as a number of Dravidian and Munda minority languages spoken in the state.

Odia has been influenced by the English, Arabic, Persian as well as Dravidian languages, especially by Telugu. Its lexicon has been enriched by borrowing from these languages as well as from Marathi and Portuguese. Many Sanskrit words entered the Odia language and literature since time immemorial occurring in two forms: one is 'tatsama' (close to the original form) and another is 'tadbhaba' (remote from the original form). Odia written form uses three diacritics i.e. *bisarga*, anuswaara, and candrabindu. LDC-IL Odia text corpus is collected in Odia script of contemporary usage.

Odia text corpus is collected from various libraries of Odisha, mostly from Bhubaneswar like Eastren Regional Language Centre's library and Harekrushna Mahtab State Library. LDC-IL tried to cover the entire domains/subdomains (categories/subcategories) in its standard list. Some subdomains like novel, short-story have huge amount of books but some subdomains like mythology, philosophy, cinema have very less amount of books. Literary texts are easily available in Odia but getting scientific/knowledge text is very difficult; even some subdomains like criminology, oceanology, geology text are too rare in Odia.

15.2 PECULIARITIES OF ODIA TEXT

The Corpus of Odia text can be broadly classified into two types: 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 that are narrative and it contains elements of fiction. Novel, short-story, play are the examples of literary text. Non-literary texts are texts whose primary purpose is to convey knowledge/information. Example of non-literary texts are text about various scientific or technical subjects, articles/papers in academic journals. In literary text, language has creative elements, cultural information, dialectical variations and ambiguities etc. But technical or scientific terms, foreign words etc. have widely appeared in non-literary texts.

15.2.1 Issues on Halanta, Ya-Phalaa & Purnachheda

It is observed that Odia is a vowel ending language. Consonants will occur only in the initial and medial positions. The consonant ending words found in Odia are either named entities or borrowed/native words from English, Arabic, Persian and Indo-Aryan language influence. When the pure consonants, i.e., stops (halanta) occur in the medial position of a word, they form a cluster as in ପୁଲିସ୍+ର = ପୁଲିସ୍ର.This issue is prevalent in Odia corpus.

Another interesting fact in the same spirit is that of 'Ya Phalaa' ('ୟ-ଫଳା') in such a way that when it occurs with the consonant 'ର'(ra), the cluster form doesn't come out properly. The combination of 'ୟ-ଫଳା' (Ya Phalaa) and 'ର'(ra) gives the result of Repha+ra (ର) = ର୍ୟ, which is not supposed to be. To concretise the fact, words like "Rank", "Racket", "Ragging" cannot be graphically represented in Odia script properly.

In Unicode, the punctuation marks such as Odia purnachheda (।) and dwichheda (॥) have not been introduced so far, and for these, we have used Devanagari dandaa, double dandaa for the same.

15.3 DATA SAMPLING NOTES

15.3.1 Principles of Data Sampling

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

15.3.2 Fieldworks Undertaken

Odia text corpus is collected from various libraries from Odisha as well as from CIIL's library, Mysore. The text materials were collected by conducting 4 fieldworks undertaken in the period from 2010 to 2012. The following resource persons were engaged in the fieldworks. They are Pramod Kumar Rout, Kshirod Kumar Das and Santosh Kumar Mohanty. The greater part of the text has been taken from the Eastren Regional Language Centre's library and Harekrushna Mahtab State Library, Bhubaneswar.

Overall, the following libraries served as the source of the Odia text corpus.

1. Library, Eastren Regional Language Centre, Bhubaneswar.
2. Harekrushna Mahtab State Library, Bhubaneswar.
3. Library, Kedarnath Gaveshana Pratisthan, Bhubaneswar.
4. Parija Library, Utkal University, Bhubaneswar.
5. Prahllad Mallick Central Library, Marshaghai College, Kendrapara.
6. Library, Central Institute of Indian Languages, Mysore.

Collected text materials published from various places within Odisha and New Delhi as well as other country like Russia.

An attempt has been made to cover the entire domains and subdomains in its standard list. Some subdomains like novel, short-story have huge amount of books but some subdomains like cinema, weather, philosophy have very less amount of books. Literary texts are easily available in Odia but getting scientific/ knowledge text is very difficult. Some subdomains like criminology, oceanology, geology text are too rare in Odia.

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

Some time Xerox attendants refused to photocopy randomly selected pages because of the long queue waiting and it takes up more time for them to turn the pages compared to continuous page photocopying they are accustomed to. It was another issue that the fieldworker had to carry a huge list of photocopy bundles with them which was many a times cumbersome to travel with.

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

15.3.3 Data Inputting

All the text has been typed in Unicode compatible font using the InScript Keyboard directly into the XML files. The data has been inputted by Sunil Kumar Pattanayak, Swahashree Sahoo, Rupa H. S. and Radhika M.. Among them Sunil Kumar Pattanayak and Swahashree Sahoo are the native speakers of Odia.

15.3.4 Validation and Normalization Workshops

Linguistic Data Consortium for Indian Languages (LDC-IL) conducted workshops for data validation and normalization. The experts unanimously suggested that the Odia text corpus should remain true to the text.

15.3.5 Proofreading

Odia text data has been proofread by both internal resource persons and the resource persons engaged in the programmes for Corpus Normalization and the Short-Term Goal Oriented Projects (Text Corpus Cleaning Workshops). The following account shows the workshop facet:

9. Corpus Normalization-Odia: 7th June 2010 to 11th June 2010.
10. Short Term Goal Oriented Project- Odia Language Text Corpus Cleaning: 13th November 2012 to 28th December 2012.
11. Short Term Goal Oriented Project- Odia Language Text Corpus Cleaning: 2nd September 2013 to 30th September 2013.

It was so decided and followed across the languages that text manipulation be avoided thoroughly and only the typo errors committed during the input process have been corrected with reference to the source materials/hard copies. The source printed materials collected for the corpus are contemporary, mainly published after 1990. The following resource persons attended in the above-mentioned workshops for Odia corpus. They are Pramod Kumar Rout, Kshirod Kumar Das, Santosh Kumar Mohanty, Kuni Mallick, Sudhir Kumar Barik, Lingaraj Meher and Mohan Kar respectively.

15.4 TRANSLITERATION IN LDC-IL ODIA TEXT CORPUS

For easy reference and uniformity of metadata, some entries in the metadata file, namely *'Title','Headline', 'Author', 'Editor', 'Translator'* are 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	ଶ	ଷ	ସ	ହ
ya	Ya	ra	la	La	va	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

Table 15-1: LDC-IL Odia Transliteration Schema

15.5 COPYRIGHT CONSENTS

The Odia text corpus have been collected from various sources and the copyright for the same stays with different sources. However, for the purposes of this corpus, consents have been sought from all the stakeholders. Most of the copyrights belong to private parties (around 94%) with only 06% belonging to the government agencies, either state or the central.

15.6 OVERVIEW OF REPRESENTED DOMAINS

LDC-IL Odia Text Corpus size is: 15,88,287 words and 1,03,04,173 characters drawn from 206 different titles, including the extracts from newspapers.The data can be categorized into two classes of typed+cleaned and crawled. Only typed+cleaned data is available for Odia.

Text Type	Word Count	Keystroke/Character Count
Typed + Cleaned	15, 88, 287	1,03,04,173
Crawled	0	0
Total	**15, 88, 287**	**1, 03,04, 173**

Table 15-2: Representation of the Typed and Crawled Text in Odia Raw Text Corpus

The following table gives a summary of the typed and cleaned text of the Odia Raw Text Corpus. The representation of the five domains/categories covered has been shown in the table below:

#	Domain	Word Count	Percentage
1	Aesthetics	5,11,887	32.23%
2	Commerce	19,616	1.24%
3	Mass Media	8,02,100	50.50%
4	Science & Technology	31,589	1.99%
5	Social Sciences	2,23,095	14.05%
	Total	**15, 88, 287**	**100.00%**

Table 15-3: Representation of the Domains in Odia Raw Text Corpus

As each domain has several subdomains and total number of subdomains are 32, the following table shows the representation of the several domains, both within the domain and across all the domains.

15.6.1 Aesthetics

The Aesthetics domain/category of LDC-IL Odia text corpus covers 10 subdomains/subcategories bearing a total of 5,11,887 words along with the overall percentage of 32.23%. The representational details are given in the table below.

#	Subdomain	Word Count	Percentage (within Subdomain)	Overall Percentage
1	Autobiographies	14,552	2.84%	0.92%
2	Biographies	9,964	1.95%	0.63%
3	Cinema	486	0.09%	0.03%
4	Culture	9,612	1.88%	0.61%
5	Folklore	1,737	0.34%	0.11%
6	Literature-Criticism	35,499	6.93%	2.24%
7	Literature-Novels	2,50,336	48.90%	15.76%
8	Literature-Short Stories	1,86,531	36.44%	11.74%
9	Literature-Travelogues	2,943	0.57%	0.19%
10	Mythology	227	0.04%	0.01%
	Total	**5, 11, 887**	**100.00%**	**32.23%**

Table 15-4: Representation of Aesthetics Domain

15.6.2 Commerce

The Commerce domain/category of LDC-IL Odia text corpus covers only 1 subdomain/subcategory bearing a total of 19,616 words along with the overall percentage of 1.24%. The representational details are given in the table below.

#	Subdomain	Word Count	Percentage (within Subdomain)	Overall Percentage
1	Business	19,616	100.00%	1.24%

Table 15-5: Representation of Commerce Domain

15.6.3 Mass Media

The Mass Media domain/category of LDC-IL Odia text corpus covers 5 subdomains/subcategories bearing a total of 8,02,100 words along with the overall percentage of 50.50%. The representational details are given in the table below.

#	Subdomain	Word Count	Percentage (within Subdomain)	Overall Percentage
1	Editorial	1,30,416	16.26%	8.21%
2	General News	6,43,286	80.20%	40.50%
3	Social	4,508	0.56%	0.28%
4	Sports News	20,387	2.54%	1.28%
5	Weather	3,503	0.44%	0.22%
	Total	**8, 02,100**	**100.00%**	**50.50%**

Table 15-6: Representation of Mass Media Domain

15.6.4 Sceience and Technology

The Science and Technology domain/category of LDC-IL Odia text corpus covers 4 subdomains/subcategories bearing a total of 31,589 words along with the overall percentage of 1.99%. The representational details are given in the table below.

#	Subdomain	Word Count	Percentage (within Subdomain)	Overall Percentage
1	Agriculture	4, 277	13.54%	0.27%
2	Astrology	822	2.60%	0.05%
3	Medicine	8, 225	26.04%	0.52%
4	Text Book (Science)	18, 265	57.82%	1.15%
	Total	**31, 589**	**100.00%**	**1.99%**

Table 15-7: Representation of Science and Technology Domain

15.6.5 Social Sciences

The Social Science domain/category of LDC-IL Odia text corpus covers 12 subdomains/subcategories bearing a total of 2, 23,095 words along with the overall percentage of 14.05%. The representational details are given in the table below.

#	Subdomain	Word Count	Percentage (within Subdomain)	Overall Percentage
1	Economics	32,162	14.42%	2.02%
2	Education	5,766	2.58%	0.36%
3	Food and Wellness	7,088	3.18%	0.45%
4	Health and Family Welfare	1,288	0.58%	0.08%
5	History	13,016	5.83%	0.82%
6	Linguistics	8,066	3.62%	0.51%
7	Philosophy	1,748	0.78%	0.11%
8	Political Science	44,032	19.74%	2.77%
9	Religion/Spiritual	264	0.12%	0.02%
10	Sociology	11,662	5.23%	0.73%
11	Sports	68,682	30.79%	4.32%
12	Text Book (Social Science)	22,483	10.08%	1.42%
	Total	**2, 23, 095**	**100.00%**	**14.05%**

Table 15-8: Representation of Social Science Domain

15.7 REFERENCES

Mahapatra, Bijay Prasad. 2007, *A SYNCHRONIC GRAMMAR OF ORIYA (Standard Spoken and Written)*. Mysore: Central Institute of Indian Languages.

Mohanty, Santosh Kumar. 2018. 'Unicode Odia Lipirupara Byabahaara'. in *Esana*, *Vol. 76*, Cuttack: Institute of Odia Studies.

******

16 PUNJABI RAW TEXT CORPUS

Poonam Dhillon, Rajesha N, Manasa G, Narayan Choudhary, L. Ramamoorthy

16.1 INTRODUCTION

Punjabi is the principal and administrative language of Punjab. Punjabi is a tonal language with three tones: high falling, low rising and level. Punjabi is not only spoken in Punjab in India it is also a language of Lehnda Punjab in Pakistan. In Pakistani Punjabi is the second most widely-spoken language in Pakistan but has no official status. Punjabi is an Indo-Aryan language. It is derived from Sanskrit through Prakrit languages and later *Apabhrans*. There was no such form of Punjabi language in the beginning that we see today. With the flow of time, it has emerged in the present form. This same Punjabi language is being written in two epigraphs in Gurmukhi and Shahmukhi script. In our Eastern Punjab it is being used in Gurmukhi and Lehnda Punjab (Pakistan) using Shahmukhi script.

Punjabi language with Gurmukhi script:

In 16th century Guru Angad Dev Ji, the second Sikh guru was standardised the Gurmukhi alphabet from the Landa alphabet. Gurmukhi is written from right to left. This script has 10 vowels and 29 consonants and 5 Perso-arabic consonants. It also has two semivowels / y / and / v /.

Punjabi language with Shahmukhi script:

The Shahmukhi alphabet is a version of Perso-Arabic alphabet and used to write Punjabi in Pakistan. Shahmukhi is written from left to right. This script has 10 vowels and 46 consonants and 10 mixed words.

Punjabi is written in Shahmukhi scripts as well. ‘Shahmukhi’ is a variant of ‘Perso-Arabic’ script. LDC-IL Punjabi text corpus is collected in Gurmukhi script of contemporary usage.

Punjabi text corpus is collected from various libraries in Punjab mostly from Patiala. The greater part of the text has been taken from NRLC library and Punjabi University Patiala library. LDC-IL tried to cover the entire category in its standard list. Some categories like novel, short stories has huge amount of books but some categories like physics, chemistry, economics have very less amount of books. Literary texts are easily available in Punjabi but getting scientific text is very difficult. Some categories like epigraphy, finance, oceanology text are not available in Punjabi.

16.2 PECULIARITIES OF PUNJABI TEXT

The Corpus of Punjabi 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 that are narrative and it contains elements of fiction. Novels, short stories, plays are examples of literary text. Non-literary texts are texts 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.

16.2.1 Doubling of consonants

In Punjabi Text corpus the '◌ੰ' (GURMUKHI TIPPI) is used for the nasalization rather than '◌ਂ' (GURMUKHI SIGN ADAK BINDI).
The other predominant feature in Gurumukhi Script is the usage of '◌ੱ' (GURMUKHI ADDAK) which doubles following consonant to which it attaches. Unlike other Indian languages which uses Virama (Halanth) to make the Half-letter followed by full letter, Gurmukhi uses ADDAK.
While processing text this ADDAK has no value until the next consonant is known. This will create problem in text processing applications like transliterator, character based morph analyser etc. that are analyzing and processing character by character of the given text. The text processors need to be enabled with extra feature of checking immediate next character of the ADDAK.

The LDC-IL Corpus uses ADDAK as it occurs naturally in Punjabi Text written in Gurumukhi Script.

16.3 DATA SAMPLING NOTES

16.3.1 Principles of Data Sampling

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

16.3.2 Fieldworks Undertaken

Punjabi text corpus is collected from various libraries in Punjab, mostly from Patiala. The text materials were collected by conducting three fieldworks undertaken in the period from 2008 to 2010. The greater part of the text has been taken from NRLC, Patiala and Punjabi University Patiala library. Overall, the following libraries served as the source of the Punjabi text corpus:

- Northern Regional Language Centre, Patiala
- Punjabi University, Patiala
- Khalsa College, Patiala
- Guru Nanak Dev University, Amritsar
- Regional Campus, Jalandhar

Mostly collected text materials have been published from Punjab and New Delhi.

An attempt has been made to cover the entire category in its standard list. Some categories like criticism, novel and short stories have huge amount of books but some categories like physics, chemistry, economics have very less amount of books. Literary texts are easily available in Punjabi but getting scientific text is very difficult. Some categories like epigraphy, finance, oceanology text are not available in Punjabi.

Collecting text data from the field is a difficult job. Most of the libraries do not allow to take 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 to take many books at a time, the photocopy kiosk had issues as there was a long queue.

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

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

16.3.3 Data Inputting

All the text has been typed in Unicode using the InScript Keyboard directly onto the XML files. The data has been inputted by Harjinder Singh, Gurmeet Kaur, Harpreet Kaur, Kulwant Singh, native speaker of Punjabi, but Radhika M, Syeda Aliya Habeeba native speaker of Kannada.

16.3.4 Validation and Normalization Workshops

A 5-day workshop was conducted at Linguistic Data Consortium from November 28 to December 1, 2011 with Prof. Joga Singh Department of Linguistics, Punjabi University, Patiala, Prof. Baldev Singh Cheema Department of Punjabi, Punjabi University, Patiala and Prof. Sukhwinder Singh Sangha from Department of Punjabi, Regional Campus, Jalandhar as experts. The experts suggested that the Punjabi text corpus should remain true to the text.

16.3.5 Proofreading

Punjabi text data has been proof read by internal and external resource persons. The text has always been kept true to the printed material and typos, if any, occurring at the time of typing have only been corrected. Some text cleaning workshops were conducted using external resources wherein the Punjabi text was cleaned/proofread by the native speakers. An account of such workshops is as below:

12. July 2010
13. 24th Dec. 2012 – 28th Feb. 2013
14. 05 Aug-18 Sept 2013
15. 03 Oct-07 Dec 2016
16. 01 May-05 July 2018

The printed materials collected for the corpus is contemporary, mainly published after 1980.

16.3.6 Data Extracted from Websites

Punjabi News corpus data is extracted from News websites of "Ajit Weekly" (www.ajitweekly.com) , " Charhdikala" (www.charhdikala.com), " Europe Vich Punjabi" (www.europevichpunjabi.com), " Pardes News " (www.pardesnews.com) " Parvasi " (www.parvasi.com) " Punjab Express " (www.punjabexpress.com) , " Punjabi Webdunia " (www.punjabi.webdunia.com), and " Quami Ekta " (www.quamiekta.com). The news content was categorized based on the content of the text and archived. The period of selection of the news corpus ranges from 2008 to 2010.

16.4 TRANSLITERATIONS IN LDC-IL PUNJABI TEXT CORPUS

For easy reference and uniformity of metadata, some entries in the metadata file, namely *'Title', 'Headline', 'Author', 'Editor', 'Translator'* are transliterated from Gurmukhi to Roman letters. Numeric characters were transliterated from Gurmukhi to Hindu-Arabic system.

For such purpose the LDC-IL transliteration scheme for Gurmukhi to Roman characters is given below:

LDC-IL Transliteration Schema

Gurmukhi characters to Roman and Gurmukhi Numerals to Hindu-Arabic

Vowels

ਅ	ਆ	ਇ	ਈ	ਉ	ਊ	ਏ	ਐ	ਓ	ਔ
	ਾ	ਿ	ੀ	ੁ	ੂ	ੇ	ੈ	ੋ	ੌ
a	A	i	I	u	U	E	ai	O	au

Consonants

ਕ	ਖ	ਗ	ਘ	ਙ
ka	kha	ga	gha	ng'a
ਚ	ਛ	ਜ	ਝ	ਞ
ca	cha	ja	jha	nj'a
ਟ	ਠ	ਡ	ਢ	ਣ
Ta	Tha	Da	Dha	Na
ਤ	ਥ	ਦ	ਧ	ਨ
ta	tha	da	dha	na
ਪ	ਫ	ਬ	ਭ	ਮ
pa	pha	ba	bha	ma

Symbols

ੱ	ੰ	ਂ	ਃ
Null	m'	M	H

ਯ	ਰ	ਲ	ਵ	ੜ	ਸ਼	ਖ਼	ਗ਼	ਜ਼	ਫ਼	ਲ਼
ya	ra	la	va	Ra	sha	Kh'a	g'a	j'a	ph'a	La

Numerals (Punjabi to Hindu-Arabic)

੦	੧	੨	੩	੪	੫	੬	੭	੮	੯
0	1	2	3	4	5	6	7	8	9

The greyed out characters are obsolete. They may rarely present in the current LDC-IL corpus.

16.5 OVERVIEW OF REPRESENTED DOMAINS

LDC-IL Punjabi Text Corpus size is: 1,01,25,770 words and characters count is 5,08,24,349 drawn from 2,470 different titles, including the extracts from newspapers and magazines. The data can be categorized

into two classes of typed+cleaned and crawled. The crawled data has been crawled mainly from news websites and archived using the standard processing of LDC-IL text corpus preparation.

The following table gives a summary of the typed and crawled text of the Punjabi Raw Text Corpus.

Text Type	Word Count	Keystroke/Character Count
Typed+Cleaned	97,55,905	4,89,97,317
Crawled	3,69,865	18,27,032
Total	1,01,25,770	5,08,24,349

Table 16-1: Overivew of word count and Character count

The representation of the five major domains covered has been shown in the table below:

Domain	Domain Word Count	Percentage
Aesthetics	41,90,199	41.38%
Commerce	56,205	00.56%
Social Sciences	12,20,366	12.05%
Mass Media	42,74,922	42.22%
Science & Technology	3,84,078	03.79%
Total	1,01,25,770	100.00%

Table 16-2: Representation of the Domains in Punjabi Text Corpus

As each domain has several sub-domains, the following table shows the representation of the several domains, both within the domain and across all the domains.

16.5.1 Aesthetics

The Aesthetics category of Punjabi text corpus covers 22 sub-categories bearing a total of 41,90,199 words along with the overall percentage of 41.38%

%. The representational details are given in the table below.

#	Sub Category	Word Count	% within Subdomain	Overall Percentage
1	Autobiographies	1,45,184	3.46%	1.43%
2	Biographies	2,60,595	6.22%	2.57%
3	Cinema	73,865	1.76%	0.73%
4	Culture	1,37,921	3.29%	1.36%
5	Fine Arts-Dance	27,289	0.65%	0.27%
6	Fine Arts-Drawing	8,780	0.21%	0.09%
7	Fine Arts-Sculpture	36,964	0.88%	0.37%
8	Fine Arts-Music	35,652	0.85%	0.35%
9	Folklore	1,04,858	2.50%	1.04%
10	Humour	1,584	0.04%	0.02%
11	Literary Texts	4,27,152	10.19%	4.22%
12	Literature-Children's Literature	1,693	0.04%	0.02%
13	Literature-Criticism	14,83,799	35.41%	14.65%
14	Literature-Diaries	25,845	0.62%	0.26%
15	Literature-Letters	10,501	0.25%	0.10%
16	Literature-Novels	4,71,785	11.26%	4.66%
17	Literature-Plays	54,775	1.31%	0.54%
18	Literature-Short Stories	6,65,293	15.88%	6.57%
19	Literature-Speeches	83,710	2.00%	0.83%
20	Literature-Travelogues	1,07,341	2.56%	1.06%
21	Literature-Text Books (Schools)	23,192	0.55%	0.23%
22	Mythology	2,421	0.06%	0.02%
	Total	**41,90,199**	**100%**	**41.38%**

Table 16-3: Aesthetics category representation in Punjabi Text Corpus

16.5.2 Commerce

The Commerce category of Punjabi text corpus covers 2 sub-categories bearing a total of 56,205 words along with the overall percentage of 0.56%. The representational details are given in the table below.

#	Sub Category	Word Count	% within Subdomain	Overall Percentage
1	Business	45,159	80.35%	0.45%
2	Management	11,046	19.65%	0.11%
	Total	**56,205**	**100%**	**0.56%**

Table 16-4: Commerce category representation in Punjabi Text Corpus

16.5.3 Social Science

The Social Science category of Punjabi text corpus covers 17 sub-categories bearing a total of 12,20,366 words along with the overall percentage of 12.05%. The representational details are given in the table below.

#	Sub Category	Word Count	% within Subdomain	Overall Percentage
1	Economics	1,58,038	12.95%	1.56%
2	Education	86,299	7.07%	0.85%
3	Food and Wellness	1,261	0.10%	0.01%
4	Geography	18,317	1.50%	0.18%
5	Health and Family Welfare	16,859	1.38%	0.17%
6	History	1,69,151	13.86%	1.67%
7	Home Science	7,115	0.58%	0.07%
8	Journalism	46,378	3.80%	0.46%
9	Law	22,057	1.81%	0.22%
10	Library Science	20,977	1.72%	0.21%
11	Linguistics	1,33,404	10.93%	1.32%
12	Physical Education	54,906	4.50%	0.54%
13	Political Science	1,25,997	10.32%	1.24%
14	Public Administration	76,345	6.26%	0.75%
15	Religion/Spiritual	1,81,472	14.87%	1.79%
16	Sociology	91,775	7.52%	0.91%
17	Sports	10,015	0.82%	0.10%
	Total	**12,20,366**	**100 %**	**12.05%**

Table 16-5: Social Science category representation in Punjabi Text Corpus

16.5.4 Mass Media

The Mass Media category of Punjabi text corpus covers 14 sub-categories bearing a total of 42,74,922 words along with the overall percentage of 42.22%. The representational details are given in the table below.

#	Sub Category	Word Count	% within Subdomain	Overall Percentage
1	Business News	1,20,017	2.81%	1.19%
2	Cinema News	1,08,318	2.53%	1.07%
3	Classifieds	785	0.02%	0.01%
4	Discussions	1,610	0.04%	0.02%
5	Editorial	9,14,446	21.39%	9.03%
6	General News	22,56,487	52.78%	22.28%
7	Health	14,578	0.34%	0.14%
8	Interviews	21,589	0.51%	0.21%
9	Letters	3,094	0.07%	0.03%

10	Political	4,82,448	11.29%	4.76%
11	Religious / Spiritual News	26,957	0.63%	0.27%
12	Social	27,810	0.65%	0.27%
13	Speeches	3,054	0.07%	0.03%
14	Sports News	2,93,729	6.87%	2.90%
	Total	**42,74,922**	**100%**	**42.22%**

Table 16-6: Mass Media category representation in Punjabi Text Corpus

16.5.5 Science & Technology

The Social Science category of Punjabi text corpus covers 17 sub-categories bearing a total of 3,84,078 words along with the overall percentage of 3.79%. The representational details are given in the table below.

#	Sub Category	Word Count	% within Subdomain	Overall Percentage
1	Agriculture	42,294	11.01%	0.42%
2	Astrology	11,990	3.12%	0.12%
3	Ayurveda	40,680	10.59%	0.40%
4	Bio Chemistry	24009	6.25%	0.24%
5	Botany	21,913	5.71%	0.22%
6	Computer Sciences	44,164	11.50%	0.44%
7	Criminology	6,175	1.61%	0.06%
8	Environmental Science	22,797	5.94%	0.23%
9	Forestry	9,448	2.46%	0.09%
10	Homeopathy	29,850	7.77%	0.29%
11	Medicine	31,978	8.33%	0.32%
12	Naturopathy	6,199	1.61%	0.06%
13	Physics	19,661	5.12%	0.19%
14	Psychology	21,398	5.57%	0.21%
15	Text Book (Science)	14,584	3.80%	0.14%
16	Yoga	9,903	2.58%	0.10%
17	Zoology	27,035	7.04%	0.27%
	Total	**3,84,078**	**100%**	**3.79%**

Table 16-7: Science & Technology category representation in Punjabi Text Corpus

16.6 COPYRIGHT CONSENTS

The Punjabi text corpus have been collected from various sources and the copyright for the same stays with different sources. However, for the purposes of this corpus, consents have been sought from all the stakeholders. Most of the copyrights (around 73%) belong to private parties with only 27% belonging to the government agencies, either state or the central.

17 TAMIL RAW TEXT CORPUS

Amudha R, Premkumar L.R, Rajesha N, Manasa G, Narayan Choudhary, L. Ramamoorthy

17.1 INTRODUCTION

Tamil is one of the oldest language in the world. It is spoken in all over the world particularly in India, Sri Lanka, Mauritius, Singapore, Malaysia. The language is an official language in Tamilnadu and some of the foreign countries such as Sri Lanka and Singapore. It has official status in the Indian state of Tamilnadu and the Indian Union Territory of Puducherry. It is used as one of the languages of education in Malaysia, along with English, Malay and Mandarin. Tamil is spoken by significant minorities in the four other South Indian states of Kerala, Karnataka, Andhra Pradesh and Telangana and the Union Territory of the Andaman and Nicobar Islands. It is one of the 22 scheduled languages of India. A recorded Tamil literature has been documented for over 2000 years. The earliest period of Tamil literature, Sangam literature, is dated from ca. 300 BC – AD 300.

Tamil language inscriptions written in Brahmi script have been discovered in Sri Lanka and on trade goods in Thailand and Egypt. In 1578, Portuguese Christian missionaries published a Tamil prayer book in old Tamil script named Thambiraan Vanakkam, thus making Tamil the first Indian language to be printed and published. The Tamil Lexicon, published by the University of Madras, was one of the earliest dictionaries published in the Indian languages. According to a 2001 survey, there were 1,863 newspapers published in Tamil, of which 353 were dailies.

The Tamil script, like the other Brahmic scripts, is thought to have evolved from the original Brahmi script. The earliest inscriptions which are accepted examples of Tamil writing date to a time just after the Ashokan period. The script used by such inscriptions is commonly known as the Tamil-Brahmi, or "Tamili script", and differs in many ways from standard Ashokan Brahmi. For example, early Tamil-Brahmi, unlike Ashokan Brahmi, had a system to distinguish between pure consonants (as in m) and consonants with an inherent vowel (as in ma). In addition, according to Iravatham Mahadevan, early Tamil Brahmi used slightly different vowel markers, had extra characters to represent letters not found in Sanskrit, and omitted letters for sounds not present in Tamil such as voiced consonants and aspirates. Inscriptions from the 2nd century use a latter form of Tamil-Brahmi, which is substantially similar to the writing system described in the Tolkappiyam, an ancient Tamil grammar. Most notably, they used the puḷḷi to suppress the inherent vowel. The Tamil letters thereafter evolved towards a more rounded form, and by the 5th or 6th century, they had reached a form called the early vaṭṭeḻuttu.

The modern Tamil script does not, however, descend from that script. In the 6th century, the Pallava dynasty created a new script for Tamil, and the Grantha alphabet evolved from it, adding the Vatteluttu alphabet for sounds not found to write Sanskrit. Parallel to Pallava script a new script (Chola-Pallava script, which evolved to modern Tamil script) again emerged in Chola territory resembling the same glyph development like Pallava script, but it did not evolve from that. By the 8th century, the new scripts supplanted Vatteluttu in the Chola resp. Pallava kingdoms which lay in the north portion of the Tamil-speaking region. However, the Vatteluttu was still continued to be used in the southern portion of the Tamil-speaking region, in the Chera and Pandyan kingdoms until the 11th century, when the Pandyan kingdom was conquered by the Cholas.

17.2 PECULIARITIES OF TAMIL TEXT

The Corpus of Tamil 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 that are narrative and it contains elements of fiction. Novels, short stories, plays are examples of literary text. Non-literary texts are texts 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.

Average word length of Tamil text is comparatively higher among the scheduled languages of India. Just second to Malayalam in this regard. Tamil is highly agglutinative and morphologically rich language; hence the saturation level of Tamil i.e. the new words coming into corpus for a unit amount of input is much higher compared to other languages. One needs to have much larger text corpora for good coverage of words.

Tamil has 12 vowels and 18 consonants. The language doesn't have letter for voiced sounds like other Dravidian languages. Only the pronunciation would be varied according to the context. For Example, in word 'kakka', initial position of 'k' and germination 'kk' will be pronounced as voice –less sound 'k'. But the occurrence of followed by the nasal sound, 'k' will be pronounced as voiced sound 'g' and 'k' will be pronounced as 'x' when it occurs in between vocalic.

Furthermore, there are six more letters which are called Grantha letters. They are *ja, sha, sa, ha , ksha, srI,* where *ksha* and *srI* are consonant clusters. These letters are used for writing Sanskrit or Prakrit words. There is no initial clusters and no stop, க், ச், ட், த், ப், ற், ending words. The five letters namely, ங், ஞ், ந், வ், ழ் do not occur word finally.

17.3 DATA SAMPLING NOTES

17.3.1 Principles of Data Sampling

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

17.3.2 Field Works Undertaken

Tamil text corpus is collected from various libraries in Tamilnadu, mostly from Chennai. The text materials were collected by conducting four field works undertaken in the period from 2008 to 2012. The greater part of the text has been taken from Kannimara library. Overall, the following libraries served as the source of the Tamil text corpus:

1. Kamarajar University, Madurai.
2. Kannimara Library, Chennai
3. Tamil University Library, Thanjavur
4. International Institute of Tamil Studies
5. CIIL –Library, Central Institute of Indian Language, Mysore
6. Southern Regional Language Center Library, Mysore
7. Grant-in-Aid, Central Institute of Indian Language, Mysore
8. NTS Library, Central Institute of Indian Language, Mysore

Collected text materials have been published at various places within Tamilnadu and other states of India such as Karnataka, Kerala, Maharashtra, Delhi as well as other countries such as Srilanka, Malaysia, USA etc.

An attempt has been made to cover the entire category in its standard list. Some categories like novel, short stories have huge amount of books but some categories like physics, chemistry, economics, scientific text, epigraphy, finance, oceanology have very less amount of books. Literary texts are easily available in Tamil.

Collecting text data from the field is a difficult job. Most of the libraries do not allow to take 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 to take many books at a time, the photocopy kiosk had issues as there was a long queue.

Some time Xerox attendants refused to photocopy randomly selected pages because of the long queue waiting and it takes up more time for them to turn the pages compared to continuous page photocopying they are accustomed to. It was another issue that the field worker/linguist had to carry a huge list of photocopy bundles with them which was many times cumbersome to travel with.

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

17.3.3 Data Inputting

All the text has been typed in Unicode using the InScript Keyboard directly onto the XML files. The data has been inputted by Manivasuki. D, Poorna Mary C, and Moksha Rani a native speaker of Tamil.

17.3.4 Validation and Normalization Workshops

A workshop was conducted at Linguistic Data Consortium from June-2010 to 10-July-2010 in presence of subject experts Prof. C. Karthikeyan, Department of Tamil University, Thanjavur, Prof. G. Ravisankar, Department of Linguistics, PILC, Pondicherry and Prof. Sudarshan, Department of Linguistics, PILC, Pondicherry. The experts suggested that the Tamil text corpus should remain true to the text.

17.3.5 Proofreading

Tamil text data has been proofread by internal resource persons and also resource persons from outside by conducting short-term project at LDC-IL. The text has always been kept true to the printed material and types, if any, occurring at the time of typing have only been corrected.

The printed materials collected for the corpus is contemporary, mainly published after 1904.

17.3.6 Data Extracted from Web Sites

The Tamil News corpus data has been extracted from the following website:

askmathi.googlepages.com, dinamalar.com, dinamani.com (http://dinamalar.com), (http://in.tamil.yahoo.com), (http://in.tamil.yahoo.com/News/National), (http://jeyamohan.in), (http://tamil.webdunia.com), tamil.webdunia.com (http://webdunia.com, jeyamohan.in), kalachuvadu.com, mathimaran.wordpress.com, nakkheeran.in, Sportsdinamalar.com, tamil.sify.com, tamilskynews.com, thatstamil.oneindia.in, theekkathir.in, truetamilans.blogspot.com,vikatan.com, www.aaraamthinai.com, www.dinamalar.com/weeklys, www.dinamani.com, www.dinamani.com/edition, www.puthinam.com, and www.tamilish.com, aaraamthinai.com. The news content was categorized based on the content of the text and archived. The period of selection of the news corpus ranges from 2008 to 2010.

17.4 TRANSLITERATIONS IN LDC-IL TAMIL TEXT CORPUS

For easy reference and uniformity of metadata, some entries in the metadata file, namely *'Title', 'Headline', 'Author', 'Editor', 'Translator'* are 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

Aytam: Tamil has a phoneme called the aytam, written as 'ஃ', which predominately used in old Tamil. Some Tamil grammarians classified it as a dependent phoneme, but it is very rare in modern Tamil. The aytam, in modern Tamil, is also used to convert p to f when writing English words using the Tamil script. Ayutha ezhuthu is rarely used in the Tamil language. Aaytha ezhuthu, also known by a variety of names: 'muppaal pulli', 'thaninilai', 'aghenam', etc, is a unique and special character to Tamil language and script, occring in a few words like 'அஃது' (ahdhu) & 'எஃகு' (ehhu)

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

17.5 OVERVIEW OF REPRESENTED DOMAINS

LDC-IL Tamil Text Corpus size is: **1,09,31,902** words drawn from **1,963** different titles, including the extracts from newspapers. The data can be categorized into two classes of typed cleaned and crawled. The crawled data has been crawled mainly from news websites and archived using the standard processing of LDC-IL text corpus preparation.

The following table gives a summary of the typed and crawled text of the Tamil Raw Text Corpus.

Text Type	Word Count	Keystroke/Character Count
Typed+Cleaned	9872341	90422391
Crawled	1059561	9624028
Total	**1,09,31,902**	**10,00,46,419**

Table 17-1: The typed and crawled text of the Tamil Raw Text Corpus

Domain	Word Count	Percentage
Aesthetics	5595316	51.18%
Commerce	83148	0.76%
Mass Media	2100226	19.21%
Official Document	12768	0.12%
Science and Technology	886532	8.11%
Social Sciences	2253912	20.62%
Total	**1,09,31,902**	**100.00%**

Table 17-2: Representation of the Domains in Tamil Raw Text Corpus

As each domain has several sub-domains, the following table shows the representation of the several domains, both within the domain and across all the domains.

17.5.1 Aesthetics

The Aesthetics Category of LDC-IL Tamil Raw text corpus covers 29 sub domains. The details of the representation of subdomains is given below

#	Subdomain	Word Count	Percentage within Subdomain	Overall Percentage
1	Autobiographies	34949	0.62%	0.32%
2	Biographies	186179	3.33%	1.70%
3	Cinema	158177	2.83%	1.45%
4	Culture	81774	1.46%	0.75%
5	Fine Arts-Dance	106406	1.90%	0.97%
6	Fine Arts-Drawing	11651	0.21%	0.11%
7	Fine Arts-Hobbies	2231	0.04%	0.02%
8	Fine Arts-Music	36149	0.65%	0.33%
9	Fine Arts-Musical Instruments	15581	0.28%	0.14%
10	Fine Arts-Sculpture	61577	1.10%	0.56%
11	Folklore	93196	1.67%	0.85%
12	Handicrafts	14882	0.27%	0.14%
13	Humour	9117	0.16%	0.08%
14	Literary Texts	62803	1.12%	0.57%
15	Literature-Children's Literature	31662	0.57%	0.29%
16	Literature-Criticism	598266	10.69%	5.47%
17	Literature-Diaries	16510	0.30%	0.15%
18	Literature-Epics	18125	0.32%	0.17%
19	Literature-Essays	1953945	34.92%	17.87%
20	Literature-Letters	31072	0.56%	0.28%
21	Literature-Novels	1301291	23.26%	11.90%
22	Literature-Plays	34095	0.61%	0.31%
23	Literature-Poetry	28392	0.51%	0.26%
24	Literature-Science Fiction	35815	0.64%	0.33%
25	Literature-Short Stories	588408	10.52%	5.38%
26	Literature-Speeches	16093	0.29%	0.15%
27	Literature-Text Books (School)	26646	0.48%	0.24%
28	Literature-Travelogues	35498	0.63%	0.32%
29	Mythology	4826	0.09%	0.04%
	Total	**5595316**	**100.00%**	**51.18%**

Table 17-3: Aesthetics Category Representation

17.5.2 Commerce

The Commerce Category of LDC-IL Tamil Raw text corpus covers 5 subdomains. The details of the representation of subdomanins is given below

#	Subdomain	Word Count	Percentage within Subdomain	Overall Percentage
1	Accountancy	5400	6.49%	0.05%
2	Banking	22515	27.08%	0.21%
3	Finance	1001	1.20%	0.01%
4	Industry	8331	10.02%	0.08%
5	Management	45901	55.20%	0.42%
	Total	**83148**	**100.00%**	**0.76%**

Table 17-4: Commerce Category Representation

17.5.3 Mass Media

The Mass Media Category of LDC-IL Tamil Raw text corpus covers 17 subdomains. The details of the representation of subdomanins is given below

#	Subdomain	Word Count	Percentage within Subdomain	Overall Percentage
1	Article	146084	6.96%	1.34%
2	Business News	29858	1.42%	0.27%
3	Cinema News	37027	1.76%	0.34%
4	Classifieds	10324	0.49%	0.09%
5	Discussions	46177	2.20%	0.42%
6	Editorial	165176	7.86%	1.51%
7	General News	1084080	51.62%	9.92%
8	Health	5163	0.25%	0.05%
9	Interviews	83326	3.97%	0.76%
10	Letters	220	0.01%	0.00%
11	Obituary	13487	0.64%	0.12%
12	Political	192753	9.18%	1.76%
13	Religious/Spiritual News	8242	0.39%	0.08%
14	SMS	175	0.01%	0.00%
15	Social	229088	10.91%	2.10%
16	Speeches	3110	0.15%	0.03%
17	Sports News	45936	2.19%	0.42%
	Total	**2100226**	**100.00%**	**19.21%**

Table 17-5: Mass Media Category Representation

17.5.4 Official Document

The Official Document Category of LDC-IL Tamil Raw text corpus covers one subdomain. The details of the representation of subdomanins is given below

Subdomain	Word Count	Percentage within Subdomain	Overall Percentage
Parliamentary/Assembly Debates	12768	100.00%	0.12%

Table 17-6: Official Document Category Representation

17.5.5 Science and Technology

The Science and Technology Category of LDC-IL Tamil Raw text corpus covers 34 subdomains. The details of the representation of subdomanins is given below

#	Subdomain	Word Count	Percentage within Subdomain	Overall Percentage
1	Agriculture	111492	12.58%	1.02%
2	Architecture	26039	2.94%	0.24%
3	Astrology	40202	4.53%	0.37%
4	Astronomy	41359	4.67%	0.38%
5	Ayurveda	28660	3.23%	0.26%
6	Bio Chemistry	16894	1.91%	0.15%
7	Biology	3041	0.34%	0.03%
8	Botany	22303	2.52%	0.20%
9	Chemistry	7474	0.84%	0.07%
10	Computer Sciences	40042	4.52%	0.37%
11	Criminology	3808	0.43%	0.03%
12	Engineering-Electrical	8527	0.96%	0.08%
13	Engineering-Electronics Communication	344	0.04%	0.00%
14	Engineering-Mechanical	513	0.06%	0.00%
15	Environmental Science	7563	0.85%	0.07%
16	Film Technology	16110	1.82%	0.15%
17	Forestry	9923	1.12%	0.09%
18	Geology	17514	1.98%	0.16%
19	Homeopathy	9224	1.04%	0.08%
20	Language Technology	5750	0.65%	0.05%
21	Mathematics	19609	2.21%	0.18%
22	Medicine	165592	18.68%	1.51%
23	Micro Biology	4378	0.49%	0.04%
24	Naturopathy	23599	2.66%	0.22%
25	Oceanology	19722	2.22%	0.18%
26	Physics	23507	2.65%	0.22%
27	Psychology	41301	4.66%	0.38%
28	Sexology	23829	2.69%	0.22%
29	Statistics	491	0.06%	0.00%
30	Text Book (Science)	10444	1.18%	0.10%
31	Textile Technology	16296	1.84%	0.15%
32	Veterinary	28524	3.22%	0.26%
33	Yoga	14761	1.67%	0.14%
34	Zoology	77697	8.76%	0.71%
	Total	**886532**	**100.00%**	**8.11%**

Table 17-7: Science and Technology Category Representation

17.5.6 Social Sciences

The Social Sciences Category of LDC-IL Tamil Raw text corpus covers 24 subdomains. The details of the representation of subdomanins is given below

#	Subdomain	Word Count	Percentage within Subdomain	Overall Percentage
1	Anthropology	28691	1.27%	0.26%
2	Archeology	55237	2.45%	0.51%
3	Economics	95044	4.22%	0.87%
4	Education	281997	12.51%	2.58%
5	Epigraphy	6791	0.30%	0.06%
6	Fisheries	22869	1.01%	0.21%
7	Food and Wellness	7588	0.34%	0.07%
8	Geography	59372	2.63%	0.54%
9	Health and Family Welfare	58395	2.59%	0.53%
10	History	428100	18.99%	3.92%
11	Home Science	22206	0.99%	0.20%
12	Journalism	208532	9.25%	1.91%
13	Law	102249	4.54%	0.94%
14	Library Science	23973	1.06%	0.22%
15	Linguistics	63305	2.81%	0.58%
16	Personality Development	451	0.02%	0.00%
17	Philosophy	57849	2.57%	0.53%
18	Physical Education	33405	1.48%	0.31%
19	Political Science	76516	3.39%	0.70%
20	Public Administration	20612	0.91%	0.19%
21	Religion/Spiritual	340860	15.12%	3.12%
22	Sociology	207593	9.21%	1.90%
23	Sports	46787	2.08%	0.43%
24	Text Book (Social Science)	5490	0.24%	0.05%
	Total	**2253912**	**100.00%**	**20.62%**

Table 17-8: Social Sciences Category Representation

17.6 COPYRIGHT CONSENTS

The Tamil text corpus has been collected from various sources and the copyright for the same stays with different sources. However, for the purposes of this corpus, consent has been sought from all the stakeholders. Most of the copyrights (around 93%) belong to private parties with only 07% belonging to the government agencies, either state or the central.

18 TELUGU RAW TEXT CORPUS

Sajila S, Rajesha N, Manasa G, Narayan Choudhary, L. Ramamoorthy

18.1 INTRODUCTION

Telugu is the principal and official language of Andhra Pradesh and Telangana. It was also referred to as `Tenugu' in the past. Telugu language is the largest member of the Dravidian language family. Telugu, primarily spoken in south-eastern India, is the official language of the states of Andhra Pradesh and Telangana. Among the Dravidian languages, Telugu is spoken by the largest population. Based on 2011 census after Hindi and Bengali, Telugu is the third most frequently spoken Indian language. Telugu also has official language status in the Yanam district of the union territory of Puducherry.

Telugu language has four major dialects namely (i) Northern Telugu dialect spoken in Telangana region (10 districts) (ii) Southern Telugu dialect spoken in Rayalaseema region (4 districts), Nellore and Prakasam districts (iii) Eastern Telugu dialect spoken in Visakhapatnam, Vijayanagaram and Srikakulam districts and (iv) Central Coastal Telugu dialect which is considered as modern Standard Telugu dialect (Krishnamurti and Gwynn 1985) spoken in Guntur, Krishna, East and West Godavari. Its vocabulary is very much influenced by Sanskrit. In the course of time, some Sanskrit expressions used in Telugu got so naturalized that people regarded them as pure Telugu words. With the advent of the Muslim rule, several Persian and Arabic words entered into the Telugu language. Telugu script is originated from Brahmi script. The Brahmi script was used by Mauryan kings. The Bhattiprolu script is a variant of the Brahmi script which has been found in old inscriptions The Bhattiprolu Brahmi script evolved to become the Telugu script by 5th century.

LDC-IL Telugu text corpus is collected in Telugu script of contemporary usage. Telugu text corpus is collected from various libraries in Andhra Pradhesh. Telugu text corpus is collected from various libraries in Andhra Pradesh, mostly from Hyderabad, Vishakahppatanam, Kuppam, Guntoor, Thirupathi and Ananthpur. LDC-IL tried to cover the entire category in its standard list. Some categories like novel, short stories have huge amount of books but some categories like physics, chemistry, economics have very less amount of books. Literary texts are easily available in Telugu but getting scientific text is very difficult. Some categories like epigraphy, finance, Commerce, oceanology text are rare in these libraries.

18.2 PECULIARITIES OF TEXT

The Corpus of Telugu 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 that are narrative and it contains elements of fiction. Novels, short stories, plays are examples of literary text. Non-literary texts are texts 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.

18.3 DATA SAMPLING NOTES

18.3.1 Principles of Data Sampling

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

18.3.2 Field Works Undertaken

Telugu text corpus is collected from various libraries in Andhra Pradhesh, mostly from Hyderabad, Vishakhapatnam, Kuppam, Guntoor, Thirupathi and Ananthpur. All text materials were collected by conducting four field works undertaken in the period from 2010 to 2012.

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

- Central University of Hyderabad, Hyderabad.
- Osmania University, Hyderabad.
- Sri Venkateswara University, Tirupati
- Sri Krishnadevaraya University, Ananthpur.
- Acharya Nagarjuna University, Guntur
- Andhra University, Visakhapatnam
- Southern Regional Language Center Library, Mysore

Collected text materials have been published at various places within Andhra Pradhesh ,Telangana and other states of India such as Karnataka, Tamilnadu, Maharashtra, Delhi.

An attempt has been made to cover the entire category in its standard list. Some categories like novel, short stories have huge amount of books but some categories like physics, chemistry, economics have very less amount of books. Literary texts are easily available in Telugu but getting scientific text is very difficult. Some categories like epigraphy, finance, oceanology text are too rare in Telugu.

Collecting text data from the field is a difficult job. Most of the libraries do not allow to take 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 to take 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 takes up more time for them to turn the pages compared to continuous page photocopying they are accustomed to. It was another issue that the field worker/linguist had to carry a huge list of photocopy bundles with them which was many times cumbersome to travel with.

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

18.3.3 Data Inputting

All the text has been typed in Unicode using the InScript Keyboard directly onto the XML files. The data has been inputted by Mrs.Rajeshwari, a native speaker of Telugu.

18.3.4 Validation and Normalization Workshops

No validation workshops are done for Telugu.

18.3.5 Proof reading

Telugu text data has been proof read by internal resource persons. The text has always been kept true to the printed material and typos, if any, occurring at the time of typing have only been corrected.

18.3.6 Data Extracted from Web Sites

Telugu News corpus data is extracted from the following news websites, eenadu, (http://www.eenadu.net), Sakshi (http://www.sakshi.com) , Andhrajyothi (http:www.andhrajyothy.com). The news content was categorized based on the content of the text and archived. The period of selection of the news corpus ranges from 2008 to 2012.

18.4 TRANSLITERATIONS IN LDC-IL TEXT CORPUS

For easy reference and uniformity of metadata, some entries in the metadata file, namely *'Title', 'Headline', 'Author', 'Editor', 'Translator'* are transliterated from Telugu to Roman letters. Numeric characters were transliterated from Telugu to Hindu-Arabic system.

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

LDC-IL Transliteration Scheme
Telugu characters to Roman and Telugu Numerals to Hindu-Arabic

Vowels and Vowel Signs*																
అ	ఆ	ఇ	ఈ	ఉ	ఊ	ఋ	ౠ	ఌ	ఎ	ఏ	ఐ	ఒ	ఓ	ఔ		
	ా	ి	ీ	ు	ూ	ృ	ౄ	ౢ	ె	ే	ై	ొ	ో	ౌ	ం	ః
a	A	i	I	u	U	x	X	q	e	E	ai	o	O	au	M	H

Consonants				
క	ఖ	గ	ఘ	ఙ
ka	kha	ga	gha	ng'a
చ	ఛ	జ	ఝ	ఞ
ca	cha	ja	jha	nj'a
ట	ఠ	డ	ఢ	ణ
Ta	Tha	Da	Dha	Na
త	థ	ద	ధ	న
ta	tha	da	dha	na
ప	ఫ	బ	భ	మ
pa	pha	ba	bha	ma

య	ర	ల	వ	ళ	శ	ష	స	హ	ఱ
Ya	ra	la	va	La	sha	Sa	sa	ha	ṛ

Numerals									
౦	౧	౨	౩	౪	౫	౬	౭	౮	౯
0	1	2	3	4	5	6	7	8	9

*The greyed out characters are obsolate in use, and may rarely present inLDC-IL corpus.

18.5 OVERVIEW OF REPRESENTED DOMAINS

LDC-IL Telugu Text Corpus size is: 3,010,993 Words drawn from 737 different titles, including the extracts from newspapers. The data can be categorized into two classes of typed+cleaned and crawled. The crawled data has been crawled mainly from news websites and archived using the standard processing of LDC-IL text corpus preparation.

The following table gives a summary of the typed and crawled text of the Telugu Raw Text Corpus.

Text Type	Word Count	Keystroke/Character Count
Typed+Cleaned	2982155	24668907
Crawled	28838	245914
Total	3010993	24914821

Table 18-1 Representation of the typed and crawled text in Telugu Text Corpus

The representation of the six major domains covered has been shown in the table below:

Domain	Word Count	Percentage
Aesthetics	1,687,968	56.06%
Commerce	45,130	1.50%
Official Documents	6,708	0.22%
Social Sciences	841,429	27.95%
Mass Media	14,656	0.49%
Science & Technology	415,102	13.79%
Total	**3,010,993**	**100**

Table 18-2 Representation of the Domains in Telugu Text Corpus

As each domain has several sub-domains, the following table shows the representation of the several domains, both within the domain and across all the domains.

18.5.1 Aesthetics

The Aesthetics category of Telugu text corpus covers 23 sub-categories bearing a total 1,687,968 words along with the overall percentage of 56.06%. The representational details are given in the table below.

Subdomain	Word Count	Percentage (within Subdomain).	Overall Percentage
Autobiographies	66209	3.92%	2.20%
Biographies	146962	8.71%	4.88%
Culture	30562	1.81%	1.02%
Fine Arts-Dance	48284	2.86%	1.60%
Fine Arts-Music	22377	1.33%	0.74%
Fine Arts-Sculpture	30848	1.83%	1.02%
Folklore	116784	6.92%	3.88%
Fine Arts-Handicrafts	2526	0.15%	0.08%
Humour	9938	0.59%	0.33%
Mythology	96303	5.71%	3.20%
Literature-Children's Literature	20,601	1.22%	0.68%
Literature-Criticism	96601	5.72%	3.21%

Literature-Epics	5507	0.33%	0.18%
Literature-Essays	241217	14.29%	8.01%
Literature-Novels	199920	11.84%	6.64%
Literature-Plays	114999	6.81%	3.82%
Literature-Poetry	14179	0.84%	0.47%
Literature-Science Fiction	8244	0.49%	0.27%
Literature-Short Stories	375227	22.23%	12.46%
Literature-Speeches	23761	1.41%	0.79%
Literature-Text Books (School)	9851	0.58%	0.33%
Literature-Travelogues	4927	0.29%	0.16%
Photography	2141	0.13%	0.07%
Total	1687968	100%	56.06%

Table 18-3 Aesthetics Category Representation

18.5.2 Commerce

The Commerce category of Telugu text corpus covers 5 sub-categories bearing a total 45,130 words along with the overall percentage of 1.50%. The representational details are given in the table below.

Subdomain	Word Count	Percentage (within Subdomain).	Overall Percentage
Accountancy	9046	20.04%	0.30%
Banking	5888	13.05%	0.20%
Career and Employment	528	1.17%	0.02%
Finance	6181	13.70%	0.21%
Industry	23487	52.04%	0.78%
Total	45130	100%	1.50%

Table 18-4 Commerce Category Representation

18.5.3 Mass Media

The Mass Media of Telugu text corpus covers 5 sub-categories bearing total 14656 words along with the overall percentage of 0.49%. The representational details are given in the table below.

Subdomain	Word Count	Percentage (within Subdomain).	Overall Percentage
Article	2728	18.61%	0.09%
Classifieds	858	5.85%	0.03%
General News	578	3.94%	0.02%
Political	9981	68.10%	0.33%
Social	511	3.49%	0.02%
Total	14656	100%	0.49%

Table 18-5 Mass Media Category Representation

18.5.4 Official Documents

The Official Documents category of Telugu text corpus covers 2 sub-categories bearing total 6708 words along with the overall percentage of 0.22%. The representational details are given in the table below.

Subdomain	Word Count	Percentage (within Subdomain).	Overall Percentage

Legislature	3797	56.60%	0.13%
Parliamentary/Assembly Debates	2911	43.40%	0.10%
Total	6708	100%	0.22%

Table 18-6 Official Documents Category Representation

18.5.5 Science and Technology

The Science and Technology of Telugu text corpus covers 13 sub-categories bearing total 415102 words along with the overall percentage of 13.79%. The representational details are given in the table below.

Subdomain	Word Count	Percentage (within Subdomain).	Overall Percentage
Astrology	18747	4.52%	0.62%
Ayurveda	3141	0.76%	0.10%
Biology	4206	1.01%	0.14%
Botany	2265	0.55%	0.08%
Film Technology	35978	8.67%	1.19%
Geology	56416	13.59%	1.87%
Homeopathy	2743	0.66%	0.09%
Medicine	121255	29.21%	4.03%
Psychology	8521	2.05%	0.28%
Text Book (Science)	22002	5.30%	0.73%
Textile Technology	9986	2.41%	0.33%
Yoga	4682	1.13%	0.16%
Zoology	125160	30.15%	4.16%
Total	415102	100%	13.79%

Table 18-7 Science and Technology Category Representation

18.5.6 Social Sciences

The Social Sciences category of Telugu text corpus covers 19 sub-categories bearing total 841429 words along with the overall percentage of 27.95%. The representational details are given in the table below.

Subdomain	Word Count	Percentage (within Subdomain).	Overall Percentage
Anthropology	8757	1.04%	0.29%
Archaeology	1423	0.17%	0.05%
Demography	18776	2.23%	0.62%
Economics	2012	0.24%	0.07%
Education	622	0.07%	0.02%
Fisheries	826	0.10%	0.03%
Geography	19835	2.36%	0.66%
Health and Family Welfare	41859	4.97%	1.39%
History	17173	2.04%	0.57%
Journalism	109570	13.02%	3.64%
Law	32877	3.91%	1.09%
Library Science	22901	2.72%	0.76%
Linguistics	154262	18.33%	5.12%
Philosophy	169724	20.17%	5.64%
Political Science	102312	12.16%	3.40%

Public Administration	30511	3.63%	1.01%
Religion/Spiritual	87470	10.40%	2.91%
Sociology	12767	1.52%	0.42%
Sports	7752	0.92%	0.26%
Total	841429	100%	27.95%

Table 18-8 Social Science Category Representation

18.6 COPYRIGHT CONSENTS

The Telugu text corpus has been collected from various sources and the copyright for the same stays with different sources. However, for the purposes of this corpus, consent have not yet received 62% letter sent for copyright concern. Most of the copyrights (around 91%) belong to private parties with only 8% belonging to the government agencies, either state or the central.

19 URDU RAW TEXT CORPUS

Mansoor Khan, Shahnawaz Alam, Bi. Bi. Mariyam, Rajesha N, Manasa G, Narayan Choudhary, L. Ramamoorthy

19.1 INTRODUCTION

Urdu is a significant language of the Indian sub-continent. Region-wise, Urdu language co-existed along side other languages, in the north, north-west and parts of eastern India, though understood and spoken occasionally in the rest of India also.

The name Urdu was first used by the poet Ghulam Hamadani Mushafi around 1780. From the 13th century until the end of the 18th century Urdu was commonly known as Hindi. The language was also known by various other names such as Hindavi and Dehlavi. Hindustani in Persian script was used by Muslims and Hindus, but was current chiefly in Muslim influenced society. The communal nature of the language lasted until it replaced Persian as the official language in 1837 and was made co-official, along with English. Hindustani was promoted in British India by British policies to counter the previous emphasis on Persian. This triggered a Hindu backlash in northwestern India, which argued that the language should be written in the native Devanagari script. This literary standard called "Hindi" replaced Urdu as the official language of Bihar in 1881, establishing a sectarian divide of "Urdu" for Muslims and "Hindi" for Hindus, a divide that was formalized with the division of India and Pakistan after independence (though there are Hindu poets who continue to write in Urdu to this day, with post-independence examples including Gopi Chand Narang and Gulzar).

The Muslims of North India were as indifferent as the Hindus to the cultivation of Khaři Boli in the medieval period. Although there are distinguished Muslim contributions to e.g. Awadhi, Braj or Punjabi poetry, the huge bulk of medieval Indo-Muslim literature is written in Persian.

By another of those paradoxical developments in which the history of Urdu so abound, the first substantial tradition of writing in Urdu was founded not in the North but in the Deccan, where the lingua franca of Delhi had been introduced into the quite alien linguistic territory of Telugu and the other Dravidian languages by the Muslim invasions of the 13th century. In the 16th and 17th centuries, the Deccan was divided among several powerful Muslim kingdoms, and the rulers of Bijapur and Golkunda (modern Hyderabad) in particular were notable poets and patrons not only of Persian but also of the archaic local variety of Urdu known as ‘Dakani. Motivated in part doubtless by a wish to assert their separate identity from the Mughal Empire which was to absorb them by 1687, these courts produced considerable quantities of Dakani verse, although their archaic language effectively separates them from the mainstream of classical Urdu.

Urdu is written in Perso-Arabic script. LDC-IL Urdu text corpus is collected in Perso-Arabic script of contemporary usage.

Urdu text corpus is collected from various libraries in Uttar Pradesh mostly from Aligarh Muslim University. The greater part of the text has been taken from Maulana Azad Library, Aligarh Muslim University, Aligarh, and Delhi University Campus library. LDC-IL tried to cover the entire category in its standard list. Some categories like novel, short stories have huge amount of books but some categories like physics, chemistry, economics have very less amount of books. Literary texts are easily available in Urdu but getting scientific text is difficult. Some categories like epigraphy, finance, oceanology text are too rare in Urdu.

19.2 PECULIARITIES OF URDU TEXT

The Corpus of Urdu text can be broadly classified into two: literary text and non-literary text. These two explicitly shows their differences in terms of frequency of word usage and variety that it brings into corpus. Literary texts are texts that are narrative and it contains elements of fiction. Novels, short stories, plays are examples of literary text. Non-literary texts are texts 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.

19.2.1 The writing system used in the language

The writing system Urdu has is the right-to-left alphabet, a modification of the Persian alphabet known as Perso-Arabic, which is, so to claim, itself a derivative of the Arabic alphabet. The Urdu alphabet has 58 letters. Urdu has its nomenclature for the 'type of script', i.e., ABJAD. And, the script is consonantal in large amount; whereas, the vowel information is represented by combining marks that appear above or below the base consonant. It can be noted that the consonants are largely reliable phonetically despite its difficulty for the vowel sound to realize. On the other hand, we can say that there is mostly a one-to-one correspondence between letters and sounds. Urdu is only one of an Indian language, which is written from right to left.

19.3 DATA SAMPLING NOTES

19.3.1 Principles of Data Sampling

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

19.3.2 Fieldworks Undertaken

Urdu text corpus is collected from various libraries in Uttar Pradesh, mostly from Aligarh, Delhi and Bhopal. The text materials were collected by conducting three fieldworks undertaken in the period from 2010 to 2012. The greater part of the text has been taken from Aligarh Muslim University Library and other Departmental libraries of Aligarh Muslim University, Aligarh.
Overall, the following libraries served as the source of the Urdu text corpus:

- Maulana Azad Library, Aligarh Muslim University, Aligarh
- Department of Urdu Library, Aligarh Muslim University, Aligarh.
- Department of Linguistics Library, Aligarh Muslim University, Aligarh.
- Department of Islamic Studies Library, Aligarh Muslim University, Aligarh.
- Hakeem Ajmal Khan Tibya College Library, Aligarh Muslim University, Aligarh.
- Delhi University Library, Delhi.
- Jawaharlal University Library, Delhi.
- Jamia Hamdard University Library, Delhi.
- Some other personal Libraries from Professors and other people.

Collected text materials have been published at various places within Aligarh, Delhi and other states of India such as Andhra Pradesh, Madhya Pradesh and Telangana State, as well as other countries such as Bangladesh and Pakistan etc.

An attempt has been made to cover the entire category in its standard list. Some categories like novel, short stories have huge amount of books but some categories like physics, chemistry, economics have very less amount of books. Literary texts are easily available in Urdu but getting scientific text is very difficult. Some categories like epigraphy, finance, oceanology text are too rare in Urdu.

Collecting text data from the field is a difficult job. Most of the libraries do not allow to take 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 to take 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 takes up more time for them to turn the pages compared to continuous page photocopying they are accustomed to. It was another issue that the field worker/linguist had to carry a huge list of photocopy bundles with them which was many times cumbersome to travel with.

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

19.3.3 Data Inputting

All the text has been typed in Unicode using the InScript Keyboard directly onto the XML files. The data has been inputted by Miss. B.B. Mariyam, Miss. Naseerunnisa, Miss. Ayesha Talath and Khadeeja Tasneem, who are native speakers of Urdu.

19.3.4 Validation and Normalization Workshops

A 5-day workshop was conducted at Linguistic Data Consortium for Indian Languages held at August 2-9, 2010 with Dr. Mohammed Zia Ullah, University of Mysore, Dr. Aftab Ahmad Faridi, Aligarh Muslim University, Ms. Paiker Fatima, Aligarh Muslim University, Ms. Sheeba Aziz, Aligarh Muslim University, Aligarh and Ms. Farah Jawed, Aligarh Muslim University, as experts. The experts suggested that the Urdu text corpus should remain true to the text.

19.3.5 Proofreading

Urdu text data has been proofread by internal resource persons and other resource persons who have been called by LDC-IL for short term program for 45 working days. The text has always been kept true to the printed material and typos, if any, occurring at the time of typing have only been corrected.

The printed materials collected for the corpus is contemporary, mainly published after 1990.

19.3.6 Data Extracted from Web Sites

Urdu News corpus data is extracted from News websites of *"Roznama Rashtriya Sahara" (*https://www. roznamasahara.com*), "Qaumi Awaz" (*https://www.qaumiawaz.com/*), "Daily Aaj" (https://epaper.dailyaaj.com.pk/index.htm)* . The news content was categorized based on the content of the text and archived. The period of selection of the news corpus ranges from 2010 to 2012.

19.4 OVERVIEW OF REPRESENTED DOMAINS/CATEGORIES

LDC-IL Urdu Text Corpus size is: 51,61,927 Words and character count is 2,39,47,905 drawn from 739 different titles and 3 titles including the extracts from newspapers. The data can be categorized into two classes of typed+cleaned and crawled. The crawled data has been crawled mainly from news websites and archived using the standard processing of LDC-IL text corpus preparation.

The following table gives a summary of the typed and crawled text of the Urdu Raw Text Corpus.

Text Type	Word Count	Keystroke/Character Count
Typed+Cleaned	4254716	19596417
Crawled	907211	4351488
Total	**51,61,927**	**2,39,47,905**

Table 19-1: Representation of the typed and crawled text of the Urdu Raw Text Corpus

The representation of the five major domains covered has been shown in the table below:

Domain	Word Count	Percentage
Aesthetics	2616382	**50.69%**
Commerce	28601	**0.55%**
Mass Media	843477	**16.34%**
Science and Technology	348082	**6.74%**
Social Sciences	1325385	**25.68%**
Total	**51,61,927**	**100**

Table 19-2: Representation of the Domains in Urdu Text Corpus

As each domain has several sub-domains/sub-categories, the following table shows the representation of the several domains, both within the domain and across all the domains.

19.4.1 Aesthetics

The aesthetic domain of Urdu text corpus covers 21 subdomains bearing a total of 26,16,382 words along with the overall percentage of 50.69%. The representational details are given in the table below.

#	Subdomain	Word Count	% within Subdomain	Overall Percentage
1	Autobiographies	45774	1.75%	0.89%
2	Biographies	279390	10.68%	5.41%
3	Cinema	159160	6.08%	3.08%
4	Culture	3314	0.13%	0.06%
5	Fine Arts-Music	597	0.02%	0.01%
6	Fine Arts-Sculpture	15863	0.61%	0.31%
7	Humour	10834	0.41%	0.21%
8	Literary Texts	26989	1.03%	0.52%
9	Literature-Children's Literature	121395	4.64%	2.35%
10	Literature-Criticism	1098831	42.00%	21.29%
11	Literature-Epics	4136	0.16%	0.08%
12	Literature-Essays	113292	4.33%	2.19%
13	Literature-Letters	17530	0.67%	0.34%

14	Literature-Novels	193691	7.40%	3.75%
15	Literature-Plays	55264	2.11%	1.07%
16	Literature-Poetry	1876	0.07%	0.04%
17	Literature-Science Fiction	20619	0.79%	0.40%
18	Literature-Short Stories	223213	8.53%	4.32%
19	Literature-Speeches	23568	0.90%	0.46%
20	Literature-Text Books (School)	195980	7.49%	3.80%
21	Literature-Travelogues	5066	0.19%	0.10%
	Total	**26,16,382**	**100%**	**50.69%**

Table 19-3: Aesthetics Category Representation

19.4.2 Commerce

The Commerce category of Urdu text corpus covers 4 subdomains bearing a total of 28,601 words along with the overall percentage of 0.55%. The representational details are given in the table below.

#	Subdomain	Word Count	% (within Subdomain)	Overall Percentage
1	Accountancy	3373	11.79%	0.07%
2	Banking	15378	53.77%	0.30%
3	Finance	4025	14.07%	0.08%
4	Industry	5825	20.37%	0.11%
	Total	**28,601**	**100**	**0.55**

Table 19-4: Commerce Category Representation

19.4.3 Mass Media

The Mass Media category of Urdu text corpus covers 6 subdomains bearing a total of 8,43,477 words along with the overall percentage of 16.34%. The representational details are given in the table below.

#	Subdomain	Word Count	% (within Subdomain)	Overall Percentage
1	Classifieds	5143	0.61%	0.10%
2	Editorial	123335	14.62%	2.39%
3	General News	5739	0.68%	0.11%
4	Obituary	2135	0.25%	0.04%
5	Political	511812	60.68%	9.92%
6	Sports News	195313	23.16%	3.78%
	Total	**8,43,477**	**100**	**16.34**

Table 19-5: Mass Media Category Representation

19.4.4 Science and Technology

The Science and Technology category of Urdu text corpus covers 19 subdomains bearing a total of 3,48,082 words along with the overall percentage of 6.74%. The representational details are given in the table below.

#	Subdomain	Word Count	within Subdomain	Overall Percentage
1	Agriculture	12139	3.49%	0.24%
2	Astronomy	5745	1.65%	0.11%

3	Ayurveda	4260	1.22%	0.08%
4	Biology	5641	1.62%	0.11%
5	Chemistry	6541	1.88%	0.13%
6	Computer Sciences	25917	7.45%	0.50%
7	Engineering-Chemical	4905	1.41%	0.10%
8	Forestry	5821	1.67%	0.11%
9	Geology	4908	1.41%	0.10%
10	Homeopathy	17905	5.14%	0.35%
11	Logic	7642	2.20%	0.15%
12	Mathematics	4728	1.36%	0.09%
13	Medicine	6077	1.75%	0.12%
14	Physics	48622	13.97%	0.94%
15	Psychology	62573	17.98%	1.21%
16	Sexology	37555	10.79%	0.73%
17	Statistics	4419	1.27%	0.09%
18	Text Book (Science)	69856	20.07%	1.35%
19	Zoology	12828	3.69%	0.25%
	Total	**3,48,082**	**100.00**	**6.74**

Table 19-6: Science and Technology Category Representation

19.4.5 Social Sciences

The Social Science category of Urdu text corpus covers 19 subdomains bearing a total of 13,25,385 words along with the overall percentage of 25.68%. The representational details are given in the table below.

#	Subdomain	Word Count	% (within Subdomain)	Overall Percentage
1	Archeology	5209	0.39%	0.10%
2	Demography	9292	0.70%	0.18%
3	Economics	72631	5.48%	1.41%
4	Education	288608	21.78%	5.59%
5	Epigraphy	12332	0.93%	0.24%
6	Fisheries	5513	0.42%	0.11%
7	Geography	12497	0.94%	0.24%
8	History	80303	6.06%	1.56%
9	Home Science	4821	0.36%	0.09%
10	Journalism	28503	2.15%	0.55%
11	Law	22921	1.73%	0.44%
12	Library Science	5590	0.42%	0.11%
13	Linguistics	37643	2.84%	0.73%
14	Philosophy	59456	4.49%	1.15%
15	Political Science	35926	2.71%	0.70%
16	Public Administration	9834	0.74%	0.19%
17	Religion/Spiritual	558725	42.16%	10.82%
18	Sociology	36343	2.74%	0.70%
19	Text Book (Social Science)	39238	2.96%	0.76%
	Total	**13,25,385**	**100.00%**	**25.68%**

Table 19-7: Social Science Category Representation

19.5 COPYRIGHT CONSENTS

The Urdu text corpus has been collected from various sources and the copyright for the same stays with different sources. However, for the purposes of this corpus, consent has been sought from all the stakeholders. Most of the copyrights (around 65%) belong to private parties with only 35% belonging to the government agencies, either state or the central.

20 LDC-IL RAW SPEECH CORPORA: AN OVERVIEW

Narayan Choudhary, Rajesha N, Manasa G, L. Ramamoorthy

20.1 INTRODUCTION

Lack of basic linguistic resources have been the first and foremost bottleneck in development of language technology for Indian languages. When text data itself has been available for most of the Indian languages, one could not even think of the speech data. India is one of the foremost multilingual country where multilingualism is ingrained and most people speak more than one language with more than 75 languages having more than one million speakers (as per 2011 Census of India data). As per a study[3] of KPMG and Google released in 2017, the internet user base grew at a compound annual growth rate (CAGR) of 41% between 2011 and 2016 to reach 234 million users at the end of 2016 and this trend is likely continue. It is also estimated that internet users in Indian language will account for nearly 75% of India's internet user base by 2021.

Despite this, the availability of technology in Indian languages have been on close to null. This is mainly due to the reason that the technology developing agencies find it either too difficult to come up with the language support on various applications for Indian languages or it is economically not a viable solution. However, recent analyses from various quarters have shown that the latter is not correct and the major issue is availability of the linguistic resources based on which language technology and language support for various types of applications proves to be a bottleneck for the developing community, be it industry or otherwise.

Considering this as an issue, the Government of India has taken several initiatives to provide the basic ingredients which may prove as a catalyst for the development of language technology in Indian languages. As part of the this initiative, a scheme named Linguistic Data Consortium for Indian Languages (LDC-IL) was established by the Ministry of Human Resource and Development at Central Institute of Indian Languages, Mysore.

The goal of LDC-IL was to develop linguistic resources for all Indian languages with the initial focus more on the scheduled languages of India. These linguistic resources may be as deemed fit by the language technology developing community.

Based upon the recommendations of the Project Advisory Committee which includes ex-officio members from MeitY, IITs Ministry of HRD, Director and other academicians from

[3] https://assets.kpmg.com/content/dam/kpmg/in/pdf/2017/04/Indian-languages-Defining-Indias-Internet.pdf

reputed Institutes/Universities working in this area as well as major and minor industrial entitites working in this area, the LDC-IL decided to embark upon developing the text and speech corpus for the scheduled languages of India.

There have been several types of datasets prepared under LDC-IL. This document serves as a generic documentation for the raw speech corpus of the LDC-IL being released for several languages.

20.2 LDC-IL SPEECH CORPUS

LDC-IL speech corpus is collected after careful deliberations on what type of speech corpus is required for various types of speech based linguistic analysis that may suit multifarious needs of the research and development community.

After several meetings with the experts from around India and abroad, it was decided that LDC-IL should focus on not just developing a speech corpus for a particular need, rather to get the data that would be useful for various tasks such as ASR, STT, linguistic analysis, speech therapy and so on.

Keeping this in mind, various types of content were created *a priori* before the speech recordings took place. The content of these datasets have been prepared in consultation with the experts from the language as well as linguists giving inputs to ensure that no specific sound patterns are missed out.

For example, it has been ensured that the speech datasets contain all the phones and allophones of the language and ample examples are available in the language to prove their phonemic status in the language. To ensure that the corpus is good for an ASR, it is ensured that the continuous speech is present in natural environment.

20.3 CONTENT TYPE DESCRIPTIONS

Each content type has a number of files with each file containing standard content. A sub-set of these files in each of the content types selected randomly constitute a subset that are given to a speaker for reading out in natural flow. A few full sets (namely W3B, W4, and W5) are also read in full by certain selected speakers in each age group.

There are three age group ranges selected for LDC-IL datasets. These are ‘16 to 20 years’, ‘21 to 50 years’ and ‘above 50 years’. Attempt has been made to collect equal number of male and female data from each of the age groups.

The list of the datasets and their notation is given in the table below:

SL	Notation	Conent Type
01.	T1	Contemporary Text (News)

02.	T2	Creative Text
03.	S	Sentence
04.	D	Date
05.	W1	Command and Control Words
06.	W2	Place Name
07.	W2	Person Name
08.	W3A	Most Frequent Word-Part
09.	W3B	Most Frequent Word-FullSet
10.	W4	Phonetically Balanced-Fullset
11.	W5	Form and Function Word-Fullset

Table 20-1: LDC-IL Speech Data Content Types

Detailed descriptions of each of the content types are given in the following sub-sections.

20.3.1 T1: Contemporary Text

The Contemporary Text (news) data is given the notation of T1. News items have been selected from the LDC-IL news corpora. The text is contemporary in nature as the news items such have been picked over a period from 2005 to 2012, either from news websites or from the print editions newspapers of the respective language.

The domain information is present in the news items as well as the news items deal various topics such as political news, editorials, sports news and so on. Given that the news items have been collected from local news reported for each language, the style may be considered as colloquial or belonging to the news reporting style.

Each LDC-IL dataset 'Contemporary Text 'contains minimum of 500 words per speaker, which is rarely repeated. Since it is the continuous text, it constitutes the largest part of the speech corpora, in terms of data size and time duration.

20.3.2 T2: Creative Text

'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. The text may be any standard text which is descriptive in nature. It exhibits the language style of a particular period from which the text is taken.

Creative text were prepared in two types. In the first 6 or 8 esaasyes or shortstories were prepared and randomly One of these selected randomly from the set, is assigned to one speaker for reading out. The same story may be read by multiple speakers.

In the other approach a distinct text is given to each individual

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

20.3.3 D: Date

Languages tend to speak out the date in a specific and many a times in a particular manner which may not always conform to the grammatical structure of the language. To capture it, LDC-IL tried to document how a date is spoken in each of the languages.

The normal way is put a question before the informat the answer of which must be in a date format. Normally the following six questions were placed before the informant and the informants would answer minimum one of the questions.

1. What is tomorrow's date?
2. When is Gandhi Jayanthi observed?
3. What is the date today?
4. When do we celebrate our Independence Day?
5. What is your date of birth?
6. On which date you go to market?

20.3.4 S: Sentences

To ensure that all the types of syntactic structures are covered in the speech data, a set of sentences have been constructed with the help of language experts and linguists for each of the languages. It is ensured that all possible sentence structures are covered including all types of tenses, aspects, moods, compound and complex sentences and so on.

These sentences are isolated sentences and not part of a continuous speech. While care has been taken to extract sentences from the text corpus of the corresponding language, sometimes sentences have also been modified to ensure that the specific valid sentence structure of the language is present.

Very long sentences are avoided while selecting or constructing the sentences, so that the informant can read the sentences easily. The words used in these sentences are common words which are found in day-to-day life. Each sentence in the list contains minimum four words. The sentences are not too long so that each sentence does not span for more than a line in the prompting sheet. Care is taken to avoid abusive or taboo words.

Each speaker is given 25 sentences out of this sentence list for reading out.

20.3.5 W1: Command and Control Words

Spoken language usually contains a lot of sentences that are commands or use a lot of control words. This happens mostly in the conversational speech. Even though the LDC-IL speech corpus at present does not contain the conversation speech, an attempt has been made by including common command and control sentences/phrases carefully crafted with the help of respective language experts and linguists.

These include imperative sentences, optative sentences as well as other controlling phrases which may come as a reply to an interrogative sentence. Each of the languages have a set of command and control sentences created before the speech data is recorded. Each speaker is given

a list of 30 command and control sentences randomly selected from the set. Each of these phrases/sentences are repeated three times by each speaker while recording.

20.3.6 W2: Proper Noun (Person Names and Place Names)

Recognizing proper nouns by using an ASR system is a complex task. For example, voice recognition application in mobile phone may have a few hundreds of names to distinguish when placing a call through voice command. Native speakers use different pronunciations depending on their language of origin and familiarity with the language. The speakers use different pronunciation for native and foreign names ranging from a nativised pronunciation to a totally foreignised pronunciation. All this adds to the complexity of an ASR system in recognizing proper nouns. To address this issue LDC-IL speech data has been collected to include person names and place names.

20.3.6.1 Person Names

Person names are included to capture the native pronunciations. The names are taken from people from different walks of life like Politicians, Film Actors and Directors, Writers, Kings and Queens, Astrologers, Historical Personalities, Scientists, Sports persons etc.

20.3.6.2 Place Names

Place names are included to capture the native pronunciations. This data set contains Indian place names. These include main cities, district names and popular tourist destinations from all over India. Some local place names are also included like names of villages, taluk headquarters, district names, local forest reserves, local tourist and pilgrimage destinations etc.

Each speaker typically pronounce 20 person names and 10 place names, out of the total Proper Noun wordlist of the particular language. Each word is uttered three times

20.3.7 W3: Most Frequent Words

Most frequent word list is the regularly and repeatedly used list of words. Since these words are used most frequently in a language, it is imperative to have these words in our dataset.

The most frequent words dataset is derived from LDC-IL Corpus. However, it may be noted that when the most frequest word list was extracted, the text corpus was rather small. So, the frequency list might change if it is compared to the current LDC-IL text corpus.

20.3.7.1 Most Frequent Words Part-W3A

The most frequent words of a language are randomly picked from a list of around 1000 most frequent wordlist of a language. Each speaker records randomly selected 30 words from this list. Each word is uttered thrice.

20.3.7.2 Most Frequent Words Part-W3B

Two speakers, one male and one female, pronounces the full set of 1000 most frequent words. This is done for each dialect of the language, if available.

20.3.8 W4: Phonetically Balanced Vocabulary

To cover all possible phonemic occurrences of a language, the "phonetically balanced Vocabulary" is prepared. It is a list of words in which the occurrence of a phoneme in initial medial and final positions of that language can be represented.

The pronunciation of the phoneme is varied according to the position of the phoneme in a word and the influence of the following and proceeding phoneme. The pronunciation of initial position is different from middle and final positions. For example the phoneme '**pa**' is used in different forms while pronouncing words like

- '**pa**llaʊi'- 'pa' inherent vowel at initial position (CV initial)
- '**pra**kaṭa' - 'p' as pure consonant in conjunction with 'ra' in initial position, (CCV Initial)
- '**spa**ndana',- 'pa' with inherent vowel preceded by a consonant at medial position (CCV Initial)
- '**pa**rika**lpa**ne'- 'pa' inherent vowel at initial position (CV initial) and 'pa' with inherent vowel preceded by a consonant in the medial position (CCV Medial)
- 'aː**pta**' - 'p' with followed by a consonant in the final position (CCV medial)

Using the articulation score as the measure, *phonetically balanced lists have* been used to test differences among transmission systems and to test the effects of noise. The phonetically balanced words used in word recognition testing contain speech sounds that occur in the same frequency as those of conversational speech.

20.3.9 W5: Form and Function words

Form and function words dataset is a closed class list of words. They are quite limited in a language. These constitute mostly the indeclinable words of a language. Form words are static, bearing some content with them. They are meaningful and are actually the building blocks of a language.

The Form and Function dataset includes Grammatical function words, numerals, kinship terms, measurement terms, list of colours, days, months, seasons, directions, zodiac sings, body parts, planets etc. These words are included to the native words which may not be frequent in the overall corpus, but needs representation.

20.4 PLANNING FOR FIELDWORK

20.4.1 Dataset preparation and distribution

To ensure representativeness of the speech corpora, a conscious effort has been made to balance the speech data by taking varieties of styles into consideration. The first and foremost among at LDC-IL has been to take an expert view on the varieties of languages. For example, for Kannada it is ensured that speech varities from different regions such as Hyderabad Karnataka, Bombay Karnataka, Coastal Karnataka and Old Mysore get proportionate weightage.

LDC-IL collected the data using two approaches, with two different kind of Dataset Models They are as follows

- Dataset Model 1 (T1, T2, W1, W2, W3, W4, W5, S, D)
- Dataset Model 2 (Distinct Texts of T1 and T2)

Some Languages followed Model-1 only, and some Languages followed both Model-1 and Model-2

After the regions are identified, speech samples are collected as per the criteria shown in the table below:

Standard Speech Dataset Distribution for Each LDC-IL Fieldwork Dataset Model 1							
Content type	**Content size#**	**Content to be read by one speaker**	**Total No. of speakers**	**Age group wise no. of speaker; Female & Male equally distributed#**			**Content selection type**
				16-20	20-50	50+	
Contemporary Text	150 Texts	1 Text	150	18	90	42	Distinct Text
Creative Text	6 Texts	1 text	150	18	90	42	Random set*
Sentences	142 Sentences	25 Sentences	150	18	90	42	Random set*
Command and Control Words	82 Words	30 Words	150	18	90	42	Random set*
Person Names	489 Words	20 Words	150	18	90	42	Random set*
Place Names	511 Words	10 Words	150	18	90	42	Random set*
Most Frequent Words	1144 Words	30 Words	150	18	90	42	Random set*
Phonetically Balanced Vocabulary	390 words	Full set	6	2	2	2	Full set to be read by the speaker
Form and Function Words	432 words	Full set	6	2	2	2	Full set to be read by the speaker
1000 Most Frequent Words	1000 Words	Full set	2	0	2	0	Full set to be read by the speaker

*picked randomly by machine
#The figures shown are for illustration purpose only. The numbers may differ for each langauge. Plese reffer specific Langauge documentation for actual figures.

Table 20-2: Standard Speech Dataset Distribution for Each LDC-IL Fieldwork with Modle-1 Dataset

Speech dataset distribution for fieldwork Dataset Model 2						
Content type	**Content size**	**Content to be read by one speaker**	**Total No. of speakers**	**Age group wise no. of speaker; Female & Male equally distributed**		**Content selection type**
				16-20	21-50	
Contemporary Text (News) Text	150 Texts	1 Text	150	75	75	Distinct Text
Created Text	150 Texts	1 text	150	75	75	Distinct Text

Table 20-3: Standard Speech Dataset Distribution for Each LDC-IL Fieldwork with Modle-2 Dataset

As the data is collected from different cities across India (as per the demand of the language), it's imperative that proper preparation is made before proceeding towards the field such that day-to-day necessities of field are met with. Investigators had to make that s/he had sufficient charged batteries as well as alkaline batteries if so required, empty SD cards, laptops in proper condition, sufficient number of random and full datasets (prompt sheets) and so on. These formed as the daily routine for the linguists while in the field.

20.5 FIELD WORK

Some common guidelines and instructions were provided to the field workers before proceeding to the field. A brief of it is noted below.

20.5.1 Possible places for collecting data

Once the dataset is prepared and taken to the field, the next step is to determine places where there is an availability of speakers who can read fluently. The best possible places are schools, colleges, universities, govt. offices etc.

The Head of these organizations have to be briefed and asked permission for recording data from students, faculties etc. Certain infrastructural requirements like space, if possible power source for charging batteries etc. has to be requested from the institutions. The speakers from whom we collect data are referred as informants.

20.5.2 Field work Ethics

The informants are briefed about the procedures, nature and purpose of speech data collection. Informants are informed about the funding agency behind the data collection. In case of LDC-IL, the data collection is funded by Govt. of India. Informant are made aware of who exactly is carrying out the data collection process and what will be done with the data collected before they give their consent.

There have been situations where the informant would find it distressing that the data given by them will be segmented and further processed. In such cases, their opinions have to be respected and the investigators have to refrain from taking their data. The informants are made aware of the degree of confidentiality and anonymity that will be maintained after collecting the data. The informant are also made aware of the potential benefits of the data to the wider community. Once the informant is aware of all these information and is ready to give the data, consent is acquired in written along with certain personal details such as their educational qualification, mother tongue, place of elementary education etc.

Informants are allowed to read the dataset earlier before recording so that they can get familiar with the content of the text. It is ensured that the informants do not have any objection to the content they are about to read. For example, the informant may have objection regarding the political, social views expressed in the content. In such cases, a different dataset is offered to the informant. There are certain texts in the data set, which may pose difficulty for a certain informant to read. For Example, some informants may have difficulty in reading contents which involve dialogues between people. Such contents may differ in dialects spoken by the informant which may pose a difficult situation for them while reading. In such cases, a different dataset is offered to the informant. Complex datasets are given only to the informants who are capable of reading them and state likewise.

An attempt is made to reduce the extra noise as much as possible before recording. If necessary, test recordings are conducted before the actual recording on certain portions of the text.

Brief introduction about the informant and investigator along with details like place, time, region etc. are collected at the beginning of each recording. The conversation between investigator and informant is done in their native language so that the informant is comfortable and the natural flow of language is established.

Care is taken while recording the words, so that there is a pause between two words or between utterances of the same word. All the words of the content type W1to W5 (i.e. ‘Command and Control words’, ‘Proper Nouns’, ‘Most Frequent Words’, ‘Phonetically Balanced Vocabulary’ and ‘Form and Function words’) are repeated three times in a sequence. A pause is maintained between two sentences as well while recording.

While recording the News Item and Creative Text, the informants are briefed to read the text given, as naturally as possible. It should be as natural as reading a book or newspaper.

Informants answer to a particular question themselves regarding date format. This is done to capture how people usually pronounce the date. The investigator does not prompt any particular format.

20.6 DATA COLLECTION

The LDC-IL data is recorded using Roland EDIROL Recorder. It is a 24-bit Linear PCM (R-09) Recorder.

20.6.1 Technical Specifications for collecting data

Recording Setup:	Sample Rate : 48.0 KH
Recording Mode:	wav -16bit
Date Setup:	Current date and time.
Storage:	SD Card
Power:	•Always use rechargeable batteries (Ni-MH) for recording. Otherwise line hum will come. Never use Ni-CD battery type as it is potential for 'memory effect'. • Rechargeable batteries need to be thoroughly recharged before recording (minimum 16 hrs continuous charging).
Peak	While recording please be aware that it should not reach the peak i.e. PEAK (in the recorder) should not glow.
Recording Distance	•Keep minimum 5 cm to 25 cm distance between the microphone and the speaker and if possible use microphone holder. • The recorder should not be placed orthogonally but it should be placed diagonally. • Do not move the recorder during recording • Fix the recorder upon a table/ fixed plane if possible. • Try to have fixed the distance between the recorder and speaker • The recorder should not be placed orthogonally but it should be placed diagonally

After each recording, it is recommended to verify the recorded data, whether it is recorded in the right way. If the informant also wishes to hear the data, the investigator may oblige.

20.6.2 Metadata

The value of speech data can be determined according to the quality of metadata obtained. It is imperative to maintain metadata of the entire data collection for linguistic analysis.

After the recording is taken from the informant, personal details are collected. Care should be taken so that the signature and other formalities are completed as required.

The metadata of the speech corpus is made through the personal details taken from the informants. A typical copy of metadata sheet contains information as noted below:

Informant Data:

Name:
Dataset ID:
Address:
Gender:
Age Group: (with three options of 16 to 20, 21 to 50, and50+)
Educational Qualification: (with three options of School/Bachelors/Masters)
Place of Elementary Education:
Mother Tongue:
Dialect (if any):

Investigator Data:

Name:
Date:
Place:
Region:
Environment:

It is to note that the name and the address of the informants are discarded while archiving metadata to keep the confidentiality and anonymity.

Dataset ID: It is a unique ID given to each speaker.

The following fields are considered for the distinctiveness of each data item recorded. Each field contributes certain features which pave way for diverse research.

Gender: The Speech data is taken from both male and female to capture the difference in intensity and pitch. The difference in vocal folds size between men and women makes them different in their pitched voices. Male voice usually has low pitch whereas a female voice is of high pitch. Pitch and intensity are proportional to each other.

Age Group: Different age groups exhibit difference in pitch and loudness. As the human body ages, it undergoes changes such as lessening strength, slower movements, degeneration of body tissues etc. these factors impact the voice as well. As people age their speech slows down, syllables and words are elongated, sentences are punctuated with more pauses for air. Scientific studies also show that as male and female age, the changing larynxes changes pitch and intensity. Age also affect the hearing process, which may make a person speak louder.

Educational Qualification: This determines the fluency and speed of reading speech data.

Place of Elementary Education: This parameter determines the effects of environment and dialect of a particular place of childhood which impacts the articulation of the speech.

Mother Tongue: Mother Tongue is one of the influential factors of a native speaker, for example In Karnataka, mainly in Canara region; it can be observed that the mother tongue of native Kannada speakers may be Tulu, Konkani, Chitpavani etc. This influences the articulation of Kannada speech in these areas.

Place: Place gives better information about the speech data collected. For example, Kannada spoken in Kundapura has its own distinct variety even when it belongs to Canara region.

Date: Date describes the timeline of data collected. It becomes useful information for historic research and language evolution in time line. It also dates the technology being used in that age.

Region: Region is an influencing factor of the language. Hence keeping the information about it with the data is always useful.

Environment: The recording environment information's like Indoor, Outdoor, School, Office, etc is useful, and its marking could be helpful in determining the noise level and the kind of noise that can be expected with the data.

Each of the datasets released contain a metadata sheet which has information about each of the audio files. A total of 25 fields are there in the metadata sheet. A brief of each of these 25 fields/legends is given in the table below:

SL	Legend	Description
1	Langauge	Name of the Language
2	SpeakerID	Each speaker has a unique identity languages. 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	Notes gender, whether it is male, female or other.
6	AgeGroup	Three age groups of 16 to 20, 21 to 50, and50+
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 and linguists.
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's mostly 48.
14	BitPerSample	Bit per sample. It is mostly 16-bit.
15	Channel	How many channels. All of LDC-IL data is sterio. 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, as long as 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 happens 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 4: Metadata Legends and their Description

20.6.3 Data Transferring and Storing

After the data is collected for the day or when the SD card is full, the data needs to be transferred to the PC. It is important, to take certain precautions in this process so that the data is safely transferred. The data should be copied and pasted in the PC rather than cut and pasted. After successful transfer and rechecking the copied data, the SD card can be cleared.

For easier maintenance and organization of the data in PC, folder system is recommended for saving data. Each recorded wave file has to be labeled with corresponding speaker ID.

The investigator should try to get the required number of speakers/data before completing the field work within their schedule.

20.6.4 Observations

One of the reasons for error prone reading could be the over consciousness of the informant about being voice recorded. Despite being informed, the informant may try to read the

data in a dramatic way, but may eventually lead to normal reading after few sentences. Even after the informants give consent and their data, they may later change their mind or express their concern about the text they have read. Some may even request to discard their recordings. In such cases, the investigator has to reassure them about their given data. If they still want their data to be discarded, they have to be accommodated. It is preferable to provide complete information to the informants to avoid such situations. In many instances informants assume that they are giving auditions for Radio Jockey vacancies, or some reality shows. They should be briefed about the purpose of data collection beforehand to avoid such situations.

The investigator may be in not so hospitable environments depending upon the region they are visiting. Proper precaution and aid is to be acquired before visiting such places.

The investigator may have to face challenges in food and accommodation since he/she travels in unfamiliar places. It is recommended to be prepared for such situations. The investigator should be prepared for all such hardships and take proper measures to minimize them beforehand.

20.7 ORGANISING AND ARCHIVING THE DATA

After the field work is completed, the data has to be stored in a server as soon as possible for safe keeping. Taking a backup of the saved data is also recommended as the data collected is of vital importance.

20.7.1 Text - Speech Mapping and Naming Conventions

After the data is stored, it 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.

A Typical LDC-IL naming convention for recorded data is shown bellow.

“LDC-IL_Scheduled_Kannada_Female_16To20_News-T1_SP-0035_T1-0035.wav”
“LDC-IL_Scheduled_Kannada_Male_21To50_Sentence-S_SP-0001_S-0004.wav”

WaveSurfer, a free software, is used for segmentation of LDC-IL Speech data. It is an open source tool which can be downloaded freely from the web. While segmenting the speech data file for archiving, the introduction, content headings and any unnecessary speech are discarded. Only the dataset content is retained.

The ASR data is prepared keeping in the view, the stochastic systems such as HMMs or neural networks that do not use explicit rules for speech recognition. On the contrary, they rely

on stochastic models which are trained using some statistical optimization procedure, with very large amounts of speech corpus.

20.7.2 Observations

While segmenting a single large recording containing all the content types, there may be instances where an informant has made an error and later corrected it. In such cases, it is always a good approach to segment a recording from the end of the file in reverse order so that the correct utterance can be found before incorrect utterance; hence the incorrect utterance can be discarded/ignored. One may observe the interventions of investigator or other people between reading two data items which may also need to be discarded.

20.8 DATA VERIFICATION AND QUALLITY CONTROL

Since maping audio recordings with its corresponding text and other metadata information is a manual task. The process is prone to human errors, the data verifiction process will be done

Much of the audio text mapping is automated, but in case of distict set texts, and other metadata entry is done by human needs verification. In this,

The Audio recording of each speaker is checked against the mapped text.

Each distict text audo recording will be matched with automated entries of the same speaker to check for any mismapping of speaker.

Metadata like Gender, age group, District etc are selective part of manual data entry and could be prone to errors so verication is needed.

Metadata like Dialect entry, place, etc are keyed in by manual data entry and could be prone to errroes like typo errors so verification needs to be done.

The audio segements could be duplicated because of system/network errors and need to be checked.

At the time of data segmentation, one might have saved the whole file instead of selected part. Such casesse needs to be checked.

Some audio segements may not get migrated to the system because of wrong naming convestions. Such segements will be handpicked and corrected and migrated into the system.

21 BENGALI RAW SPEECH CORPUS

Sonali Sutradhar, Rajesha N, Manasa G, Narayan Choudhary, L. Ramamoorthy

21.1 INTRODUCTION

Bengali is the official language of West Bengal and Tripura. It belongs to the Indo-Aryan language family. Bengali is influenced by Sanskrit. After Independence, the state governments of West Bengal and Tripura started using regional languages more and more in administration. Greater use of Bengali has contributed to the growth of the language in terms of vocabulary and the number of styles and registers.

Bengali is spoken over the whole of West Bengal, Tripura and Bangladesh and in some parts of Bihar, Orissa and Assam. Bengali refugees who have settled in Andaman after 1950, have also carried the language there. Speech forms of two adjoining areas are mutually understandable, but speech forms of two remote areas (speech form of Kolkata and the speech form of Chittagong,the eastern most part of Bangladesh) are so different from each other that they are practically unintelligible mutually.

Language is the collection of more or less similar idiolects. The fundamental fact about language is its diversity. Change in language is found when we move from country to country, region to region, class to class and caste to caste. Bloomfield (1933) says that linguistic diversity is related to the density of communication or to the amount of verbal interaction among speakers. In India dialect studies in a broad sense have been initiated by G.A. Grierson, who collected evidences to understand the linguistic situation in India and to group the regional dialects into families of Language such as the Austric, Tibeto - Chinese, lndo European and Dravidian.

Dialect variation in a language is not random but systematic. There are two types of dialects; regional dialects and social dialects. Regional dialects are geographically based and social dialects originate among social groups, class ethnicity, religion etc.

Language variation reflects the language change over time and people who live in the same geographical area or maintain the same social identity share the language norms. Language change happens through three parameters like spatial, temporal and social. People never speak the same way in all time. They exploit the nuances for different purposes. People of different social classes, different occupations or different cultural groups in the same community will show variations in speech. People of different occupation have their own dialects and they use their own technical terms for better understanding. Education brings a greater difference in language style. History has contributed its own compliments to language. During wars people acquire words used by military people and in course of time these words spread through generations. It correlated with geographical factors such as un-bridged rivers, impenetrable forests, valleys, mountains, deserts etc. Marshals and artificial political barriers divide speech communities.

Language variation is due to different internal factors like semantics, vocabulary, grammar, phonological features, intonation patterns etc. along with other external factors such as region,

caste, religion, education, occupation, social stratum, style, register etc. In various levels of linguistic structure shows variations in different regional varieties of a language. Different groups of people who are living in two different areas show considerable differences in their language patterns. Bengali spoken by any group of the northern region and that of the southern region shows significant changes though features are almost uniform for any group of the respectable regions. There are many lexical items with purely regional connotations and the same forms in two areas have two different meanings and also there are forms which are taboos in one region is not so in another region. Similarly certain verbs and nouns have co-occurrence restrictions at regions.

All dialects of a language are equally efficient and expressive. In the case of Bengali the socio economic and political status of the speech community has nothing to do the standardization of the dialect. Irrespective of the socio-economic factors, all people use both the high and low varieties of Bengali for different purposes. In Bengali speech community, more of the lexical codes of the regional and caste dialects interfere with standard Bengali.

Eminent scholars and great linguists like Suniti Kumar Chatterjee and Sukumar Sen classified Bengali in 5 major dialects by their phonology and pronounciation. They are (1) Rarhi Dialect - spoken in central part of Kolkata with Birbhum as its center; (2) Bangali Dialect - spoken in and aroung Dacca, Barisal, Mymensingh; (3) Kamrupi – spoken in the north-eastern part of Bengal; (4) Barendri – spoken in northern part of Bengal; (5) Jharkhandi – spoken in the south-western part of Bengal. These divisions are superficial. Although the people of Birbhum, Murshidabad, Burdwan, Nadia, 24 Parganas and the major part of Midnapur speak the Rarhi dialect in general, but there are significant differences between the dialects used by the people of Burdwan and Nadia.

During standardization of Bengali in the late 19^{th} and 20^{th} centuries, the cultural elite was mostly from the regions of Kolkata, Hooghly, Howrah, North 24 Parganas and Nadia. What is accepted as the standard form today in West Bengal is based on the West-Central dialect while the language has been standardized today through two centuries of education and media standard.

LDC-IL has taken up the Bengali speaking areas into two regions of Standard Colloquial (West-Central dialect) and Barendri (North dialect) and collected speech data from each. After determining the regions for fieldwork, the datasets were prepared for each region.

21.2 DATASET PREPARATION FOR BENGALI

For the Regions of Standard Colloquial (West-Central) and Barendri (North) LDC-IL prepared the following dataset by which the prompt sheets were prepared.

Content Type	**Count**
Created Text	6
Date	3
Command and Control Words	238
Most Frequent Words	1000
Form and Function Words	248
Phonetically Balanced Words	475
Person Name	501
Place Name	322
Sentences	200

Table 21-1: LDC-IL Bengali 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 seleted part of the dataset prepared as follows.

Content Type	Content that Each typical prompt sheet had	Content selection type
News Text	1 Text	Distinct Text
Created 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 21-2: 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. All the speakers who provided their recordings of Bengali Speech Corpus to LDC-IL are native speakers of West Bengal and of Bengali as their first language.

The Collection of data is carried out in three phases for different regions as follows:

Region/Place	Year of data collection	Resource Person
Standard Colloquial (Central Bengal) and Barendri (North Bengal)	2008	Arundhati Sengupta
Standard Colloquial (Central Bengal) and Barendri (North Bengal)	2008	Priyanka Biswas
Standard Colloquial (Central Bengal) and Barendri (North Bengal)	2008	Sonali Sutradhar

Table 21-3: Phase of Bengali Speech Data Collection

21.3 TRANSLITERATIONS IN LDC-IL BENGALI READ CORPUS

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

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

LDC-IL Transliteration Schema
Bengali characters to Roman and Bengali 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	র	ল	শ	স	ষ	হ	ড়	ঢ়	য়	ৎ
ya	ra	la	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

21.4 SUMMARY OF THE CORPORA

In the sections below, we provide the tabular details of the different content types of the Bengali 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 Speech data has 73399 Audio segments with the duration of 130:11:14 (hh:mm:ss)

21.4.1 Summary of the Audio Segments

Each audio segment of Content type "News-T1" and "Created Text-T2" are audio reocordings of several minutes, which are readings of continuous text . Each audio segment of "Sentence-S" has a sentence. Each audio segment of "Date-D" contains the answer in date format for predetermined quiestionare. Each audio segment of the content type "Command and Control Words-W1", "Person Name-W2", "Place Name-W2", "Most Frequent Word-Part-W3A", "Most Frequent Word-FullSet-W3B", "Phonetically Balanced-W4", "Form and Function Word-W5" has a word uttered three times.

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

LDC-IL Bengali 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	Segmens	Segmens
Contemporary Text (News)-T1	450	24	142	57	27	135	65
Created Text-T2	448	23	142	57	26	135	65
Sentence-S	11239	600	3548	1423	675	3370	1623
Date-D	414	21	130	50	26	124	63
Command and Control Words-W1	13477	712	4254	1710	810	4045	1946
Person Name-W2	9012	481	2846	1138	538	2709	1300
Place Name-W2	4498	240	1414	570	270	1352	652
Most Frequent Word-Part-W3A	13525	720	4273	1715	809	4051	1957
Most Frequent Word-FullSet-W3B	5978	0	2987	0	0	2991	0
Phonetically Balanced-W4	9488	1425	2372	949	1421	948	2373
Form and Function Word-W5	4870	743	1234	494	673	493	1233

Table 21-4: Bengali Audio Segments and their Distribution

21.4.2 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 Bengali 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	35:05:07	1:51:17	11:19:33	4:22:09	2:03:08	10:23:41	5:05:19
Creative Text-T2	20:16:13	1:00:59	6:26:18	2:35:29	1:13:19	6:01:08	2:59:00
Sentence-S	16:05:22	0:49:13	5:09:50	2:06:33	0:55:33	4:45:21	2:18:52
Date-D	0:26:48	0:01:16	0:08:22	0:03:33	0:01:35	0:07:44	0:04:18
Command and Control Words-W1	14:00:24	0:43:43	4:25:05	1:51:14	0:45:00	4:19:15	1:56:07
Person Name-W2	4:56:22	0:13:57	1:34:36	0:40:15	0:15:35	1:27:41	0:44:18
Place Name-W2	1:45:35	0:04:56	0:34:03	0:14:21	0:05:35	0:31:01	0:15:39
Most Frequent Word-Part-W3A	13:33:14	0:42:49	4:14:52	1:45:01	0:44:34	4:12:42	1:53:16
Most Frequent Word-Ful`lSet-W3B	6:47:05	0:00:00	3:32:22	0:00:00	0:00:00	3:14:43	0:00:00
Phonetically Balanced-W4	11:54:02	1:27:33	4:13:34	1:02:45	1:29:26	1:04:01	2:36:43
Form and Function Word-W5	5:21:02	0:49:47	1:19:44	0:31:33	0:44:13	0:33:15	1:22:30

Table 21-5: Duration of the Bengali Speech Data

21.5 DISTINCT SET

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

21.5.1 The Contemporary Text (News) T-1

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

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution			
				Standard		Barendri	
		Female	Male	Female	Male	Female	Male
16 To 20	51	24	27	18	22	6	5
21 To 50	277	142	135	119	112	23	23
50+	122	57	65	49	56	8	9
Total	450	223	227	186	190	37	37

Table 21-6: Distribution of Bengali Contemporary Text (News) Data

21.6 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 datasets available. The random sets are given below:

21.6.1 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			
				Standard		Barendri	
		Female	Male	Female	Male	Female	Male
16 To 20	49	23	26	17	21	6	5
21 To 50	277	142	135	119	112	23	23
50+	122	57	65	49	56	8	9
Total	448	222	226	185	189	37	37

Table 21-7: Distribution of Bengali Creative Text

21.6.2 The Date-D

These are answers to one randomly selected question from a list of 3 questions to get the date format pronounced/spoken commonly by the speaker. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution			
				Standard		Barendri	
		Female	Male	Female	Male	Female	Male
16 To 20	47	26	21	21	16	5	5
21 To 50	254	125	129	104	109	21	20
50+	113	61	52	55	44	6	8
Total	414	212	202	180	169	32	33

Table 21-8: Distribution of Bengali Date Format

21.6.3 Sentences-S

The sentence content type contains a list of sentences that is a representation of all the phonemes occurring in Bengali. 25 randomly selected sentences are recorded. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender wise Distribution		Region-wise Distribution			
				Standard		Barendri	
		Female	Male	Female	Male	Female	Male
16 To 20	1275	600	675	450	550	150	125
21 To 50	6918	3548	3370	2972	2795	576	575
50+	3046	1423	1623	1225	1398	198	225
Total	11239	5571	5668	4647	4743	924	925

Table 21-9: Distribution of Bengali Sentences

21.6.4 Command and Control Words-W1

The command and control words content type contains a list of 238 words that is a representation of most of the command and control words occurring in Bengali. 30 randomly selected words is recorded from a list of words. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender wise Distribution		Region-wise Distribution			
				Standard		Barendri	
		Female	Male	Female	Male	Female	Male
16 To 20	1522	712	810	533	660	179	150
21 To 50	8299	4254	4045	3566	3355	688	690
50+	3656	1710	1946	1470	1675	240	271
Total	13477	6676	6801	5569	5690	1107	1111

Table 21-10: Distribution of Bengali Command and Control Words

21.6.5 Person Names-W2

The person name contains a list of 501 popular pan-Indian and regional person name. 20 randomly selected names are recorded from a list of names. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender wise Distribution		Region-wise Distribution			
				Standard		Barendri	
		Female	Male	Female	Male	Female	Male
16 To 20	1019	481	538	361	439	120	99
21 To 50	5555	2846	2709	2380	2237	466	472
50+	2438	1138	1300	978	1115	160	185
Total	9012	4465	4547	3719	3791	746	756

Table 21-11: Distribution of Bengali Person Names

21.6.6 Place Names-W2

The place name contains a list of 322 popular pan-Indian and regional place names. 10 randomly selected names are recorded from this list by each speaker. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution			
				Standard		Barendri	
		Female	Male	Female	Male	Female	Male
16 To 20	510	240	270	180	220	60	50
21 To 50	2766	1414	1352	1191	1122	223	230
50+	1222	570	652	490	560	80	92
Total	4498	2224	2274	1861	1902	363	372

Table 21-12: Distribution of Bengali Place Names

21.6.7 Most Frequent Words-W3A

The most frequent words contains a list of 1000 most frequent words. 30 randomly selected words is recorded from this list. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution of words		Region-wise Distribution			
				Standard		Barendri	
		Female	Male	Female	Male	Female	Male
16 To 20	1529	720	809	540	659	180	150
21 To 50	8324	4273	4051	3573	3360	700	691
50+	3672	1715	1957	1474	1681	241	276
Total	13525	6708	6817	5587	5700	1121	1117

Table 21-13: Distribution of Bengali Most Frequent Words

21.7 FULL SET

The Full sets are the master set of certain data sets which are read completely from few selected speakers. The full sets are as below:

21.7.1 Most Frequent Words-W3B

The most frequent words contains a list of 1000 most frequent words. In full set all the 1000 words is recorded from the informant. The distribution of data is as follows:

Region-wise Distribution of Native Speakers							
Age Group	Total Audio Segments	Gender-wise Distribution of words		Dialects			
				Standard		Barendri	
		Female	Male	Female	Male	Female	Male
21 To 50	5978	2987	2991	1994	1993	993	998

Table 21-14: Distribution of Bengali Most Frequent Words (Full set)

21.7.2 Phonetically Balanced Vocabulary-W4

The phonetically balanced words are a list of words where all the phones of Bengali language have occurred in all the positions of a word. In full set all the 475 words are recorded from the informant where they uttered those words three times. The distribution of data is as follows:

Region-wise Distribution of Native Speakers							
Age Group	Total Audio Segments	Gender-wise Distribution of words		Dialects			
				Standard		Barendri	
		Female	Male	Female	Male	Female	Male
16 To 20	2846	1425	1421	1425	1421	0	0
21 To 50	3320	2372	948	1897	948	475	0
50+	3322	949	2373	949	1899	0	474
Total	9488	4746	4742	4271	4268	475	474

Table 21-15: Distribution of Bengali Phonetically Balanced Words (Full set)

21.7.3 Form and Function Words-W5

The form and function words content type contains a list of 248 words that is a representation of most of the form and function words occurring in Bengali. In full set, all the 248 words are recorded from the informant where they uttered those words three times. The distribution of data is as follows:

Region-wise Distribution of Native Speakers							
Age Group	Total Audio Segments	Gender-wise Distribution of words		Dialects			
				Standard		Barendri	
		Female	Male	Female	Male	Female	Male
16 To 20	1416	743	673	743	673	0	0
21 To 50	1727	1234	493	987	493	247	0
50+	1727	494	1233	494	986	0	247
Total	4870	2471	2399	2224	2152	247	247

Table 21-16: Distribution of Bengali Form and Function words (Full set)

21.8 BENGALI NATIVE SPEAKERS DISTRIBUTIONS

LDC-IL speech data covered two variations of Bengali speech. So before going to the field the speaker distribution was an important task. As Standard Colloquial has to be taken from 7 districts of West Bengal (i.e Kolkata, North 24 Parganas, South 24 Parganas, Howrah, Hooghly, Birbhum, Nadia) maximum speaker was distributed in that region and minimum speaker was distributed in the Barendri. The distribution of speaker is as follows:

Region-wise Distribution of Native Speakers							
Age Group	Total Native Speakers	Gender-wise Distribution of Native Speakers		Dialects			
				Standard		Barendri	
		Female	Male	Female	Male	Female	Male
16 To 20	57	27	30	21	25	6	5
21 To 50	290	150	140	125	116	25	24
50+	129	59	70	51	60	8	10
Total	**476**	236	240	197	201	39	39

Table 21-17: Distribution of Bengali Native Speakers

22 BODO RAW SPPECH CORPUS

Narayan Choudhary, Hemalata Daimary, Rajesha N, Manasa G, L. Ramamoorthy

22.1 INTRODUCTION

Bodo is the language of Bodos, which are the major tribes of Indian State of Assam. The community is believed to have migrated from South-West China through Tibet and Burma. The language belongs to the Tibeto Burmese linguistic family. Bodo is one of the Tonal languages of the world. There are two clearly distinguishable kinds of tones in Bodo which are known as *Low* and *High*.

Bodos had their kingdoms in the state of Tripura, at Dimapur and Maibong of Nagaland, and in the districts of Darrang, Nagaon, and Kamrup of Assam. The Bodo language is one among the Scheduled languages and one of the official languages of Indian State of Assam. The language is closely related to Dimasa language spoken in Assam, Garo language spoken in Meghalaya and Kokborok language spoken in Tripura.

It is claimed by some intellectuals that the Deodhai script was used by ancient Bodo people. Bodo language was written in Roman and a modified Assamese script called as Purbalipi. From 1963 Bodo was introduced as a medium of instruction in school in Bodo dominated areas. In 1974, the Bodo Sahitya Sabha the apex body of the Bodos in the field of language, literature and culture decided in favor of adoption and introduction of Roman script for the Bodo language in all spheres, and started a movement demanding recognition of Roman script for Bodo language. However, through the intervention of Government of India at the center, Bodo Sahitya Sabha had to adopt Devanagari script for Bodo language in 1975, which was implemented in Education from the year 1976. The LDC-IL Bodo Speech Data is collected by reading out prompt sheets of Bodo in Devanagari Script.

The Bodo Speech data is taken from the following districts of Assam. The data is taken from Bodo mother tongue speakers. LDC-IL collected the following Regional dialects in the corresponding districts of Assam

Bodo Regional Dialect	District
BWRDWNARI	Chirang
EASTERN DIALECT	Baksa
EASTERN DIALECT	Sonitpur
EASTERN DIALECT	Udalguri
EASTERN DIALECT	Kamrup
EASTERN DIALECT	Barpeta
NON-STANDARD	Udalguri
STANDARD	Kokrajhar

Table 22-1: Bodo Speech data Collected Areas.

22.2 DATASET PREPARATION FOR BODO

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

Content Type	Count
Created Text	6
Date	3
Command and Control Words	486
Most Frequent Words	1000
Form and Function Words	304
Phonetically Balanced Words	298
Person Name	502
Place Name	324
Sentences	241

Table 22-2: LDC-IL Bodo 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 part of the dataset prepared as follows.

Content Type	Content that Each typical prompt sheet had	Content selection type
News Text	1 Text	Distinct Text
Created 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 22-3: Table of Contents in LDC-IL Dataset

The full set of

4. Phonetically Balanced Vocabulary
5. Form and Function Words
6. 1000 Most Frequent Words

were also carried to the field to get recorded by selected individuals. Once all these preparations are made, the investigator started collecting the data. The first language of all the speakers who provided their recordings for Bodo Speech Corpus to LDC-IL is Bodo.

Data is collected from the 3 field works, with details as below:

Filed Work	Investigator name
November 2009 to January 2010	Bridul Basumatary
August 2010 to October 2010	Bridul Basumatary
August 2010 to October 2010	Farson Daimary

Table 22-4: Phase of Bodo Speech Data Collection

22.3 TRANSLITERATIONS IN LDC-IL BODO READ CORPUS

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

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

The LDC-IL transliteration scheme of Devanagari to Roman and Numerals to Hindu-Arabic given below.

Vowels											
Vowel	अ	आ	इ	ई	उ	ऊ	ऋ	ए	ऐ	ओ	औ
Matra		ा	ि	ी	ु	ू	ृ	े	ै	ो	ौ
Key	a	A	i	I	u	U	x	E	Ai	O	au

Consonant								
Consonant	क	ख	ग	घ	ङ			
Key	K	kh	g	gh	ng'			
Consonant	च	छ	ज	झ	ञ			
Key	C	ch	j	jh	nj'			
Consonant	ट	ठ	ड	ढ	ण	ड़	ढ़	
Key	T	Th	D	Dh	N	D'	Dh'	
Consonant	त	थ	द	ध	न			
Key	T	th	d	dh	n			
Consonant	प	फ	ब	भ	म			
Key	P	ph	b	bh	m			
Consonant	य	र	ल	व	श	ष	स	ह
Key	Y	r	l	v	Sh	S	s	h
Consonant	श	ष	स	ह				
Key	Sh	S	s	h				

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

22.4 SUMMARY OF THE CORPORA

In the sections below, we provide the tabular details of the different content types of the Bodo 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.

22.4.1 Summary of the Utterances

The table below shows the total number of utterances and their distribution in the Bodo speech dataset.

LDC-IL Bodo 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	411	42	113	34	40	116	66
Creative Text-T2	413	42	115	34	40	116	66
Sentence-S	10257	1048	2854	840	999	2870	1646
Date-D	938	106	253	79	102	254	144
Command and Control Words-W1	12348	1233	3444	1014	1200	3479	1978
Person Name-W2	8222	834	2285	680	799	2305	1319
Place Name-W2	4115	420	1135	340	400	1160	660
Most Frequent Word-Part-W3A	12397	1261	3450	1021	1200	3482	1983
Most Frequent Word-FullSet-W3B	15999	0	5999	3000	1000	4000	2000
Phonetically Balanced-W4	5960	596	1788	1192	298	1192	894
Form and Function Word-W5	6383	607	1824	1216	304	1520	912

Table 22-5: Bodo Audio Segments and their Distribution

22.4.2 Duration of the Bodo Raw Speech Data

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

LDC-IL Bodo 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	53:47:56	5:04:18	13:23:14	6:03:32	6:03:32	13:44:47	9:28:33
Creative Text-T2	26:43:07	2:30:39	6:40:26	2:49:59	2:41:04	6:59:28	5:01:31
Sentence-S	13:58:15	0:51:33	04:12:51	02:02:30	00:47:58	04:08:47	01:54:36
Date-D	1:16:54	0:08:25	0:19:27	0:07:35	0:08:06	0:19:04	0:14:17
Command and Control Words-W1	10:21:55	1:26:18	3:43:31	1014	1:25:56	3:46:10	1978
Person Name-W2	13:04:49	0:47:43	03:55:25	01:57:15	00:45:39	03:53:33	01:45:14
Place Name-W2	04:48:42	0:18:23	01:25:30	00:41:57	00:16:44	01:27:39	00:38:29
Most Frequent Word-Part-W3A	14:34:05	1:31:11	3:49:52	1:21:38	1:27:48	3:51:25	2:32:11
Most Frequent Word-FullSet-W3B	20:07:33	00:00:00	5:56:49	4:49:10	1:08:09	5:45:53	2:27:32
Phonetically Balanced-W4	7:50:00	0:36:01	1:51:06	1:54:58	0:20:16	1:56:03	1:11:36
Form and Function Word-W5	8:28:25	0:41:15	1:55:17	2:02:18	0:21:29	2:14:10	1:13:56

Table 22-6: Duration of the Bodo Speech Data

22.5 DISTINCT SET

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

22.5.1 Contemporary Text (News) –T1

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

Age Group	Total Audio Segments	Gender-wise Distribution of text		Region-wise Distribution							
				BWRDWNARI		EASTERN DIALECT		NON-STANDARD		STANDARD	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	82	42	40	1	0	1	1	8	6	32	33
21 To 50	229	113	116	2	1	31	19	21	28	59	68
50+	100	34	66	1	10	7	8	7	8	19	40
Total	**411**	189	222	4	11	39	28	36	42	110	141

Table 22-7: Distribution of Bodo Contemporary Text (News) Data

22.6 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 datasets available. The random sets are given below:

22.6.1 Creative Text-T2

One randomly selected text of literature out of 6 texts from the prepared Bodo 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							
				BWRDWNARI		EASTERN DIALECT		NON-STANDARD		STANDARD	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	82	42	40	1	0	1	1	9	6	31	33
21 To 50	231	115	116	2	2	31	17	22	28	60	69
50+	100	34	66	1	10	7	8	6	8	20	40
Total	413	191	222	4	12	39	26	37	42	111	142

Table 22-8: Distribution of Bodo Creative Text

22.6.2 Sentences-S

The sentences content type contains a list of sentences that is a representation of almost all the phonemes occurring in Bodo. 25 Randomly selected Sentences are recorded from a list of 241 sentences. The distribution of data is as follows:

Age Group	Total Audio Segments	Genderwise Distribution		Region-wise Distribution							
				BWRDWNARI		EASTERN DIALECT		NON-STANDARD		STANDARD	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16To20	814	440	374	25	0	24	25	156	116	235	233
21 To 50	1428	737	691	48	26	232	195	218	230	239	240
50+	1202	511	691	25	163	138	143	109	145	239	240
Total	**3444**	1688	1756	98	189	394	363	483	491	713	713

Table 22-9: Distribution of Bodo Sentences

22.6.3 Date-D

The answers for 3 questions are collected from each speaker to get the Bodo date format of the informants. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution							
				BWRDWNARI		EASTERN DIALECT		NON-STANDARD		STANDARD	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16To20	17	10	7	2	0	2	2	3	2	3	3
21 To 50	21	11	10	2	2	3	2	3	3	3	3
50+	19	10	9	2	2	2	2	3	2	3	3
Total	**57**	31	26	6	4	7	6	9	7	9	9

Table 22-10: Distribution of Bengali Date Format

22.6.4 Command and Control Words-W1

The command and control words content type contains a list of 486 words that is a representation of almost all the command and control words occurring in Bodo. 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		Region-wise Distribution							
				BWRDWNARI		EASTERN DIALECT		NON-STANDARD		STANDARD	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16To20	1319	700	619	30	0	30	30	223	158	417	431
21 To 50	2554	1314	1240	58	56	405	308	377	396	474	480
50+	1792	715	1077	30	235	186	191	152	199	347	452
Total	**5665**	2729	2936	118	291	621	529	752	753	1238	1363

Table 22-11: Distribution of Bodo Command and Control Words

22.6.5 Person Name -W2

The person name contains a list of 502 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		Region-wise Distribution							
				BWRDWNARI		EASTERN DIALECT		NON-STANDARD		STANDARD	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16To20	1059	557	502	20	0	20	20	157	111	360	371
21 To 50	2221	1141	1080	38	40	338	236	300	330	465	474
50+	1377	525	852	20	173	114	139	114	144	277	396
Total	**4657**	2223	2434	78	213	472	395	571	585	1102	1241

Table 22-12: Distribution of Bodo Person Names

22.6.6 Place Name-W2

The place name contains a list of 324 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		Region-wise Distribution							
				BWRDWNARI		EASTERN DIALECT		NON-STANDARD		STANDARD	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16To20	572	302	270	10	0	10	10	82	49	200	211
21 To 50	1251	633	618	20	20	182	128	164	177	267	293
50+	740	266	474	10	87	58	74	58	73	140	240
Total	**2563**	1201	1362	40	107	250	212	304	299	607	744

Table 22-13: Distribution of Bodo Place Names

22.6.7 Most Frequent Word-Part-W3A

The most frequent words-part contains a list of 1000 most frequent words occurring in Bodo. 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		Region-wise Distribution							
				BWRDWNARI		EASTERN DIALECT		NON-STANDARD		STANDARD	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16To20	1746	917	829	30	0	30	30	244	171	613	628
21 To 50	3886	1964	1922	59	60	557	401	495	587	853	874
50+	2203	785	1418	30	262	170	218	141	215	444	723
Total	**7835**	3666	4169	119	322	757	649	880	973	1910	2225

Table 22-14: Distribution of Bodo Most Frequent Words (Part)

22.7 FULL SETS

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

22.7.1 Most Frequent Word-Full-W3B

The most frequent words contains a list of 1000 most frequent words. In full set all the 1000 words is recorded from the informant. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution							
				BWRDWNARI		EASTERN DIALECT		NON-STANDARD		STANDARD	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	1000	0	1000	0	0	0	0	0	0	0	1000
21 To 50	5000	3000	2000	0	0	1000	0	1000	1000	1000	1000
50+	4000	2000	2000	0	0	0	0	1000	1000	1000	1000
Total	**10000**	5000	5000	0	0	1000	0	2000	2000	2000	3000

Table 22-15: Distribution of Bodo Most Frequent Words - Full

22.7.2 Phonetically Balanced Vocabulary-W4

The phonetically balances vocabulary contains a list of words where all the phones of Bodo language has occurred in all the positions of a word. In full set all the 298 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		Region-wise Distribution							
				BWRDWNARI		EASTERN DIALECT		NON-STANDARD		STANDARD	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	894	596	298	0	0	0	0	298	0	298	298
21 To 50	1490	894	596	0	0	298	0	298	298	298	298
50+	1490	894	596	0	0	298	0	298	298	298	298

Total	**3874**	2384	1490	0	0	596	0	894	596	894	894

Table 22-16: Distribution of Bodo Phonetically Balanced Vocabulary

22.7.3 Form and Function Word-W5

The Form and Function Words contain a list of 304 words which is a representation of almost all the form and function words occurring in Bodo. 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 of words		Region-wise Distribution							
				BWRDWNARI		EASTERN DIALECT		NON-STANDARD		STANDARD	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	911	607	304	0	0	0	0	304	0	303	304
21 To 50	1520	912	608	0	0	304	0	304	304	304	304
50+	608	304	304	0	0	0	0	0	0	304	304
Total	**3039**	1823	1216	0	0	304	0	608	304	911	912

Table 22-17: Representation of Form and Function Word

22.8 LDC-IL BODO SPEECH DATA - NATIVE SPEAKERS DISTRIBUTION

The following table shows the distributions of Bodo Native Speakers across the regional dialects

Region-wise Distribution of Native Speakers											
Age Group	Total Native Speakers	Gender-wise Distribution of Native Speakers		Dialects							
				BWRDWNARI		EASTERN DIALECT		NON-STANDARD		STANDARD	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	86	46	40	1	0	1	1	11	6	33	33
21 To 50	258	131	127	2	2	34	20	34	34	61	71
50+	112	43	69	1	10	7	8	14	11	21	40
Total	**456**	220	236	4	12	42	29	59	51	115	144

Table 22-18: Representation of Bodo Native Speakers Distributions

23 HINDI RAW SPEECH CORPUS

Satyendra Awasthi, Madhupriya Pathak, Rajesha N, Manasa G, Narayan Choudhary, L. Ramamoorthy

23.1 INTRODUCTION

Hindi is an Indo-Aryan language, a descendent of *Sanskrit*, which is spoken in the central and northern India, in the states of *Bihar, Chhattisgarh, Delhi, Haryana, Himachal Pradesh, Jharkhand, Madhya Pradesh, Rajasthan, Uttarakhand* and *Uttar Pradesh*. It is the official language of the Union of India and is also *lingua franca* across India. Being the most intelligible language of India, it is currently reported to be spoken as the first language by 528.35 million people in India (as per 2011 census of India) i.e. a total of 43.63% of the populace of India speaks Hindi as their primary language.

According to the constitution of India the official languages are written in the Devanagari and English. Based on the provisions mentioned in the Official Language Act, Hindi is used for official activities such as communications between the Central Government and a State Government, judiciary and parliamentary proceedings.

Hindi is written in Devenagari script, a Left to Right script which is a descendent of *Brahmi* script. The script is also used to write several other languages of India and neighboring countries such as Nepali, Marathi, Maithili etc.

Many ethnolects, sociolects and other varieties of Hindi are in practice in the Hindi-belt, which are very different from each other in terms of phonological nuances and structural features. However, only major varieties which are intelligible throughout the Hindi plain due to the variety continuum, and are accepted as the dialects/sister languages of Hindi by the academia are considered in the fieldwork.

As claimed by various linguists of repute, Hindi can be divided into following varieties based on the phonological nuances:

#	Regions	Dialects
1.	Eastern Hindi	Awadhi, Bagheli, Chhattisgarhi
2.	Western Hindi	Bangru, Braj, Bundeli, Haryanvi, Kannauji, Khariboli
3.	Bihari Hindi	Bhojpuri, Magahi
4.	Rajasthani Hindi	Marwadi, Mewati, Malvi, Dhundhandi
5.	Pahari Hindi	Pahari (Himachal Pradesh), Kumaoni, Garhwali

Table 23-1: Hindi Regions and Dialects

LDC-IL divided the Hindi speaking areas into these five regions and collected speech data from three regions. Hindi speech corpus is representative in the terms that it reflects all the nuances of dialects it was collected from. Many phonological nuances such as no-distinction in pronunciation of sibilants in ‘Awadhi speakers’ speech, simplification of consonant clusters in ‘Bihari speakers’ speech can be easily seen across the corpus.

23.2 DATASET PREPARATION FOR HINDI

Three fieldworks were conducted in Eastern Hindi and Bihari Hindi regions. Speech data comprising of 434 speakers was collected in these fieldworks. The fieldworks covered the areas as follows.

Eastern Hindi (Awadhi belt)

- Allahabad, UP
- Balrampur, UP
- Lucknow, UP
- Sitapur, UP

Bihari Hindi (Bhojpuri belt and Magahi belt)

- Bhabua, Bihar
- Bhojpur, Bihar
- Buxar, Bihar
- Chandauli, UP
- Gaya, Bihar
- Ghazipur, UP
- Gopalganj, Bihar
- Gorakhpur, UP
- Jehanabad, Bihar
- Kaimur, Bihar
- Nalanda, Bihar
- Varanasi,U

LDC-IL collected the Hindi speech data using two approaches. Two different kinds of Dataset Models were prepared as follows.

- Dataset preparation Model 1 (T1, T2, W1, W2, W3, W4, W5, S, D)
- Dataset preparation Model 2 (Distinct Texts of T1 and T2)

23.2.1 Model 1 (T1, T2, S, D, W1, W2, W3, W4, W5)

For the Regions of Awadhi belt, Bhojpuri belt and Magahi belt LDC-IL prepared the following dataset by which the prompt sheets were prepared.

Notation	Content Type	Count
T1	Contemporary Text (News)	400
T2	Created Text	6
D	Date	2
S	Sentences	500
W1	Command and Control Words	250
W2	Person Name	500
W2	Place Name	324
W3	Most Frequent Words	1000
W4	Phonetically Balanced Words	800
W5	Form and Function Words	200

Table 23-2: Content Types of Dataset Model-1

The command and control word list also includes such form and function words which is not covered in form and function word list. 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 typical prompt sheet had a distinct news item and selected part of the dataset as follows.

Content Type	Content that Each typical prompt sheet had	Content selection type
News Text	1 Text	Distinct Text
Created 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 23-3: Prompt Sheet Distribution of Contents

The full set of

7. Phonetically Balanced Words
8. Form and Function Words
9. 1000 Most Frequent Words

Were also carried to the field to get recorded by selected individuals.

23.2.2 Model 2 (T1, T2)

For the fieldwork of Delhi-NCR region (Khariboli belt), LDC-IL attempted a different approach of dataset preparation, and tried to concentrate on capturing more distinct continues text recordings rather than word segments. The prompt sheet for Delhi-NCR were prepared as follows.

Content Type	Content in Each typical prompt sheet	Content selection type
News Text	1Text	Distinct Text
Created Text	1 text	Distinct Text

Table 23-4: Content Types of Dataset Model-2

23.3 DATA COLLECTION DETAILS

Once all these preparations are made, the investigator started collecting the data.
The Collection of data is carried out in four different regions as follows.

Region/Place	Year of data collection	Resource Person
Uttar Pradesh	2008	Anjali Sinha
Uttar Pradesh	2008	Jitendra Kumar Singh
Bihar	2008	Dheeraj Kumar Mishra
Delhi-NCR	2012	Satyaendra Awasthi

Table 23-5: Data Collection Fieldwork Details

23.4 TRANSLITERATIONS IN LDC-IL HINDI READ CORPUS

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

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

LDC-IL Transliteration Schema Hindi-Devanagari characters to Roman and Hindi Numerals to Hindu-Arabic											
Vowels and Vowel Signs											
अ	आ	इ	ई	उ	ऊ	ऋ	ए	ऐ	ओ	औ	ऑ
	ा	ि	ी	ु	ू	ृ	े	ै	ो	ौ	ॉ
a	A	i	I	u	U	x	E	ai	O	au	ao
Consonant								Signs			
क्	ख्	ग्	घ्	ङ्				ं	ः	ँ	
k	kh	g	gh	ng'				M	H	m'	
च्	छ्	ज्	झ्	ञ्			क़	ख़	ग़	ज़	फ़
c	ch	j	jh	nj'			k'a	Kh'a	g'a	j'a	ph'a
ट्	ठ्	ड्	ढ्	ण्	ड़्	ढ़्					
T	Th	D	Dh	N	D'	Dh'					
त्	थ्	द्	ध्	न्							
t	th	d	dh	n							
प्	फ्	ब्	भ्	म्							
p	ph	b	bh	m							
य्	र्	ल्	व्	श्	ष्	स्	ह्				
y	r	l	v	sh	S	s	h				
Numerals (Devanagari to Roman mapping)											
०	१	२	३	४	५	६	७	८	९		
0	1	2	3	4	5	6	7	8	9		

23.5 SUMMARY OF THE CORPORA

In the sections below, we provide the tabular details of the different content types of the Hindi 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.

23.5.1 Summary of the Utterances

The table below shows the total number of utterances and their distribution in the Hindi speech dataset.

LDC-IL Hindi 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	455	29	161	26	42	136	61
Created Text-T2	463	31	163	24	41	140	64
Date-D	765	29	277	45	56	238	120
Sentence-S	10182	375	3747	627	697	3125	1611
Command and Control Words-W1	12282	450	4605	771	866	3714	1876
Person Name-W2	8171	278	3058	517	559	2498	1261
Place Name-W2	4085	140	1524	260	279	1253	629
Most Frequent Word-Part-W3A	12320	449	4628	760	840	3758	1885
Most Frequent Word-FullSet-W3B	6994	0	2000	2000	0	1994	1000
Phonetically Balanced-W4	14384	798	3994	3200	800	1598	3994
Form and Function Word-W5	3594	200	1000	795	200	400	999

Table 23-6: Audio Segments and their Distribution

23.5.2 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 Hindi 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)
News-T1	35:32:38	3:51:19	12:22:27	1:42:41	4:35:13	10:41:16	4:02:23
Created Text-T2	27:03:47	3:48:01	6:36:49	1:03:10	4:05:24	8:23:46	3:06:37
Date-D	0:58:08	0:01:55	0:16:21	0:03:16	0:13:55	0:14:23	0:08:18
Sentence-S	9:18:25	0:55:50	3:14:14	0:32:46	0:35:22	2:36:04	1:24:09
Command and Control Words-W1	9:37:52	0:29:47	3:26:30	0:30:59	0:57:18	2:40:02	1:33:16

Person Name-W2	11:16:28	0:35:01	3:58:12	0:52:38	1:00:12	3:04:13	1:46:12
Place Name-W2	3:14:44	0:06:31	1:11:12	0:12:10	0:12:38	1:01:21	0:30:52
Most Frequent Word-Part-W3A	8:54:39	0:31:07	3:14:28	0:29:06	0:31:59	2:49:59	1:18:00
Most Frequent Word-FullSet-W3B	4:30:14	0	1:23:23	1:18:39	0	1:10:44	0:37:28
Phonetically Balanced-W4	10:10:44	0:45:52	3:32:13	2:00:22	0:39:23	0:59:51	2:13:03
Form and Function Word-W5	2:22:35	0:11:03	0:47:37	0:28:53	0:08:24	0:13:57	0:32:41

Table 23-7: Duration of the Collected Data

23.6 DISTINCT 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

23.6.1 Contemporary Text (News) T-1

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

Age Group	Total Text (One distinct text / speaker)	Gender-wise Distribution of text		Region-wise Distribution							
				Awadhi belt		Bhojpuri belt		Magahi belt		Khariboli belt	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	71	29	42	0	13	8	8	6	7	15	14
21 To 50	297	161	136	51	38	49	46	49	38	12	14
50+	87	26	61	3	19	18	21	5	21	0	0
Total	**455**	216	239	54	70	75	75	60	66	27	28

Table 23-8: Distribution of Contemporary Text (News) Data

23.6.2 The Creative Text-T2

Distinct Text Extracts from literary books are recorded from the informants to get the speech data of literary text. These types of distinct creative text are collected from Khariboli belt. The distribution of data is as follows:

Total Text (One distinct text/speaker)	Gender-wise Distribution of text		Khariboli belt (Distinct Set)	
	Female	Male		
29	15	14	Female	Male
26	12	14	15	14
0	0	0	12	14
55	27	28	0	0
29	15	14	27	28

Table 23-9: Distribution of Creative Text (Distinct)

23.7 RANDOM SET

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

23.7.1 The Creative Text-T2

One randomly selected text of literature out of six 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 Text (One distinct text/speaker)	Gender-wise Distribution of text		Region-wise Distribution					
				Awadhi belt (Random Set)		Bhojpuri belt (Random Set)		Magahi belt (Random Set)	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	43	16	27	0	14	8	7	8	6
21 To 50	277	151	126	51	38	49	46	51	42
50+	88	24	64	3	20	18	21	3	23
Total	408	191	217	54	72	75	74	62	71

Table 23-10: Distribution of Created Text (Random)

23.7.2 The Date-D

To get the varieties of date formats of the native speakers, two questions were answered. The distribution of data is as follows:

Age Group	(Total questionnaire (Two questions per speaker)	Gender-wise Distribution of text		Region-wise Distribution					
				Awadhi belt		Bhojpuri belt		Magahi belt	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	85	29	56	0	26	16	16	13	14
21 To 50	515	277	238	101	76	98	92	78	70
50+	165	45	120	6	40	34	42	5	38
Total	**765**	351	414	107	142	148	150	96	122

Table 23-11: Distribution of Date Format

23.7.3 Sentences-S

The sentences content type contains a list of sentences that is a representation of almost all the phonemes occurring in Hindi. 25 Randomly selected sentences are recorded from a list of 500 sentences. The distribution of data is as follows:

Age Group	Total Sentences	Gender-wise Distribution of text	Region-wise Distribution		
			Awadhi belt	Bhojpuri belt	Magahi belt

		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	1072	375	697	0	347	200	200	175	150
21 To 50	6872	3747	3125	1280	949	1223	1150	1244	1026
50+	2238	627	1611	75	505	450	525	102	581
Total	**10182**	4749	5433	1355	1801	1873	1875	1521	1757

Table 23-12: Distribution of Sentences

23.7.4 Command and Control Words-W1

The command and control words content type contains a list of 250 words that is a representation of almost all the command and control words occurring in Hindi. 30 randomly selected words are recorded from a list of words. The distribution of data is as follows:

Age Group	Total Words three utterance each	Gender-wise Distribution of text		Region-wise Distribution					
				Awadhi belt		Bhojpuri belt		Magahi belt	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	1316	450	866	0	417	240	239	210	210
21 To 50	8319	4605	3714	1523	1137	1469	1379	1613	1198
50+	2647	771	1876	90	564	540	630	141	682
Total	12282	5826	6456	1613	2118	2249	2248	1964	2090

Table 23-13: Distribution of Command and Control words

23.7.5 Person Names-W2

The person name contains a list of 500 popular Pan Indian and regional person name. 20 randomly selected names are recorded from a list of names. The distribution of data is as follows:

Age Group	Total Words three utterance each	Gender-wise Distribution of text		Region-wise Distribution					
				Awadhi belt		Bhojpuri belt		Magahi belt	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	877	278	599	0	260	160	160	118	139
21 To 50	5556	3058	2498	1022	759	980	921	1056	818
50+	1778	517	1261	60	378	360	422	97	461
Total	**8211**	3853	4358	1082	1397	1500	1503	1271	1418

Table 23-14: Distribution of Person Names

23.7.6 Place Names-W2

The place name contains a list of 324 popular Pan Indian and regional place name. 10 randomly selected names are recorded from a list of names. The distribution of data is as follows:

Age Group	Total Words	Gender-wise Distribution	Region-wise Distribution		
			Awadhi belt	Bhojpuri belt	Magahi belt

	three utterance each	of text							
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	419	140	279	0	130	80	80	60	69
21 To 50	2777	1524	1253	510	380	490	460	524	413
50+	889	260	629	30	190	180	211	50	228
Total	4085	1924	2161	540	700	750	751	634	710

Table 23-15: Distribution of Place Names

23.7.7 Most Frequent Words-W3A

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

Age Group	Total Words three utterance each	Gender-wise Distribution of text		Region-wise Distribution					
				Awadhi belt		Bhojpuri belt		Magahi belt	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	1289	449	840	0	390	240	240	209	210
21 To 50	8386	4628	3758	1559	1141	1468	1381	1601	1236
50+	2645	760	1885	90	568	522	636	148	681
Total	**12320**	**5837**	**6483**	**1649**	**2099**	**2230**	**2257**	**1958**	**2127**

Table 23-16: Distribution of Most Frequent Words-Part (Radom Selection)

23.8 FULL SET

The full sets are the master set of certain datasets which are read completely from few selected speakers in each groups. The full sets are as below:

23.8.1 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 Words three utterance each	Gender-wise Distribution of text		Region-wise Distribution					
				Awadhi belt		Bhojpuri belt		Magahi belt	
		Female	Male	Female	Male	Female	Male	Female	Male
21 To 50	3994	2000	1994	0	0	1000	997	1000	997
50+	3000	2000	1000	2000	1000	0	0	0	0
Total	**6994**	**4000**	**2994**	**2000**	**1000**	**1000**	**997**	**1000**	**997**

Table 23-17: Distribution of Most Frequent Word (Full Set)

23.8.2 Phonetically Balanced Vocabulary-W4

The phonetically balanced vocabulary contains a list of words where almost all the phones of Hindi language have occurred in all the possible positions of a word. In full set all the 800 words is recorded from the informant where they uttered those words three times. The distribution of data is as follows:

Age Group	Total Words three utterance each	Gender-wise Distribution of text		Region-wise Distribution			
				Awadhi belt		Bhojpuri belt	
		Female	Male	Female	Male	Female	Male
16 To 20	1598	798	800	0	0	798	800
21 To 50	5592	3994	1598	3196	800	798	798
50+	7194	3200	3994	2400	3195	800	799
Total	14384	7992	6392	5596	3995	2396	2397

Table 23-18: Distribution of Phonetically Balanced Vocabulary

23.8.3 Form and Function Words-W5

The form and function words content type contains a list of 200 words that is a representation of almost all the form and function words occurring in Hindi. 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 Words three utterance each	Gender-wise Distribution of text		Region-wise Distribution			
				Awadhi belt		Bhojpuri belt	
		Female	Male	Female	Male	Female	Male
16 To 20	400	200	200	0	0	200	200
21 To 50	1400	1000	400	800	200	200	200
50+	1794	795	999	597	799	198	200
Total	3594	1995	1599	1397	999	598	600

Table 23-19: Distribution of Form and Function Words

23.9 NATIVE SPEAKERS DISTRIBUTIONS

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

Region-wise Distribution of Native Speakers											
Age Group	Total Native Speakers	Gender-wise Distribution of Native Speakers		Region							
				Awadhi belt		Bhojpuri belt		Magahi belt		Khariboli belt	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	75	31	44	0	14	8	9	8	7	15	14
21 To 50	317	174	143	56	39	49	46	57	44	12	14
50+	97	29	68	6	24	18	21	5	23	0	0
Total	489	234	255	62	77	75	76	70	74	27	28

Table 23-20: Regional Distribution of Native Speakers

23.10 MOTHER TONGUE DISTRIBUTIONS OF INFORMANTS

The following table shows the distribution of mother tongue of the native speakers in LDC-IL speech data.

Mother Tongue of the native speaker	Geographical Dialect Distribution of LDC-IL Hindi Speech Corpus				Total speaker
	Awadhi belt	Bhojpuri belt	Khariboli belt	Magahi belt	
Awadhi	62	-	-	-	62

Bhojpuri	-	151	-	-	151
Magahi	-	-	-	124	124
Hindi	77	-	55	20	152
Total	**139**	**151**	**55**	**144**	**489**

Table 23-21: Distribution of Mother Tongue

24 KANNADA RAW SPEECH CORPUS

Rajesha N, Vijayalaxmi F. Patil, Manasa G, Chetan Baji, Narayan Choudhary, L. Ramamoorthy

24.1 INTRODUCTION

One of the most ancient languages of India and a prominent language among the Dravidian language family, and notified as a classical language by the Govt. of India, Kannada is widely spoken in the state of Karnataka and also in some of the border areas of other adjacent states. Kannada is the administrative language of the state of Karnataka. The Kannada language uses a script by its own name which a part of Bramhi script family. The langauge is highly agultinative in nature.

Scholars have described four very broad categories of geographical dialects. They are Mysuru Kannada (Kannada of Old Mysore Region), Dharwad Kannada (Kannada spoken in Mumbai Karnataka Region), Mangaluru Kannada (Kannada Spoken in the coastal region of Karnataka) and Gulbarga Kannada (Kannada spoken in Hyderabad Karnataka Region). (Kettle. F, 1993; Kamath, 2002, 2001; Buchanan, 1807). Of course each one of these consists of sub-dialects that have their own distinctive feature. Many of these distinctions occur because the dialects are strongly influenced by their neighboring languages. Tamil, Marathi, Telugu and Malayalam have shaped their vocabulary and less intensely their grammar. The differences among these geographical dialects are well documented.

The state of Karnataka is formed by integrating Kannada speaking areas. These 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 Kannada, spoken in these areas. For example, the Hyderabad Karnataka region is highly influenced by Urdu as it was the part of the erstwhile Nizam Princely State of Hyderabad. Mumbai Karnataka region was a part of Bombay Presidency where the predominant languages were Marathi and Guajarati. Therefore, Marathi has a great influence on Kannada of this region. Canara Region was divided between Bombay Presidency and Madras presidency where the predominant languages are Tulu and Konkani. Old Mysore region was a part of erstwhile Mysore Kingdom ruled by Wadiyar Dynasty where Kannada was the administrative Language.

LDC-IL divided the Kannada speaking areas into these four regions and collected speech data from each. A place from which LDC-IL Kannada Speech Data is collected from respective regions is listed in the table below:

Region→	Hyderabad Karnataka	Canara	Mumbai Karnataka	Old Mysore
Places →	1. Kalburgi 2. Bidar 3. Raichur	1. Mangalore (South Canara) 2. Udupi-Kundapur 3. Bhatkal (North Canara)	1. Hubli-Dharwad 2. Gadag-Betageri 3. Haveri-Hirekerur	1. Mysore-Krishnaraja Nagara 2. Mandya 3. Hassan 4. Chikmagalur-Sringeri 5. Bangalore Urban

Table 24-1: Regions and Places Covered for Kannada Speech Data

24.2 DATASET PREPARATION FOR KANNADA

For the selected regions, Hyderabad Karnataka, Canara, Mumbai Karnataka and Old Mysore. 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	82
Most Frequent Words	1,144
Form and Function Words	432
Phonetically Balanced Words	390
Person Name	489
Place Name	511
Sentences	142

Table 24-2: 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
News Text	1 Text	Distinct Text
Created 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 24-3: 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 four phases for different regions as follows:

Region	Year	Field Investigator
Hyderabad Karnataka	2008	Rajesha N
Canara	2009	Rajesha N

Mumbai Karnataka	2009	Rajesha N
Old Mysore	2012	Malini N. Abhyankar

Table 24-4: Four Phases of Speech Data Collection

24.3 TRANSLITERATIONS IN LDC-IL KANNADA READ CORPUS

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

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

LDC-IL Transliteration Schema

Kannada characters to Roman and Kannada Numerals to Hindu-Arabic

Vowels

ಅ	ಆ	ಇ	ಈ	ಉ	ಊ	ಋ	ೠ	ಌ	ೡ	ಎ	ಏ	ಐ	ಒ	ಓ	ಔ
	ಾ	ಿ	ೀ	ು	ೂ	ೃ	ೄ	ೢ	ೣ	ೆ	ೇ	ೈ	ೊ	ೋ	ೌ
a	A	i	I	u	U	x	X	q	Q	e	E	ai	o	O	au

Consonants

ಕ	ಖ	ಗ	ಘ	ಙ
ka	kha	ga	gha	ng'a
ಚ	ಛ	ಜ	ಝ	ಞ
ca	cha	ja	jha	nj'a
ಟ	ಠ	ಡ	ಢ	ಣ
Ta	Tha	Da	Dha	Na
ತ	ಥ	ದ	ಧ	ನ
ta	tha	da	dha	na
ಪ	ಫ	ಬ	ಭ	ಮ
pa	pha	ba	bha	ma
ಯ	ರ	ಲ	ವ	ಶ	ಷ	ಸ	ಹ	ಳ	ೞ	ೱ
ya	ra	la	va	sha	Sa	sa	ha	La	Za	Ra

Symbols

ಀ	ಁ	ಂ	ಃ	಼	ಽ
M'	m'	M	H	J	G

Numerals (Kannada to Hindu-Arabic)

೦	೧	೨	೩	೪	೫	೬	೭	೮	೯
0	1	2	3	4	5	6	7	8	9

Table 24-5: Transliteration Scheme of Kannada to Roman

Note: The letters in gray cells are obsolete in usage or only used for Sanskrit language written in Kannada Script. These letters may rarely present in the corpus.

24.4 SUMMARY OF THE CORPUS

In the sections below, we provide the tabular details of the different content types of the Kannada 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 179:32:52 (hh:mm:ss) comprising 99,109 audio segments.

24.4.1 Summary of the Audio Segments

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

LDC-IL Kannada 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	600	36	180	84	36	180	84
Creative Text-T2	600	36	180	84	36	180	84
Sentence-S	14887	890	4459	2086	910	4460	2082
Date-D	1200	72	360	168	72	360	168
Command and Control Words-W1	17988	1080	5396	2516	1084	5392	2520
Person Name-W2	12009	718	3600	1678	732	3601	1680
Place Name-W2	6032	379	1807	842	361	1803	840
Most Frequent Word-Part-W3A	18065	1114	5416	2523	1080	5408	2524
Most Frequent Word-FullSet-W3B	8000	0	4000	0	0	4000	0
Phonetically Balanced-W4	9360	1560	1560	1560	1560	1560	1560
Form and Function Word-W5	10368	1728	1728	1728	1728	1728	1728

Table 24-6: Audio Segments and their Distribution

24.4.2 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 Kannada 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	66:06:09	4:10:05	20:01:27	09:27:13	03:44:31	19:32:27	09:10:26
Creative Text-T2	33:09:20	2:08:15	09:47:05	04:46:15	01:53:13	09:51:45	04:42:47
Sentence-S	13:58:15	0:51:33	04:12:51	02:02:30	00:47:58	04:08:47	01:54:36
Date-D	01:16:22	0:04:55	00:21:38	00:11:53	00:04:12	00:23:14	00:10:30
Command and Control Words-W1	12:31:43	0:45:25	03:43:29	01:48:43	00:43:55	03:49:35	1:40:36
Place Name-W2	04:48:42	0:18:23	01:25:30	00:41:57	00:16:44	01:27:39	00:38:29
Person Name-W2	13:04:49	0:47:43	03:55:25	01:57:15	00:45:39	03:53:33	01:45:14
Most Frequent Word-Part-W3A	12:21:24	0:46:10	03:38:13	01:44:35	00:42:48	03:49:52	01:39:46
Most Frequent Word-FullSet-W3B	06:45:56	0:00:00	03:23:48	00:00:00	00:00:00	03:22:08	00:00:00
Phonetically Balanced-W4	06:47:23	1:10:40	00:53:28	01:08:16	01:06:45	01:15:41	01:12:33
Form and Function-Word-W5	08:42:49	1:27:34	01:31:17	01:17:43	01:33:10	01:27:18	01:25:47

Table 24-7: Duration of the Collected Data

24.5 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.

24.5.1 The Contemporary Text (News)- T1

Distinct Text Extracts from Newsapers 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							
				Hyderabad Karnataka		Canara		Mumbai Karnataka		Old Mysore	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 to 20	72	36	36	9	9	9	9	9	9	9	9
21 to 50	360	180	180	45	45	45	45	45	45	45	45
50+	168	84	84	21	21	21	21	21	21	21	21
Total	**600**	**300**	**300**	**75**	**75**	**75**	**75**	**75**	**75**	**75**	**75**

Table 24-8 Distribution of Contemporary Text (News) Data

24.6 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.

24.6.1 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							
				Hyderabad Karnataka		Canara		Mumbai Karnataka		Old Mysore	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 to 20	72	36	36	9	9	9	9	9	9	9	9
21 to 50	360	180	180	45	45	45	45	45	45	45	45
50+	168	84	84	21	21	21	21	21	21	21	21
Total	**600**	**300**	**300**	**75**	**75**	**75**	**75**	**75**	**75**	**75**	**75**

Table 24-9: Distribution of Kannada Creative Text

24.6.2 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							
				Hyderabad Karnataka		Canara		Mumbai Karnataka		Old Mysore	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 to 20	144	72	72	18	18	18	18	18	18	18	18
21 to 50	720	360	360	90	90	90	90	90	90	90	90
50+	336	168	168	42	42	42	42	42	42	42	42
Total	**1200**	**600**	**600**	**75**	**75**	**75**	**75**	**75**	**75**	**75**	**75**

Table 24-10: Distribution of Kannada Date Format

24.6.3 The Sentences-S

The Sentences contain a list of sentences that is a representation of all most all the phonemes occurring in Kannada. 25 Randomly selected Sentences are recorded from a list of 142 sentences. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution							
				Hyderabad Karnataka		Canara		Mumbai Karnataka		Old Mysore	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 to 20	1800	890	910	222	238	224	225	222	225	222	222
21 to 50	8919	4459	4460	1117	1119	1112	1112	1114	1112	1116	1117
50+	4168	2086	2082	523	517	523	523	519	520	521	522
Total	**14887**	**7435**	**7452**	**1862**	**1874**	**1859**	**1860**	**1855**	**1857**	**1859**	**1861**

Table 24-11: Distribution of Kannada Sentences

24.6.4 Command and Control Words-W1

The Command and Control Words contain a list of 82 words that is a representation of all most all the command and control words occurring in Kannada. 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							
				Hyderabad Karnataka		Canara		Mumbai Karnataka		Old Mysore	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 to 20	2164	1080	1084	271	276	270	270	270	269	269	269
21 to 50	10788	5396	5392	1349	1348	1350	1349	1349	1348	1348	1347
50+	5036	2516	2520	626	630	630	630	630	630	630	630
Total	**17988**	**8992**	**8996**	**2246**	**2254**	**2250**	**2249**	**2249**	**2247**	**2247**	**2246**

Table 24-12: Distribution of Kannada Command and Control Words

24.6.5 Person Names –W2

The Person Names contain a list of 489 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							
				Hyderabad Karnataka		Canara		Mumbai Karnataka		Old Mysore	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 to 20	1450	720	732	179	192	180	180	180	180	179	180
21 to 50	7202	3600	3605	898	899	900	899	900	904	902	900
50+	3357	1679	1682	420	419	417	420	420	420	420	421
Total	**12009**	**5999**	**6019**	**1497**	**1510**	**1497**	**1499**	**1500**	**1504**	**5999**	**6019**

Table 24-13: Distribution of Kannada Person Names

24.6.6 Place Names-W2

The Place Names contain a list of 511 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							
				Hyderabad Karnataka		Canara		Mumbai Karnataka		Old Mysore	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 to 20	740	379	361	109	90	90	90	90	90	90	91
21 to 50	3610	1807	1803	452	449	450	450	452	454	453	450
50+	1682	842	840	210	209	212	211	210	210	210	210
Total	**6033**	**3028**	**3005**	**771**	**750**	**752**	**751**	**752**	**754**	**753**	**751**

Table 24-14: Distribution of Kannada Place Names

24.6.7 Most Frequent Words-PART-W3A

The Most Frequent Words-part contain a list of 1,144 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							
				Hyderabad Karnataka		Canara		Mumbai Karnataka		Old Mysore	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 to 20	2194	1114	1080	304	270	270	270	270	270	270	270
21 to 50	10824	5416	5408	1354	1349	1351	1354	1358	1353	1353	1352
50+	5047	2523	2524	630	630	633	633	631	631	629	630
Total	**18065**	**9053**	**9012**	**2288**	**2249**	**2254**	**2257**	**2259**	**2254**	**2252**	**2252**

Table 24-15: Distribution of Kannada Most Frequent Words - Part

24.7 FULL SET

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

24.7.1 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:

Total Audio Segments from Speakers of 21-50 Age group	Gender-wise Distribution		Region-wise Distribution							
			Hyderabad Karnataka		Canara		Mumbai Karnataka		Old Mysore	
	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
8000	4000	4000	1000	1000	1000	1000	1000	1000	1000	1000

Table 24-16: Distribution of Kannada Most Frequent Words – Full

24.7.2 The Phonetically Balanced Vocabulary-W4

The Phonetically Balanced Vocabulary contain a list of words where all most all the phones of Kannada language has occurred in all the possible positions of a word. In full set all the 390 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		Region-wise Distribution							
				Hyderabad Karnataka		Canara		Mumbai Karnataka		Old Mysore	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 to 20	3120	1560	1560	390	390	390	390	390	390	390	390
21 to 50	3120	1560	1560	390	390	390	390	390	390	390	390
50+	3120	1560	1560	390	390	390	390	390	390	390	390
Total	**9360**	**4680**	**4680**	**1170**	**1170**	**1170**	**1170**	**1170**	**1170**	**1170**	**1170**

Table 24-17: Distribution of Kannada Phonetically Balanced Vocabulary

24.7.3 The Form and Function Words-W5

The Form and Function Words contain a list of 432 words which is a representation of all most all the form and function words occurring in Kannada. 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		Region-wise Distribution							
				Hyderabad Karnataka		Canara		Mumbai Karnataka		Old Mysore	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 to 20	3456	1728	1728	432	432	432	432	432	432	432	432
21 to 50	3456	1728	1728	432	432	432	432	432	432	432	432
50+	3456	1728	1728	432	432	432	432	432	432	432	432
Total	**10368**	**5184**	**5184**	**1296**	**1296**	**1296**	**1296**	**1296**	**1296**	**1296**	**1296**

Table 24-18: Distribution of Form and Function words

24.8 NATIVE SPEAKERS DISTRIBUTIONS

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

Region-wise Distribution of Native Speakers											
Age Group	Total Native Speakers	Gender-wise Distribution of Native Speakers		Regions							
				Hyderabad Karnataka		Canara		Mumbai Karnataka		Old Mysore	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 to 20	88	44	44	11	11	11	11	11	11	11	11
21 to 50	384	192	192	48	48	48	48	48	48	48	48
50+	184	92	92	23	23	23	23	23	23	23	23
Total	**656**	**328**	**328**	**82**	**82**	**82**	**82**	**82**	**82**	**82**	**82**

Table 24-19: Distribution of Kannada Native Speakers

24.9 MOTHER TONGUE DISTRIBUTION OF THE NATIVE SPEAKERS

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

Mother Tongue of the Native Speaker	Geographical Dialect Distribution of LDC-IL Kannada Speech Corpus				Total speaker
	Hyderabad Karnataka	Canara	Mumbai Karnataka	Old Mysore	
Kannada	156	68	152	159	535
Konkani	-	46	-	-	46
Tulu	-	40	-	-	40
Marathi	2	2	7	-	11
Sankethi	-	-	-	5	5
Telugu	2	1	1	-	4
Urdu	2	2	-	-	4
Malayalam	-	2	-	-	2
Tamil	-	-	2	-	2
Hindi	1	-	1	-	2
Lambani	1	-	1	-	2
Chitpavani	-	2	-	-	2
Kodava	-	1	-	-	1
Total	**164**	**164**	**164**	**164**	**656**

Table 24-20: Representation of Mother Tongue Distribution of the Kannada Native Speakers

24.10 REFERENCES

1. Kittel, F (1993), *A Grammar of the Kannada Language Comprising the Three Dialects of the Language (Ancient, Medieval and Modern)*. New Delhi, Madras: Asian Educational Services.
2. Kamath, Suryanath U. (2002), *A Concise History of Karnataka from Pre-historic Times to the Present*. Bangalore: Jupiter Books.
3. Buchanan, Francis Hamilton (1807). *A Journey from Madras through the Countries of Mysore, Canara, and Malabar. Volume 3*. London: Cadell.

25 KONKANI RAW SPEECH CORPUS

Bhageshree Khandale, Saurabh Varik, Rajesha N, Manasa G, Narayan Choudhary, L. Ramamoorthy

25.1 INTRODUCTION

Konkani is the principal and administrative language of Goa. Konkani is an Indo-Aryan language belonging to the Indo-European family of languages and is spoken along the western coast of India. The Konkani language is spoken widely in the western coastal region of India known as Konkan. This consists of the Konkan division of Maharashtra, the state of Goa, and the Uttara Kannada (formerly North Canara), Udupi, and Dakshina Kannada (formerly South Canara) districts of Karnataka, together with many districts in Kerala (such as Kasargod, Kochi, Alappuzha, Trivandrum, and Kottayam). It is one of the 22 scheduled languages mentioned in the 8th schedule of the Indian Constitution and the official language of the Indian state of Goa. The first Konkani inscription is dated 1187 A.D. It is a minority language in Karnataka, Maharashtra and Kerala,Dadra and Nagar Haveli and Daman and Diu.

Konkani is a member of the southern Indo-Aryan language group. It retains elements of Vedic structures and shows similarities with both western and eastern Indo-Aryan languages. It is inflexive, and less distant from Sanskrit as compared to other modern Indo-Aryan languages. Linguists describe Konkani as a fusion of variety of Prakrits. This could be attributed to the confluence of immigrants that the Konkan coast has witnessed over the years. Konkani developed with overall Sanskrit complexity and grammatical structure, which eventually developed into a lexical fund of its own. The second wave of Indo-Aryans is believed to have been accompanied by Dravidians from the Deccan plateau.

The Konkani language has 16 basic vowels (excluding an equal number of long vowels), 36 consonants, 5 semi-vowels, 3 sibilants, 1 aspirate, and many diphthongs. Like the other Indo-Aryan languages, it has both long and short vowels and syllables with long vowels may appear to be stressed. Different types of nasal vowels are a special feature of the Konkani language. Konkani grammar is similar to other Indo-Aryan languages. Notably, Konkani grammar is also influenced by Dravidian languages. Konkani is a language rich in morphology and syntax. It cannot be described as a stress-timed language, nor as a tonal language.

The vocabulary from Konkani comes from a number of sources. The main source is Prakrits. There are many indications that Konkani is more closer to Sanskrit than any other widely spoken Indian languages. So Sanskrit as a whole has played a very important part in Konkani vocabulary. Other sources of vocabulary are Arabic, Persian, and Turkish. Finally Kannada, Marathi, and Portuguese have enriched its lexical content. following are the konkani Dialects groups: Canara Konkani, Goan Konkani, Individual dialects: Malvani, Mangalorean, Chitpavani, Antruz, Bardeskari, Saxtti and Pednekari.

At present Konkani is written in four scripts: Devanagari, Roman, Kannada and Malayalam Because Devanagari is the official script used to write Konkani in Goa and Maharashtra, most Konkanis (especially Hindus) in those two states write the language in Devanagari. In the state of

Karnataka the Konkani is taught and used in Kannada script. Konkani in Malayalam Script is used by the Konkani people in Kerala State.

25.2 DATASET PREPARATION FOR KONKANI

For the selected Regions, North Goa, South Goa, Karwar and Sindhudurg, 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	326
Most Frequent Words	1000
Person Name	614
Place Name	742
Sentences	425

Table 25-1: Representation of Content Type

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 in each typical prompt sheet	Content selection type
Contemporary Text (News Text)	1 Text	Distinct Text
Created 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 25-2: Representation of Prompt Sheet

The full set of

10. Phonetically Balanced Vocabulary of 425 Words
11. Form and Function Words of 537 words
12. 1000 Most Frequent Wordlist

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

Once all these preparations were made, the investigator started collecting the data. The Collection of data is carried out in three phases.

- Saurabh Varik 2009
- Saurabh Varik 2010

- Yashwant Gawas 2010

Some data is collected from Konakanis who visited CIIL Mysore at different times.

25.3 TRANSLITERATIONS IN LDC-IL KONKANI READ CORPUS

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

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

LDC-IL Transliteration Schema

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

Vowels and Vowel Signs																	
अ	आ	इ	ई	उ	ऊ	ऋ	ॠ	ऌ	ॡ	ऍ	ऎ	ए	ऐ	ऑ	ऒ	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	ऴ
ya	ra	la	va	sha	Sa	sa	ha	La	Za

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

25.4 SUMMARY OF THE CORPUS

In the sections below, we provide the tabular details of the different content types of the Konkani 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 156:37:51 (hh:mm:ss) comprising 72,938 audio segments.

25.4.1 Summary of the Audio Segments

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

LDC-IL Konkani 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)	477	40	152	61	26	136	62
Creative Text-T2	480	37	153	62	26	140	62
Sentence-S	12050	999	3871	1553	649	3423	1555
Date-D	953	80	306	122	49	272	124
Command and Control Words-W1	14944	1165	4943	1828	780	4369	1859
Person Name-W2	9588	778	3084	1224	519	2740	1243
Place Name-W2	4812	390	1545	612	261	1379	625
Most Frequent Word-Part-W3A	16376	1170	5631	1837	780	5090	1868
Most Frequent Word-FullSet-W3B	5998	1000	1000	1000	998	1000	1000
Phonetically Balanced-W4	2975	0	850	425	425	850	425
Form and Function Word-W5	4285	537	1072	537	537	1067	535

Table 25-3: Representation of Audio Segments of Konkani Raw Speech Data

25.4.2 Duration of the Konkani Raw Speech Data

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

LDC-IL Konkani 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)	49:52:09	4:27:19	16:24:18	5:47:57	3:15:38	14:07:57	5:49:00
Creative Text-T2	22:09:05	1:51:59	6:48:40	2:50:24	1:28:41	6:27:56	2:41:25
Sentence-S	15:51:11	1:20:55	5:05:23	1:58:41	0:59:09	4:32:28	1:54:35
Date-D	01:50:39	0:08:32	0:35:16	0:15:02	0:05:01	0:32:32	0:14:16
Command and Control Words-W1	16:11:02	1:13:49	5:17:28	2:01:42	0:45:24	4:45:55	2:06:44
Person Name -W2	15:55:43	1:18:00	5:09:25	2:04:41	0:48:32	4:32:16	2:02:49
Place Name-W2	05:31:03	0:26:04	1:46:32	0:42:34	0:15:56	1:35:57	0:44:00
Most Frequent Word-Part-W3A	16:03:13	1:08:47	5:26:59	1:50:45	0:41:38	5:01:25	1:53:39
Most Frequent Word-FullSet-W3B	05:55:07	0:53:11	1:13:31	0:46:00	1:01:07	1:03:57	0:57:21
Phonetically Balanced-W4	02:49:36	0:00:00	0:51:42	0:21:03	0:24:37	0:48:11	0:24:03
Form and Function Word-W5	04:29:03	0:39:27	0:59:22	0:26:33	0:38:38	0:59:21	0:45:42

Table 25-4: Representation of Konkani Raw Speech Data Duration

25.5 DISTINCT SET

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

25.5.1 Contemporary Text (News)

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

Age Group	Total Audio Segments	Gender-wise Distribution of text		Region-wise Distribution							
				NORTH GOA		SOUTH GOA		KARWARI KONKANI		SINDHUDURG	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 to 20	66	40	26	14	9	14	13	10	4	2	0
21 to 50	288	152	136	63	35	42	53	46	41	1	7
50+	123	61	62	11	9	27	25	23	28	0	0
Total	**477**	**253**	**224**	**88**	**53**	**83**	**91**	**79**	**73**	**3**	**7**

Table 25-5: Representation of Konkani Contemporary text (News)

25.6 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 datasets available. The random sets are given below:

25.6.1 Creative Text-T2

One randomly selected text of literature out of 6 texts from the prepared Konkani 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 of text		Region-wise Distribution							
				NORTH GOA		SOUTH GOA		KARWARI KONKANI		SINDHUDURG	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 to 20	63	37	26	12	9	13	13	10	4	2	0
21 to 50	293	153	140	64	38	43	54	45	41	1	7
50+	124	62	62	11	9	28	25	23	28	0	0
Total	**480**	**252**	**228**	**87**	**56**	**84**	**92**	**78**	**73**	**3**	**7**

Table 25-6: Representation of Konkani Creative Text

25.6.2 Date Format

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

Age Group	Total Audio Segments	Gender-wise Distribution of dates		Region-wise Distribution							
				NORTH GOA		SOUTH GOA		KARWARI KONKANI		SINDHUDURG	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 to 20	129	80	49	28	18	28	23	20	8	4	0
21 to 50	578	306	272	128	72	84	104	92	82	2	14
50+	246	122	124	22	18	54	50	46	56	0	0
Total	**953**	**508**	**445**	**178**	**108**	**166**	**177**	**158**	**146**	**6**	**14**

Table 25-7: Representation of Konkani Date format

25.6.3 Sentences

The sentences content type contains a list of sentences that is a representation of almost all the phonemes occurring in Konkani. 25 Randomly selected Sentences is recorded from a list of 425 sentences. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution of Sentences		Region-wise Distribution							
				NORTH GOA		SOUTH GOA		KARWARI KONKANI		SINDHUDURG	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 to 20	1648	999	649	349	225	350	324	250	100	50	0
21 to 50	7294	3871	3423	1621	899	1075	1323	1150	1026	25	175
50+	3108	1553	1555	276	231	700	624	577	700	0	0
Total	**12050**	**6423**	**5627**	**2246**	**1355**	**2125**	**2271**	**1977**	**1826**	**75**	**175**

Table 25-8: Representation of Konkani Sentences

25.6.4 Command And Control Words

The command and control words content type contains a list of 326 words that is a representation of almost all the command and control words occurring in Konkani. 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 of words		Region-wise Distribution							
				NORTH GOA		SOUTH GOA		KARWARI KONKANI		SINDHUDURG	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 to 20	1945	1165	780	415	270	420	390	270	120	60	0
21 to 50	9312	4943	4369	2243	1376	1290	1553	1380	1230	30	210
50+	3687	1828	1859	329	270	810	750	689	839	0	0
Total	**14944**	**7936**	**7008**	**2987**	**1916**	**2520**	**2693**	**2339**	**2189**	**90**	**210**

Table 25-9: Representation of Konkani Command and Control words

25.6.5 Person Name

The person name contains a list of 614 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 of Person Names		Region-wise Distribution							
				NORTH GOA		SOUTH GOA		KARWARI KONKANI		SINDHUDURG	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 to 20	1297	778	519	279	179	279	260	180	80	40	0
21 to 50	5824	3084	2740	1285	721	859	1059	920	820	20	140
50+	2467	1224	1243	220	183	540	500	464	560	0	0
Total	**9588**	**5086**	**4502**	**1784**	**1083**	**1678**	**1819**	**1564**	**1460**	**60**	**140**

Table 25-10: Representation of Konkani Person Names

25.6.6 Place Name

The place name contains a list of 742 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 of Place Names		Region-wise Distribution							
				NORTH GOA		SOUTH GOA		KARWARI KONKANI		SINDHUDURG	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 to 20	651	390	261	140	91	140	130	90	40	20	0
21 to 50	2924	1545	1379	643	359	430	537	462	413	10	70
50+	1237	612	625	111	94	270	251	231	280	0	0
Total	**4812**	**2547**	**2265**	**894**	**544**	**840**	**918**	**783**	**733**	**30**	**70**

Table 25-11: Representation of Konkani Place Names

25.6.7 Most Frequent Word-Part

The most frequent words-part contains a list of 1000 most frequent words occurring in Konkani. 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 of words		Region-wise Distribution							
				NORTH GOA		SOUTH GOA		KARWARI KONKANI		SINDHUDURG	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 to 20	1950	1170	780	420	270	420	390	270	120	60	0
21 to 50	10721	5631	5090	2929	2050	1290	1596	1382	1234	30	210
50+	3705	1837	1868	330	277	811	751	696	840	0	0
Total	**16376**	**8638**	**7738**	**3679**	**2597**	**2521**	**2737**	**2348**	**2194**	**90**	**210**

Table 25-12: Representation of Konkani Most Frequent Words-Part

25.7 FULL SET

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

25.7.1 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. The distribution of data is as follows:

Age Group	Total Audio Segments	SOUTH GOA REGION	
		Female	Male
16 To 20	1998	1000	998
21 To 50	2000	1000	1000
50+	2000	1000	1000
Total	**5998**	**3000**	**2998**

Table 25-13: Representation of Konkani Most Frequent Word-Full

25.8 PHONETICALLY BALANCED VOCABULARY

The Phonetically Balanced words contain a list of words where almost all the phonemes of Konkani language has occurred in all the possible positions of a word. In full set all the 425 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 of words		Region-wise Distribution			
				NORTH GOA		SOUTH GOA	
		Female	Male	Female	Male	Female	Male
16 to 20	425	0	425	0	0	0	425
21 to 50	1700	850	850	425	425	425	425
50+	850	425	425	0	0	425	425
Total	**2975**	**1275**	**1700**	**425**	**425**	**850**	**1275**

Table 25-14: Representation of Konkani Phonetically Balanced Vocabulary

25.9 FORM AND FUNCTION WORD

The Form and Function Words contain a list of 537 words which is a representation of almost all the form and function words occurring in Konkani. 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 of words		Region-wise Distribution			
				NORTH GOA		SOUTH GOA	
		Female	Male	Female	Male	Female	Male
16 to 20	1074	537	537	0	0	537	537
21 to 50	2674	1072	1602	537	530	535	1072
50+	537	537	0	0	0	537	0
Total	**4285**	**2146**	**2139**	**537**	**530**	**1609**	**1609**

Table 25-15: Representation of Konkani Form And Function Word

25.10 NATIVE SPEAKERS DISTRIBUTIONS

The following table shows the distributions of Konkani Native Speakers across the regions

Region-wise Distribution of Native Speakers											
Age Group	Total Native Speakers	Gender-wise Distribution of Native Speakers		Regions							
				NORTH GOA		SOUTH GOA		KARWARI KONKANI		SINDHUDURG	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 to 20	71	42	29	14	9	16	16	10	4	2	0
21 to 50	304	160	144	66	38	47	58	46	41	1	7
50+	129	65	64	11	9	31	27	23	28	0	0
Total	**504**	**267**	**237**	**91**	**56**	**94**	**101**	**79**	**73**	**3**	**7**

Table 25-16: Representation of Konkani Native Speakers Distributions

25.11 MOTHER TONGUE DISTRIBUTION OF THE NATIVE SPEAKERS

The following table shows the distribution of mother tongue of the native speakers in LDC-IL speech data.

Mother Tongue of the Native Speaker	Geographical Dialect Distribution of LDC-IL Konkani Speech Corpus				Total speaker
	NORTH GOA	SOUTH GOA	KARWARI KONKANI	SINDHUDURG	
Konkani	147	193	9	0	349
Kannada	0	0	140	0	140
Marathi	0	1	1	10	12
Urdu	0	0	2	0	2
Total	**147**	**194**	**152**	**10**	**503**

Table 25-17: Representation of Mother Tongue Distribution of the Konkani Native Speakers

26 MAITHILI RAW SPEECH CORPUS

Dinesh Mishra, Rajesha N, Manasa G, Narayan Choudhary, L. Ramamoorthy

26.1 INTRODUCTION

Maithili is an Indio-Aryan language, a direct descendent of Sanskrit, which is spoken in the states of Bihar, Jharkhand and part of Nepal. It is one of the scheduled languages of India. The name Maithili is derived from the word Mithila, an ancient Kingdom of which King Janaka was the ruler (see Ramayana). Maithili is also of the name of Sita, the wife of King Rama and daughter of King Janaka. Scholars in Mithila used Sanskrit for their literary work and Maithili was the language of the common folk (Abahatta).

It can be observed that Mithila region has been rich in cultural heritage which has produced a distinct cultural landscape over the years. Its evolution has been influenced by geographical isolation surrounded by the three big rivers and lofty mountains. The region remained secluded peaceful and least influenced tract.

Maithili dates back to the 14th century. The "Varna Ratnakara" is the earliest known prose text, preserved from 1507, and is written in "Mithiliksar" script. Maithili was traditionally written in their own script which is known as Mithilakshar or Tirhuta. This script is similar to Bengali-Assamese script. Devanagari script is most commonly used since the 20th century. It was also written in the local variant of Kaithi script. The Tirhuta (Mithilakshar) and Kaithi scripts are both currently included in Unicode.

In the 19th century, linguistic scholars considered Maithili as a dialect of Bihari languages and grouped it with other languages spoken in Bihar. Hoernle compared it with Gaudian languages and recognized that it shows more similarities with Bengali languages than with Hindi. Grierson recognized it as a distinct language and published the first grammar in 1881.

In 2003, Maithili was included in the Schedule of the Indian Constitution as a recognized Indian language, which allows it to be used in education, government, and other official contexts in India. The Maithili language is included as an optional paper in the UPSC Exam.

In India, Maithili is Spoken mainly in Bihar and Jharkhand in the districts of Darbhanga, Madhubani, Samastipur, Muzaffarpur, Sitamarhi, Begusarai, Khagaria, Purnia, Katihar, Kishanganj, Sheohar, Bhagalpur, Madhepura, Araria, Supaul, Vaishali, Saharsa (Bihar) Ranchi, Bokaro, Jamshedpur, Dhanbad, and Deoghar (Jharkhand). The geographic region comprising of these districts is also called as Mithilanchal Region. Darbhanga and Madhubani Constitute cultural and linguistic centers. Native speakers also reside in Patna, Delhi, Kolkata, Mumbai and Bengaluru.

In 1965, Maithili was officially accepted by Sahitya Academy, an organization dedicated to the promotion of Indian literature. In March 2018, Maithili received the second official language status in the Indian state of Jharkhand.

Presently Maithili language is predominately written in the Devanagari. Mithilakshar Script is also in practice. Both the Scripts are Left to Right scripts which are descendent of *Brahmi* script. The Devanagari

script is also used to write several other languages of India and neighboring countries such as Nepal. The dataset prepared for LDC-IL Maithili Speech data is in Devanagari script.

Many ethnolects, sociolects and other varieties of Maithili are in practice in the Mithilanchal area, which are very different from each other in terms of phonological nuances and structural features. These divisions throw light on variation in the spoken language from core to regional boundary. Accordingly, the purity in pronunciation also varies.

Maithili has many social dialects in India and Nepal, to name a few, Dehati, Kisan, Bantat, Barmeli, Musar, Tati, Kortha and Jolaha. All these dialects are intelligible to native Maithili speakers.

Maithili varies greatly in geographic dialects. The standard form of Maithili is Sotipura or Central Maithili or Madhubani dialect which is mainly spoken in Darbhanga and Madhubani districts in Bihar, Indian.

Bajjika dialect of Maithili is spoken in Samastipur, Sitamarhi, Muzaffarpur, Vaishali East Champaran and West Champaran districts of Bihar in India. Bajjika is listed as a distinct language in Nepal and overlaps by 76-86% with Maithili dialects spoken in Dhanusa, Morang, Saptri and Sarlahi districts.

Thēthi dialect is spoken mainly in Kosi, Purnia, and Munger divisions of Bihar, India and some adjoining districts of Nepal.

The Chika-Chiki dialect of the south of the Ganga, specially of the districts of Bhagalpur, Monghyr, and Santhal Parganas is descended from Maithili. It is the result of a well-marked dialect from its frequent use of the-syllable “Chhika”, which is based on which the conjugation of the verb substantive is conjugated.

LDC-IL divided Mithilnachal area of India into four geographical regions based on the regional dialects. They are namely Sotipura, Bajjika dialect, Thēthi dialect, Chika-Chiki. The speech data is collected from three regions. Maithili Speech corpus is collected from Darbhanga, Madhubani for Sotipura Dialect; Madhepura, Purnia, Saharsa for Thēthi dialect and Samastipur for Bajjika Dialect. The data is collected from the native speakers of Mithilanchal area with Maithili as their mother tongue

26.2 DATASET PREPARATION FOR MAITHILI

For the selected Sotipura, Bajjika dialect, Thēthi Regions, 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	187
Most Frequent Words	1000
Person Name	500
Place Name	324
Sentences	208

Table 26-1: LDC-IL Speech Dataset

Distinct News Items were prepared to get the audio recording of contemporary text along with the aforementioned content types.

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
News Text	1 Text	Distinct Text
Created 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 26-2: Table of Contents in LDC-IL Dataset

Once all these preparations were made, the investigator started collecting the data. The Collection of data is carried out in two phases:

Region	Year	Field Investigator
Sotipura	2008	Savita Kiran
Bajjika	2008	Savita Kiran
Thēthi	2012	Arun Kumar Singh

Table 26-3: Four Phases of Speech Data Collection

26.3 TRANSLITERATIONS IN LDC-IL MAITHILI READ CORPUS

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

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

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

Vowels and Vowel Signs														
अ	आ	इ	ई	उ	ऊ	ऋ	ए	ऐ	ऑ	ओ	औ			
	ा	ि	ी	ु	ू	ृ	े	ै	ॉ	ो	ौ	ं	ः	ँ
a	A	i	I	u	U	x	e	ai	ao	o	au	M	H	m'

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

य	र	ल	व	श	ष	स	ह	ड़	ढ़
ya	Ra	la	va	sha	Sa	sa	ha	D	Dh

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

26.4 SUMMARY OF THE CORPUS

In the sections below, we provide the tabular details of the different content types of the Maithili 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 71:26:42 (hh:mm:ss) comprising 35109 audio segments.

26.4.1 Summary of the Audio Segments

The total number of Audio Segments and their distribution in the Maithili speech dataset is shown below.

LDC-IL Kannada 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	290	15	91	39	12	90	43
Creative Text-T2	294	14	91	40	16	90	43
Sentence-S	7449	371	2331	1002	400	2270	1075
Date-D	584	30	185	78	31	176	84
Command and Control Words-W1	8924	450	2785	1197	480	2725	1287
Person Name-W2	5917	300	1840	799	320	1817	841
Place Name-W2	2952	150	910	400	160	907	425
Most Frequent Word-Part-W3A	8699	442	2739	1162	470	2683	1203

Table 26-4: Maithili Audio Segments and their Distribution

26.4.2 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 Kannada 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	22:29:21	0:55:52	6:52:27	2:50:53	0:51:38	7:32:24	3:26:07
Creative Text-T2	15:34:55	0:46:57	4:58:28	1:59:31	0:51:42	4:35:05	2:23:12
Sentence-S	8:09:30	0:20:31	3:53:55	0:20:31	0:22:55	2:10:36	1:01:02
Date-D	0:31:38	0:01:30	0:09:48	0:03:45	0:01:38	0:09:55	0:05:02
Command and Control Words-W1	7:07:33	0:20:57	2:11:52	0:57:27	0:21:20	2:15:22	1:00:35
Person Name-W2	7:49:32	0:23:46	2:33:25	1:04:08	0:23:00	2:23:50	1:01:23
Place Name-W2	2:47:50	0:08:15	0:53:25	0:22:27	0:08:09	0:52:45	0:22:49
Most Frequent Word-Part-W3A	6:56:23	0:20:03	2:13:10	0:53:48	0:20:55	2:13:30	0:54:57

Table 26-5: Duration of the Maithili Speech Data

26.5 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

26.5.1 The Contemporary Text (News)- T1

Distinct Text Extracts from Newsapers 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					
				SOTIPURA		BAJJIKA		THETHI	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	27	15	12	2	2	5	2	8	8
21 To 50	181	91	90	26	20	19	27	46	43
50+	82	39	43	5	14	13	5	21	24
Total	290	145	145	33	36	37	34	75	75

Table 26-6: Distribution of Maithili Contemporary Text (News) Data

26.6 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

26.6.1 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					
				SOTIPURA		BAJJIKA		THETHI	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	30	14	16	2	5	5	3	7	8
21 To 50	181	91	90	27	20	20	28	44	42
50+	83	40	43	5	14	14	5	21	24
Total	294	145	149	34	39	39	36	72	74

Table 26-7: Distribution of Maithili Creative Text

26.6.2 The Date-D

The answer to one randomly selected question from the list of 2 questions is collected, 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					
				SOTIPURA		BAJJIKA		THETHI	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	61	30	31	4	10	10	5	16	16
21 To 50	361	185	176	52	41	41	51	92	84
50+	162	78	84	10	26	26	10	42	48
Total	584	293	291	66	77	77	66	150	148

Table 26-8: Distribution of Date Format

26.6.3 The Sentences-S

The sentences content type contains a list of sentences that is a representation of all most all the phonemes occurring in Maithili. 25 Randomly selected Sentences are recorded from a list of 208 sentences. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution					
				SOTIPURA		BAJJIKA		THETHI	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	771	371	400	50	125	122	75	199	200
21 To 50	4601	2331	2270	659	525	524	670	1148	1075
50+	2077	1002	1075	128	351	349	125	525	599
Total	7449	3704	3745	837	1001	995	870	1872	1874

Table 26-9: Distribution of Sentences

26.6.4 Command and Control Words-W1

The command and control words content type contains a list of 187 words that is a representation of all most all the command and control words occurring in Maithili. 30 randomly selected words are recorded from a list of words. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution					
				SOTIPURA		BAJJIKA		THETHI	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	930	450	480	60	150	150	90	240	240
21 To 50	5510	2785	2725	780	629	629	810	1376	1286
50+	2484	1197	1287	150	423	418	145	629	719
Total	8924	4432	4492	990	1202	1197	1045	2245	2245

Table 26-10: Distribution of Command and Control Words

26.6.5 Person Names –W2

The person name contains a list of 500 popular Pan Indian and regional person name. 20 randomly selected names are recorded from a list of names. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution					
				SOTIPURA		BAJJIKA		THETHI	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	620	300	320	40	100	100	60	160	160
21 To 50	3657	1840	1817	500	420	420	537	920	860
50+	1640	799	841	100	280	279	81	420	480
Total	5917	2939	2978	640	800	799	678	1500	1500

Table 26-11: Distribution of Personal Names

26.6.6 Place Names-W2

The place name contains a list of 324 popular Pan Indian and regional place name. 10 randomly selected names are recorded from a list of names. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution					
				SOTIPURA		BAJJIKA		THETHI	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	310	150	160	20	50	50	30	80	80
21 To 50	1817	910	907	250	210	201	267	459	430
50+	825	400	425	50	144	140	41	210	240
Total	2952	1460	1492	320	404	391	338	749	750

Table 26-12: Distribution of Place Names

26.6.7 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 a list of names. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution					
				SOTIPURA		BAJJIKA		THETHI	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	912	442	470	58	142	144	88	240	240
21 To 50	5422	2739	2683	728	606	632	789	1379	1288
50+	2365	1162	1203	148	354	384	127	630	722
Total	8699	4343	4356	934	1102	1160	1004	2249	2250

Table 26-13: Distribution of Most Frequent Words

26.7 NATIVE SPEAKERS DISTRIBUTIONS OF MAITHILI

Age Group	Total Speakers	Gender-wise Distribution		Region-wise Distribution					
				SOTIPURA		BAJJIKA		THETHI	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	31	15	16	2	5	5	3	8	8
21 To 50	186	94	92	27	21	21	28	46	43
50+	83	40	43	5	14	14	5	21	24
Total	300	149	151	34	40	40	36	75	75

Table 26-14: Distribution of Maithili Native Speakers

26.8 REFERENCE

Brass, P. R. (2005). *Language, Religion and Politics in North India*. Lincoln: iUniverse. ISBN 0-595-34394-5.

Baleshwar Thakur, George Pomeroy, Chris Cusack, Sudhir K Thakur. *City Society and Planning*. Volume (II), pp-429, ISBN-10:81-8069-460-7

Chaudhary, Radhakrishna. (2010). *A Survey of Maithili Literature*. (pp-12) ISBN: 978-93-80538-36-5;

Grierson, George Abraham. *Seven Grammars of the Dialects and Sub-dialects of the Bihari Language.* (1883–87). ISBN 81-7835-451-9

27 MALAYALAM RAW SPEECH CORPUS

Rejitha K.S, Saritha S.L, Sajila S, Rajesha N, Manasa G, Narayan Choudhary, L. Ramamoorthy

27.1 INTRODUCTION

On July 1, 1949, Travancore and Kochi joined to form the unified Travancore-Cochin state. But Kerala continued to be politically divided till the 1950s, even with the geographic similarities and solidarity of language. On 1 January 1950, Travancore-Cochin was recognised as a state. On 1 November 1956, the state of Kerala was formed by the States Reorganisation Act merging the three distinct areas such as Malabar district, Cochin and Travancore and taluk of Kasargod which is in South Canara. Four southern taluks like Thovala, Agastheswaram, Kalkkulam and Vilavancode separated from Travancore-Cochin which was merged with Tamilnadu.

Malayalam is the official language of Kerala and Laccadive Islands. It belongs to the Dravidian language family. Malayalam is closely related to Tamil and it is more influenced by Sanskrit than Tamil. After Independence, the state governments of Kerala started using regional languages more and more in administration. Greater use of Malayalam has contributed to the growth of the language in terms of vocabulary and the number of styles and registers.

Language is the collection of more or less similar idiolects. The fundamental fact about language is its diversity. Change in language is found when we move from country to country, region to region, class to class and caste to caste. Bloomfield (1933) says that linguistic diversity is related to the density of communication or to the amount of verbal interaction among speakers. In India dialect studies in a broad sense have been initiated by G.A. Grierson, who collected evidences to understand the linguistic situation in India and to group the regional dialects into families of Language such as the Austric, Tibeto - Chinese, lndo European and Dravidian.

Dialect variation in a language is not random but systematic. There are two types of dialects; regional dialects and social dialects. Regional dialects are geographically based and social dialects originate among social groups, class ethnicity, religion etc.

Language variation reflects the language change over time and people who live in the same geographical area or maintain the same social identity share language norms. Language change happens through three parameters like spatial, temporal and social. People never speak the same way in all time. They exploit the nuances for different purposes. People of different social classes, different occupations or different cultural groups in the same community will show variations in speech. People of different occupation have their own dialects and they use their own technical terms for better understanding. Education brings a greater difference in language style. History has contributed its own compliments to language. During wars people acquire words used by military people and in course of time these words spread through generations. Dialect variation is also due to political reasons like people are tried to preserve the dialects of their ruling kings. It correlated with geographical factors such as un-bridged rivers, impenetrable

forests, valleys, mountains, deserts etc. Marshals and artificial political barriers divide speech communities.

Language variation is due to different internal factors like semantics, vocabulary, grammar, phonological features, intonation patterns etc. along with other external factors region, cast, religion, education, occupation, social stratum, style, register etc. Various levels of linguistic structures show variations in different regional varieties of a language. Different groups of people who are living in two different areas show considerable differences in their language patterns. Malayalam spoken by any group of the northern region and that of the southern region shows that significant features are almost uniform for any group of the respectable regions. There are many lexical items with purely regional connotations and the same forms in two areas have two different meanings and also there are forms which are considered as taboos in one region, may not be perceived as same in another region. Similarly certain verbs and nouns have co-occurrence restrictions at regions.

All dialects of a language are equally efficient and expressive. In the case of Malayalam the socio economic and political status of the speech community has nothing to do with the standardization of the dialect. Irrespective of the socio-economic factors, all people use both the high and low varieties of Malayalam for different purposes. In Malayalam speech community, more of the lexical codes of the regional and caste dialects interfere with standard Malayalam.

Kerala formerly into Travancore, Cochin and Malabar resulted in lesser communication among the people of these three kingdoms. The sea separates Lakshadweep from Kerala and to lesser communication resulted in a distinct Malayalam dialect called Lakshadweep Malayalam. The Malayalam dialects show both regional and social variations and sometimes they overlap.

Malayalam has a number of social varieties depending on caste and religion. There are differences in the speech of Christians, Hindus, and Muslims within a single geographic area. The language of the high castes in Hindus is more influenced by Sanskrit than the language of the lower castes. Dialects spoken by Christians have more loan words from Portuguese, Latin, and English than other dialects. Dialects spoken by the Muslim population have many borrowing words from Arabic and Urdu. Lots of lexical items and many idiomatic expressions in modern Malayalam are of English origin. The influence of some other languages like Tamil, Prakrit, Pali, Marathi, Hindi, Persian, Dutch and French can be seen in the course of its evolution and transformation. In Kasargod, people are using 7 other languages such as Tulu, Kannada, Beary, Konkani, Urdu, Hindi, Gujarati, Marathi and Tamil along with Malayalam.

The existence of Southern, Central and Northern dialects of Malayalam is fairly obvious even to common people. But under each broad group there are a number of sub varieties, as shown by the dialect survey of Malayalam (Ezhava-Tiyya, Subramaniam 1974) which identified 12 dialect areas of Malayalam spoken by Ezhavas and Tiyyas. In all probability the other caste dialects also closely follow the geographical stratification found in Ezhava-Tiyya dialects. In other words, the

twelve dialect areas identified by the survey can be generalized to include all Malayalam regional dialects. The twelve areas are (1) South Travancore, (2) Central Travancore, (3) West Vempanad, (4) North Travancore, (5) Cochin, (6) South Malabar, (7) South Eastern Palghat, (8) North Western Palghat, (9) Central Malabar, (10) Wayanad, (11) North Malabar and (12) Kasargod. Another significant dialect is the Malayalam spoken in Lakshadweep which is not connected to the dialect of the mainland.

According to the formation of Kerala and the language of Travancore, Cochin and Malabar regions are influenced by different internal and external factors so LDC-IL considered Malayalam has three specifically different varieties, thus collected speech data from Thiruvananthapuram, Ernakulam and Kozhikode.

LDC-IL divided the Malayalam speaking areas into these three regions and collected speech data from each. After determining the regions for fieldwork, the datasets were prepared for each region.

27.2 DATASET PREPARATION FOR MALAYALAM

LDC-IL collected the Malayalam speech data using two approaches. Two different kind of Dataset Models were prepared as follows

- Dataset preparation Model 1 (T1, T2, W1, W2, W3, W4, W5, S, D)
- Dataset preparation Model 2 (Distinct Texts of T1 and T2)

27.2.1 Model 1 (T1, T2, S, D, W1, W2, W3, W4, W5)

For the Regions of Thiruvananthapuram and Ernakulam LDC-IL prepared the following dataset by which the prompt sheets were prepared.

Notation	Content Type	Count
T1	News	300
T2	Created Text	6
D	Date	2
S	Sentences	303
W1	Command and Control Words	58
W2	Person Name	599
W2	Place Name	300
W3	Most Frequent Words	1000
W4	Phonetically Balanced Words	518
W5	Form and Function Words	545

Table 27-1: Representation of Model 1 Content Type

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
News Text	1 Text	Distinct Text
Created 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 27-2: Representation of Model 1 Prompt Sheet

The full set of

13. Phonetically Balanced Words
14. Form and Function Words
15. 1000 Most Frequent Words

were also carried to the field to get recorded by selected individuals.

27.2.2 Model 2 (T1, T2)

For the fieldwork of Kozhikode LDC-IL attempted a different approach of dataset preparation, and tried to concentrate on capturing more distinct continues text recordings rather than word segments. The prompt sheet for Kozhikode were prepared as follows.

Content Type	Content in Each typical prompt sheet	Content selection type
News Text	1Text	Distinct Text
Created Text	1 text	Distinct Text

Table 27-3: Representation of Model 2 Prompt Sheet

Once all these preparations are made, the investigator started collecting the data. All the speakers who provided their recordings Malayalam Speech Corpus to LDC-IL are native speakers of Kerala and of Malayalam as their first language.

27.3 DATA COLLECTION DETAILS

The Collection of data is carried out in three phases for different regions as follows.

Region/Place	Year of data collection	Resource Person
Thiruvananthapuram	2008-09	Saritha S.L.
Ernakulam	2009-10	Saritha S.L.

Kozhikode	2012-13	Rejitha K.S. & Midhun P.G.

Table 27-4: Representation of Data Collection Details

27.4 TRANSLITERATIONS IN LDC-IL MALAYALAM READ CORPUS

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

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

LDC-IL Transliteration Schema

Malayalam characters to Roman and Malayalam Numerals to Hindu-Arabic

Vowels															
അ	ആ	ഇ	ഈ	ഉ	ഊ	ഋ	ഌ	ൠ	ൡ	എ	ഏ	ഐ	ഒ	ഓ	ഔ
	ാ	ി	ീ	ു	ൂ	ൃ	ൢ	ൄ	ൣ	െ	േ	ൈ	ൊ	ോ	ൌ
A	A	i	I	u	U	x	X	q	Q	e	E	ai	o	O	au

Consonants				
ക	ഖ	ഗ	ഘ	ങ
Ka	kha	ga	gha	ng'a
ച	ഛ	ജ	ഝ	ഞ
Ca	cha	ja	jha	nj'a
ട	ഠ	ഡ	ഢ	ണ
Ta	Tha	Da	Dha	Na
ത	ഥ	ദ	ധ	ന
Ta	tha	da	dha	na
പ	ഫ	ബ	ഭ	മ
Pa	pha	ba	bha	ma

Symbols			
◌ം	◌ഁ	ം	ഃ
m`	m'	M	H

യ	ര	ല	വ	ശ	ഷ	സ	ഹ	ള	ഴ	റ	ഩ
Ya	ra	la	va	sha	Sa	sa	ha	La	Za	Ra	TTTa

Chillu Letters								
ൺ	ൻ	ർ	ൽ	ൾ	ൿ	ൔ	ൟ	ൎ
N'	n'	R'	l'	L'	k'	M'	y'	Z'

Numerals (Malayalam to Hindu-Arabic)									
൦	൧	൨	൩	൪	൫	൬	൭	൮	൯
0	1	2	3	4	5	6	7	8	9

The greyed out characters are obsolete. They may rarely present in the current LDC-IL corpus.

27.5 SUMMARY OF THE CORPORA

In the sections below, we provide the tabular details of the different content types of the Malayalam 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.

27.5.1 Summary of the Utterances

The table below shows the total number of utterances and their distribution in the Malayalam speech dataset.

LDC-IL Malayalam 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	51	123	53	48	122	52
Creative Text-T2	449	51	124	52	48	122	52
Date-D	598	26	172	106	22	168	104
Sentence-S	7452	323	2146	1342	275	2081	1285
Command and Control Words-W1	8923	388	2559	1608	330	2501	1537
Person Name-W2	5819	259	1679	1032	219	1603	1027
Place Name-W2	2906	128	826	515	110	810	517
Most Frequent Word-Part-W3A	8763	387	2509	1570	330	2425	1542
Most Frequent Word-FullSet-W3B	1979	0	990	0	0	989	0
Phonetically Balanced-W4	3096	0	1552	0	0	1029	515
Form and Function Word-W5	3236	0	2158	0	0	1078	0

Table 27-5: Representation of Audio Segments of Malayalam Raw Speech Data

27.5.2 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 Malayalam 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)	71:29:21	10:02:57	18:18:49	07:04:59	09:48:49	18:38:36	07:35:11
Creative Text-T2	54:41:20	11:37:21	14:27:14	02:16:48	10:22:47	13:38:40	02:18:30
Date-D	00:53:45	00:02:12	00:15:10	00:09:51	00:01:42	00:14:38	00:10:12
Sentence-S	06:56:46	00:19:09	01:59:41	01:18:50	00:15:48	01:51:56	01:11:22
Command and Control Words-W1	07:09:37	00:18:56	02:07:02	01:18:35	0:13:05	02:00:16	01:11:43
Person Name-W2	05:26:33	00:14:45	01:37:40	01:00:15	00:10:29	01:27:31	00:55:53
Place Name-W2	02:28:24	00:06:50	00:43:10	00:27:42	00:04:52	00:40:24	00:25:26
Most Frequent Word-Part-W3A	06:51:31	00:18:51	02:01:13	01:16:18	00:12:40	01:52:08	01:10:21
Most Frequent Word-FullSet-W3B	02:08:58	00:00:00	00:56:10	00:00:00	00:00:00	01:12:48	00:00:00
Phonetically Balanced-W4	02:40:09	00:00:00	01:24:12	00:00:00	00:00:00	00:59:27	00:16:30
Form and Function Word-W5	03:14:38	00:00:00	02:12:37	00:00:00	00:00:00	01:02:01	00:00:00

Table 27-6: Representation of Duration of Malayalam Content

27.6 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.

27.6.1 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	Gender-wise Distribution of text		Region-wise Distribution					
				Thiruvananthapuram		Ernakulam		Kozhikode	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	99	51	48	5	1	8	10	38	37
21 To 50	245	123	122	43	37	42	47	38	38
50+	105	53	52	30	35	23	17	0	0
Total	**449**	227	222	78	73	73	74	76	75

Table 27-7: Distribution of Malayalam Contemporary Text (News)

27.6.2 Creative Text

Distinct Text Extracts from literary books are recorded from the informants to get the speech data of literary text. These types of distinct creative text are collected from Kozhikode Region. The distribution of data is as follows:

Age Group	Total Audio Segments	Kozhikode Region Distribution	
		Female	Male
16 To 20	75	38	37
21 To 50	76	38	38
Total	151	76	75

Table 27-8: Distribution of Malayalam Creative Text in Distinct Set

27.7 RANDOM SET

The Random Set data composes 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:

27.7.1 Creative Text

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			
				Thiruvananthapuram		Ernakulam	
		Female	Male	Female	Male	Female	Male
16 To 20	24	13	11	5	1	8	10
21 To 50	170	86	84	44	37	42	47
50+	104	52	52	29	35	23	17
Total	298	151	147	76	75	73	74

Table 27-9: Distribution of Malayalam Creative Text in Random Set

27.7.2 Date

Each informant answers two questions to get the date format. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution			
				Thiruvananthapuram		Ernakulam	
		Female	Male	Female	Male	Female	Male
16 To 20	48	26	22	10	2	16	20
21 To 50	340	172	168	88	74	84	94
50+	210	106	104	60	70	46	34
Total	598	304	294	158	146	146	148

Table 27-10: Distribution of Malayalam Date Format

27.7.3 Sentences

The sentences content type contains a list of sentences that is a representation of almost all the phonemes occurring in Malayalam. 25 Randomly selected Sentences are recorded from a list of 303 sentences. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender wise Distribution		Region-wise Distribution			
				Thiruvananthapuram		Ernakulam	
		Female	Male	Female	Male	Female	Male
16 To 20	598	323	275	123	25	200	250
21 To 50	4227	2146	2081	1096	907	1050	1174
50+	2627	1342	1285	768	860	574	425
Total	7452	3811	3641	76	75	1824	1849

Table 27-11: Distribution of Malayalam Sentences

27.7.4 Command and Control Words

The command and control words content type contains a list of 58 words that is a representation of almost all the command and control words occurring in Malayalam. 30 randomly selected words are recorded from a list of words. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender wise Distribution		Region-wise Distribution			
				Thiruvananthapuram		Ernakulam	
		Female	Male	Female	Male	Female	Male
16 To 20	718	388	330	149	30	239	300
21 To 50	5060	2559	2501	1299	1093	1260	1408
50+	3145	1608	1537	918	1027	690	510
Total	8923	4555	4368	2366	2150	2189	2218

Table 27-12: Distribution of Malayalam Command and Control Words

27.7.5 Person Names

The person name contains a list of 599 names. 20 randomly selected names are recorded from a list of names. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender wise Distribution		Region-wise Distribution			
				Thiruvananthapuram		Ernakulam	
		Female	Male	Female	Male	Female	Male
16 To 20	478	259	219	99	20	160	199
21 To 50	3282	1679	1603	834	679	845	924
50+	2059	1032	1027	592	687	440	340
Total	5819	2970	2849	1525	1386	1445	1463

Table 27-13: Distribution of Malayalam Person Name

27.7.6 Place Names

The place name contains a list of 300 popular Pan Indian and regional place name. 10 randomly selected names are recorded from a list of names. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution			
				Thiruvananthapuram		Ernakulam	
		Female	Male	Female	Male	Female	Male
16 To 20	238	128	110	48	10	80	100
21 To 50	1636	826	810	406	348	420	462
50+	1032	515	517	295	347	220	170
Total	2906	1469	1437	749	705	720	732

Table 27-14: Distribution of Malayalam Place Names

27.7.7 Most Frequent Words - Part

The most frequent words-part contains a list of 1144 most frequent words. 30 randomly selected words are recorded from a list of words. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution			
				Thiruvananthapuram		Ernakulam	
		Female	Male	Female	Male	Female	Male
16 To 20	717	387	330	148	30	239	300
21 To 50	4934	2509	2425	1266	1035	1243	1390
50+	3112	1570	1542	881	1033	689	509
Total	8763	4466	4297	2295	2098	2171	2199

Table 27-15: Distribution of Malayalam Most frequent Words Part

27.8 FULL SET

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 below:

27.8.1 Most Frequent Words - Full

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

Total Audio Segments from 21-50 Age group	Gender-wise Distribution		Gender-wise Distribution			
			Thiruvananthapuram		Ernakulam	
	Female	Male	Female	Male	Female	Male
1979	1979	0	990	0	989	0

Table 27-16: Distribution of Malayalam Most Frequent Words = Full

27.8.2 Phonetically Balanced Words

The phonetically balanced words contain a list of words where almost all the phones of Malayalam language have occurred in all the possible positions of a word. In full set all the 518 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 of words in Thiruvananthapuram	
		Female	Male
21 To 50	2581	1552	1029
50+	515	0	515
Total	**3096**	1552	1544

Table 27-17: Distribution of Malayalam Phonetically Balanced Words

27.8.3 Form and Function Words

The form and function words content type contains a list of 545 words that is a representation of almost all the form and function words occurring in Malayalam. All the words are recorded from the informant where they uttered those words three times. The distribution of data is as follows:

Total Words three utterance each from Speakers of 21-50 Age group	Gender-wise Distribution of words in Thiruvananthapuram	
	Female	Male
3236	2158	1078

Table 27-18: Distribution of Malayalam Form Function Words

27.9 NATIVE SPEAKERS DISTRIBUTIONS

For Malayalam speech data a total of 458 speakers were collected in which 231 female speakers and 227 male speakers from three 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					
				Thiruvananthapuram		Ernakulam		Kozhikode	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	98	50	48	5	1	8	10	38	37
21 To 50	253	127	126	52	42	42	47	38	38
50+	107	54	53	31	36	23	17	0	0
Total	**458**	231	227	82	78	73	74	76	75

Table 27-19: Distribution of Malayalam Native Speakers

28 MANIPURI RAW SPEECH CORPUS

Amom Nandaraj Meetei, Yumnam Premila Chanu, Rajesha N, Manasa G, Narayan Choudhary, L. Ramamoorthy

28.1 INTRODUCTION

India, one of the most linguistically diverse countries, has five language families, namely – the Indo-Aryan, the Dravidian, the Austro-Asiatic, the Tibeto-Burman and the Andamanes respectively. It is in good health that the language policy of India is elucidated in the Constitution, implemented through various executive orders issued from time to time as well as the judicial pronouncement since 1950 focusing on the scope of being language-development oriented and language-survival oriented. In this parlance was the inclusion of the Eighth Schedule in the constitution providing formal and constitutional recognition to dominant regional languages in the sphere of administration, education, economy and social status.

28.1.1 Manipuri in the Eighth Schedule of the Constitution of India

Manipuri languages obtained its due place and recognition in the Constitution of India being included in the Eight Schedule, according to the Seventy First Amendment of the Constitution, on the 20th August, 1992. Here it is noteworthy to mention that Manipuri language, the state official language in Manipur, is also the first Tibeto-Burman (henceforth TB) language included in the said Eighth Schedule of the Constitution of India. Another mentionable point here is that it is the only language amongst the TB-languages in India which has its own scripts and written literature.

Linguistically speaking, Manipuri is the lingua-franca, and also an Inter-Tribal language amongst the speakers of different dialects and other minor languages inhabiting both the Hill and Valley areas of the state of Manipur.

Adopted as the medium of instruction and examination from Primary School level to College and University level, Manipuri has been offered as a subject of study and research not only by the Meetei native speakers alone but also by other community groups in the state of Manipur.

28.1.2 Area and Population of Manipur

Manipur has area of 22,327 sq. kms. According to census of 1981, the biggest valley area of the state, now known as the Imphal Valley is about 1,843 sq. kms which is roughly 9% of the total area of the entire state. Physically, Manipur comprises of two parts, the hills and the valley. The valley lies at the centre surrounded by hills on all sides. The hills cover about 9/10 of the total area of the State. The State has a population of almost 3 million, including the Meeteis, who are the majority ethnic group in the state and other ethnic community groups who speak a variety of Tibeto-Burman languages. In fact, the term Meetei is an endonym or autonym while Manipuri is an exonym. Manipur lies between latitude 23050' and 25030' North, and longitude 93010' and 94030' East. Having an oval shape area on the basis of geographical position of the state on the surface of the Earth, longer in north and south, and shorter in east and west, in length, the state

enjoys mild sub-tropical temperate monsoon climate with temperature varying from a little above 00C and below 350C.

28.1.3 A Cursory Glance at Manipuri Dialects

It is a generally held view that there was no particular language variety called Meeteilon (now known as Manipuri in the constitution) in its pre-historic period. What is believed so far is that there were different varieties in the form of dialects of the same speech spoken and used in different parts of the state.

Nevertheless, there is a mutual intelligibility between these dialects in which speakers of different but related dialects can readily understand each other without prior familiarity or any special effort. Gradually, a particular variety came into existence taking its certain uniform form along with typical linguistic features. This particular variety is the standard Meeteilon/Manipuri, which is the byproduct of different linguistic forces and different dialectal elements contributed to it. Henceforth, the dialect of the Kangla Imphal became its base while other dialects of the same speech continued to exhibit shares in it. Today, with cycle of time, we can observe that the intelligibility lying between Imphal dialect and other dialects such as Kakching, can be asymmetric in such a way that speakers of one dialect, say Kakching dialect, happen to understand more of the other, the Imphal dialect, than the speakers of Imphal understanding Kakching dialect.

Manipuri (locally known as Meeteilon) has dialects of Kakching, 45 kms far away from Imphal, located in the southeastern part of the state, Awang Sekmai, 17 kms from Imphal, located western part of the state, Andro, 25 kms from Imphal, located in the eastern part of the state, Phayeng, 16 kms from Imphal, located in the western part of the state, Kwatha, 102 kms from Imphal, located in the eastern most part of the state, Thangga, 48 kms from Imphal, located in the westen part of the state, etc. In this way, there is a dialect continuum or dialect chain for this Manipuri language typically occurring in a long-settled Meetei population. Standard Imphal variety within such a dialect continuum was developed and codified serving as an authority for part of it across the various geographical areas of the monolingual community. Imphal dialect is now used for official purposes, heard on radio and televisions and considered the standard form of their speech so that any standardizing changes in their speech are always oriented towards that Imphal variety. In these apparent cases, other local dialectal varieties are said to be dependent on, or heteronomous with respect to, the standard Imphal variety. This is how the formation of dialects took place and Manipuri has a standard Imphal variety, together with its dependent varieties called “dialects” of the language even though this standard variety is mutually intelligible with the other rest of the dialectal forms from the continuum.

28.1.4 Intra-regional speech variation and LDC-IL dataset

Language variation exists even in monolingual communities. The sociolinguistic elements such as social status, gender, age and ethnicity etc. happen to get reflected in the language people speak and happen to turn out to be important dimensions of the speakers’ identity in their community. Every dialect has its own unique linguistic features which the group shares with each other within the small group. It is a fact that no two people speak exactly the same exhibiting infinite source of variation in their speeches. One can notice that a sound spectrograph shows that even a single vowel could be pronounced in hundreds of minutely different manners, most of which the listeners cannot even register. However, there are certain common features of

speech which each dialect exhibits and this feature is shared by the group concerned differentiating them from the other group which again has its own common one. In the present scenario of Manipuri dialectal variations, the pronunciation, grammar, and vocabulary of Kakching or Awang Sekmai speakers of Manipuri are in some respects found quite distinct from that of people from Imphal. Since the standard Imphal dialect is always the first to be codified, the act of defining other dialects is done through contrasting them with the standard. In this way, one can capture how Kakching or Awang Sekmai dialect features contrast with the standard Imphal dialect features. In this perlance, certain linguistic features identifying regional tones and intonations, phonemic distributions (as observed in the dataset of Phonetically Balanced vocabulary-W4 and Phonetically Balanced Sentences-S, various pronunciations reflected in both regional and non-regional vocabulary items such as person names and place names (as observed in the dataset of Person names W2 and Place Names W2) etc., have been well housed based on a standard parameter of LDC-IL dataset. The LDC-IL Manipuri Raw Speech Corpus reflects the speech varieties of the same language providing the speakers' social background. In a nutshell, such speech corpus can be empirically used for developing NLP tools such as speech synthesis, speech recognition, spoken language systems, and speaker recognition /verification, etc. which generally employs algorithms working with acoustic and language modeling. On the other hand, this speech corpus can also be used for research purposes such as phonetic research, sociolinguistic research, psycholinguistics research and other language acquisition investigation purposes. Since the same dataset is applied to all the said regional dialects, vocabulary and grammar cannot be distinguished from each other; however, there are clear-cut distinctions in pronunciations observed amongst the respective dialects.

In its real sense, such language-specific resources like Manipuri Raw Speech Corpus can be used for building Manipuri TTS systems since it can be viewed as providing (i) audio recordings (ii) pronunciation lexicon and (iii) phone sets containing phonetic features for each phoneme. Since the corpus is in its raw status, the researchers/experts can specify the transcripts to the recorded segments; give letter-to-sound rules and develop phone sets that contain phonetic features for each phoneme, which will finally be used as the specific features in models of the spectrum and prosody. In fact, this Manipuri Speech Corpus is the actual linguistic resources and data for such speech synthesis task.

As one can notice that the LDC-IL Manipuri Raw Speech corpus has a total number of 620 speakers (310 Female and 310 Male) collected from three regional dialects, namely Imphal, Kakching, and Awang Sekmai respectively, the whole database can be said to be consisting of speeches with various voice characteristics and speaking styles which are more feasible features for the speech synthesis to achieve maintaining some acceptable degree of naturalness.

28.2 DATASET PREPARATION FOR MANIPURI

For the selected dialects such as Imphal, Kakching and Awang Sekmai, LDC-IL prepared the following dataset for which the prompt sheets were prepared.

DATASET CONTENTS FOR MANIPURI

Content Type	Count
Created Text	16
Date	2
Command and Control Words	249
Most Frequent Words	1000
Person Name	501
Place Name	324
Sentences	200

Table 28-1: LDC-IL Speech Dataset

Distinct News Items are prepared to get the audio recording of contemporary text along with the aforementioned content types.

Each prompt sheet that consists of a distinct news item and selected part of the dataset is prepared as follows:

Content Type	Content that Each typical prompt sheet had	Content selection type
News Text	1 Text	Distinct Text
Created 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 28-2: Table of Contents in LDC-IL Dataset

Once all these preparations are made, the investigators start collecting the data. The collection of data is carried out in four phases:

Region	Year	Field Investigator
Imphal	2008	Amom Nandaraj Meetei
Kakching	2009	Amom Nandaraj Meetei
Awang Sekmai	2010	Amom Nandaraj Meetie
Imphal	2013	Yumnam Premila Chanu

Table 28-3: Four Phases of Speech Data Collection

28.3 TRANSLITERATIONS IN LDC-IL MANIPURI READ CORPUS

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

The LDC-IL transliteration scheme of Manipuri (in Bengali scripts) to Roman is given below.

LDC-IL Transliteration Schema

Manipuri characters to Roman and Manipuri Numerals to Hindu-Arabic

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

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

Unreleased Consonants		
ক্		ঙ্
k		ng'
ম্		প্
m		p
ন্	ত্	ল্
n	t	l

Bengali	য	র	ল	শ	স	ষ	হ	ড়	ঢ়	য়	ৎ
Roman	ya	ra	la	sha	Sa	sa	ha	D'a	Dh'a	Ya	t

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

The following citation visualizes the running idea.

ContentID: S-0001

Recorded Text: মহাক্না পুনগি অসঙেবা ইসঙি তপ্না থক্লা৷

Transliteration: mahAknA punagi aseng'abA ising'a tapnA thakli.

28.4 SUMMARY OF THE CORPUS

In the sections below, we provide the tabular details of the different content types of the Manipuri Raw Speech Corpus on the basis of various yardsticks being 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 providing useful insights into the dataset. The data size is of total duration 156:11:14 (hh:mm:ss) comprising 66,231 audio segments.

28.4.1 Summary of the Audio Segments

The total number of Audio Segments along with their distribution in terms of Gender and Age for the Manipuri Speech Dataset is shown below.

LDC-IL Manipuri 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	530	29	187	55	26	170	63
Creative Text-T2	588	29	194	72	34	187	72
Sentence-S	10979	600	3277	1600	600	3301	1601
Date-D	866	48	257	128	46	261	126
Command and Control Words-W1	13129	714	3928	1919	720	3928	1920
Person Name-W2	8789	481	2625	1280	480	2641	1282
Place Name-W2	4394	240	1311	640	241	1321	641
Most Frequent Word-Part-W3A	13167	722	3929	1920	720	3956	1920
Most Frequent Word-FullSet-W3B	6992	1000	1996	0	998	1998	1000
Phonetically Balanced-W4	4518	753	753	753	753	753	753
Form and Function Word-W5	2279	380	380	380	380	379	380

Table 28-4 : Manipuri Audio Segments and their Distribution

28.4.2 Duration of the Raw Speech Data

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

LDC-IL Manipuri 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	59:47:22	02:55:21	22:46:06	04:55:44	03:31:13	20:26:36	05:12:22
Creative Text-T2	53:59:35	02:15:42	20:02:06	04:13:38	03:29:04	19:14:49	04:34:06
Sentence-S	10:01:41	00:33:51	03:03:27	01:25:40	00:31:00	02:56:09	01:31:34
Date-D	01:12:04	00:04:20	00:21:03	00:10:27	00:04:00	00:21:55	00:10:19
Command and Control Words-W1	08:00:02	00:26:41	02:26:28	01:10:38	00:24:11	02:19:52	01:12:12
Person Name-W2	07:14:04	00:24:22	02:13:58	01:03:12	00:21:39	02:05:45	01:05:08
Place Name-W2	02:46:29	00:09:34	00:51:09	00:24:34	00:08:05	00:48:13	00:24:54
Most Frequent Word-Part-W3A	06:31:30	00:05:30	02:03:59	00:59:48	00:20:23	01:59:36	01:02:14
Most Frequent Word-FullSet-W3B	02:48:42	00:26:26	00:43:29	00:00:00	00:21:08	00:50:36	00:27:03
Phonetically Balanced-W4	02:25:55	00:27:41	00:19:54	00:28:09	00:26:29	00:19:45	00:23:57
Form and Function Word-W5	01:23:50	00:14:10	00:12:07	00:12:09	00:14:18	00:19:07	00:11:59

Table 28-5: Duration of the Collected Manipuri Speech Data

28.5 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.

28.5.1 The Contemporary Text (News)- T1

Distinct Text Extracts from Newsapers 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					
				IMPHAL		KAKCHING		SEKMAI	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	55	29	26	13	10	9	8	7	8
21 To 50	357	186	171	100	84	46	42	41	44
50+	118	55	63	13	19	21	24	21	20
Total	530	270	260	125	114	76	74	69	72

Table 28-6: Distribution of Manipuri Contemporary Text (News) Data

28.6 RANDOM SET

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

28.6.1 The Created Text-T2

One randomly selected text of literature out of 16 texts from the prepared dataset is recorded from the informants to get the speech data of Created Text. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution					
				IMPHAL		KAKCHING		SEKMAI	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	63	29	34	13	18	8	8	8	8
21 To 50	381	193	188	107	101	46	42	41	44
50+	144	72	72	30	28	21	24	21	20
Total	588	294	294	149	148	75	74	70	72

Table 28-7: Distribution of Manipuri Created Text:T2

28.6.2 The Sentences-S

The Sentences-content type consists of a list of sentences that can be representative of which all the phonemes in Manipuri can occur in various positions. 25 Randomly selected Sentences are recorded from a list of 208 sentences. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution					
				IMPHAL		KAKCHING		SEKMAI	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	1200	600	600	200	200	200	200	200	200
21 To 50	6578	3277	3301	1100	1150	1151	1051	1026	1100
50+	3201	1600	1601	550	500	525	601	525	500
Total	10979	5477	5502	1850	1850	1876	1852	1751	1800

Table 28-8: Distribution of Manipuri Sentences

28.6.3 The Date-D

The answer to one randomly selected question from the list of 2 questions is recorded 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					
				IMPHAL		KAKCHING		SEKMAI	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	94	48	46	16	16	16	14	16	16
21 To 50	518	257	261	88	92	87	82	82	87
50+	254	128	126	44	40	42	46	42	40
Total	866	433	433	148	148	145	142	140	143

Table 28-9: Distribution of Manipuri Date Format

28.6.4 Command and Control Words-W1

The command and control words content type consists of a list of 187 words that is a representative of which most of the command and control words occur in Manipuri. 30 randomly selected words are recorded from a list of words. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution	Region-wise Distribution		
			IMPHAL	KAKCHING	SEKMAI

		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	1434	714	720	235	240	240	240	239	240
21 To 50	7856	3928	3928	1320	1380	1379	1259	1229	1289
50+	3839	1919	1920	659	600	630	720	630	600
Total	13129	6561	6568	2214	2220	2249	2219	2098	2129

Table 28-10: Distribution of Manipuri Command and Control Words

28.6.5 Person Names –W2

The Person Names content type consists of a list of 500 popular Pan Indian and regional person names. 20 randomly selected names are recorded from a list of names. The distribution of data is as follows:

Age Group	**Total Audio Segments**	**Gender-wise Distribution**		**Region-wise Distribution**					
				IMPHAL		KAKCHING		SEKMAI	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	961	481	480	162	160	159	160	160	160
21 To 50	5266	2625	2641	880	920	920	839	825	882
50+	2562	1280	1282	440	400	420	481	420	401
Total	8789	4386	4403	1482	1480	1499	1480	1405	1443

Table 28-11: Distribution of Maniprui Personal Names

28.6.6 Place Names-W2

The Place Names content type consists of a list of 324 popular Pan Indian and regional place names. 10 randomly selected names are recorded from a list of names. The distribution of data is as follows:

Age Group	**Total Audio Segments**	**Gender-wise Distribution**		**Region-wise Distribution**					
				IMPHAL		KAKCHING		SEKMAI	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	481	240	241	80	80	80	80	80	81
21 To 50	2632	1311	1321	440	460	460	420	411	441
50+	1281	640	641	220	200	210	241	210	200
Total	4394	2191	2203	740	740	750	741	701	722

Table 28-12: Distribution of Manipuri Place Names

28.6.7 Most Frequent Words-PART-W3A

The Most Frequent Words-Part content type consists of a list of 1000 most frequent words. 30 randomly selected words are recorded from a list of such frequent words. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution					
				IMPHAL		KAKCHING		SEKMAI	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	1442	722	720	242	240	240	240	240	240
21 To 50	7885	3929	3956	1320	1379	1380	1254	1229	1323
50+	3840	1920	1920	660	598	630	722	630	600
Total	13167	6571	6596	2222	2217	2250	2216	2099	2163

Table 28-13: Distribution of Maipuri Most Frequent Words - Part

28.7 FULL SET

The Full Sets are the master set of certain data sets which are read completely by few selected speakers in each group. The full sets are as below:

28.7.1 Most Frequent Words-Full-W3B

The Most Frequent Words-Full content type consists of a list of 1000 most frequent words. In this case, all the 1000 words are recorded from the informants. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution			
				IMPHAL		KAKCHING	
		Female	Male	Female	Male	Female	Male
16 To 20	1998	1000	998	0	0	1000	998
21 To 50	3994	1996	1998	1000	1000	996	998
50+	1000	0	1000	0	0	0	1000
Total	6992	2996	3996	1000	1000	1996	2996

Table 28-14: Distribution of Manipuri Most Frequent Word-Full

28.7.2 The Phonetically Balanced Words-W4

The Phonetically Balanced Words type contains a list of words where most of the phones of Manipuri language have occurred in all the possible positions of a word. In this case, all the 390 words are recorded from the informants in such a way that they utter those words three times. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution			
				IMPHAL		KAKCHING	
		Female	Male	Female	Male	Female	Male
16 To 20	1506	753	753	374	374	379	379
21 To 50	1506	753	753	374	374	379	379
50+	1506	753	753	374	374	379	379
Total	4518	2259	2259	1122	1122	1137	1137

Table 28-15: Distribution of Manipuri Phonetically Balanced Words-W4

28.7.3 The Form and Function Words-W5

The Form and Function Words content type contains a list of 432 words which is a representative of which most of the form and function words occur in Manipuri. All the words are recorded from the informants in such a way that they utter those words three times. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution			
				IMPHAL		KAKCHING	
		Female	Male	Female	Male	Female	Male
16 To 20	760	380	380	189	189	191	191
21 To 50	759	380	379	189	189	191	190
50+	760	380	380	189	189	191	191
Total	2279	1140	1139	567	567	573	572

Table 28-16: Distribution of Manipuri Form and Function Words-W5

28.8 NATIVE SPEAKERS DISTRIBUTIONS

The following table displays the overall distributions of the native speakers who read the data during the fieldworks undertaken at different three regions of the state of Manipur.

Age Group	Total Speakers	Gender-wise Distribution		Region-wise Distribution					
				IMPHAL		KAKCHING		SEKMAI	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	74	35	39	15	20	12	11	8	8
21 To 50	393	199	194	109	104	49	45	41	44
50+	154	76	78	32	31	23	27	21	20
Total	620	310	310	156	155	84	83	70	72

Table 28-17: Distribution of Manipuri Native Speakers

28.9 CONCLUSION

This documentation is representative of contemporary speech corpus for Manipuri in response to the corpus generation revolution undertaken by the government of India for the development of Indian scheduled languages in technological media world which was initiated in the form of the technological development works on scheduled languages in 1991. LDC-IL has created Manipuri Raw Speech Corpus, which is the backbone of automatic speech recognition and also for different speech recognition tasks in Manipuri language. The database comprises of appropriate words, sentences and paragraphs spoken by the typical users in realistic acoustic or natural environments following the data collection guidelines of LDC-IL.

28.10 REFFERENCE

Amom, N.M., Sakkan Th, Sarma, A. 2013. CORPUS Building of Endangered Languages, in *Language Endangerement in South Asia*, (Vol-I), (eds.) by Ganeshan, M, et al, Centre of Advanced Study in Linguistics, Annamalai University, Annamalainagar.

Dash, N.S. 2005. *Corpus Linguistics and Language Technology*. New Delhi: Mittal Publications.

Dash, N.S. 2006. *Speech corpora Vs. Text Corpora: Need for Separate Development*. Indian Linguistics. Vol. 67. Nos. 1-4. Pp. 65-82.

Leech, G. 1991. "The state of the art in corpus linguistics". In, Aijmer, K. and B. Altenberg (Eds.) *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London: Longman. Pp. 8-29.

29 MARATHI RAW SPEECH CORPUS

Bhageshree Khandale, Rajesha N, Manasa G, Narayan Choudhary, L. Ramamoorthy

29.1 INTRODUCTION

Marathi language is an Indo-Aryan language. It is the official language of Maharashtra state of India. Marathi has some of the oldest literature of all modern Indo-Aryan languages, dating from about the 1st century AD. Marathi is primarily spoken in Maharashtra (India) and parts of neighboring states of Gujarat, Madhya Pradesh, Goa, Karnataka (Particularly the bordering districts of Belgaum, Bidar, Gulbarga and Uttara Kannada), union-territories of Daman and Diu and Dadra and Nagar Haveli. The former Maratha ruled cities of Baroda, Indore, Gwalior, Jabalpur and Tanjore have had sizable Marathi speaking populations for centuries. Marathi is also spoken by Maharashtrian emigrants to other parts of India and overseas.

There were 83 million native Marathi speakers in India, according to the 2011 census, making it the third most spoken native language after Hindi and Bengali. Native Marathi speakers form 6.86 % of India's population. Native speakers of Marathi formed 68.93% of the population in Maharashtra, 10.89% in Goa, 7.01% in Dadra and Nagar Haveli, 4.53% in Daman and Diu, 3.38% in Karnataka, 1.7% in Madhya Pradesh and 1.52% in Gujarat. The major dialects of Marathi are called Standard Marathi and Varhadi Marathi. There are a few other sub-dialects like Ahirani, Dangi, Vadvali, Samavedi, Khandeshi, Zadi Boli, Koli and Malvani. Standard Marathi (Puneri) is the official language of the State of Maharashtra. Standard Marathi is based on dialects used by academics and the print media. Marathi is thought to be a descendant of Maharashtri, one of the Prakrit languages which developed from Sanskrit.

Khandeshi is spoken in the Khandesh region, wedged between the territory of Bhili and that of Marathi. It consists of Khandeshi proper, and the Dangri and Ahirani dialects. Zadi Boli or Zhadiboli (झाडीबोली (is spoken in Zadipranta (a forest rich region) of far eastern Maharashtra or eastern Vidarbha or western-central Gondwana comprising Gondia, Bhandara, Chandrapur, Gadchiroli and some parts of Nagpur of Maharashtra. Zadi Boli Sahitya Mandal and many literary figures are working for the conservation of this important and distinct dialect of Marathi. Varhadi (Varhādi) (or Vaidharbhi is spoken in the Western Vidarbha region of Maharashtra. In Marathi, the retroflex lateral approximant is common, while in the Varhadii dialect, it corresponds to the palatal approximant, making this dialect quite distinct. Such phonetic shifts are common in spoken Marathi and, as such, the spoken dialects vary from one region of Maharashtra to another. Malvani is a dialect of Konkani with significant Marathi influences and loanwords. Though Malvani does not have a unique script, scripts of the other languages native to the regions its speakers inhabit are used. Devanagari is used by most of the speakers.

Although there are many scripts and languages in India but not much research work is done for handwritten Marathi characters. Marathi handwritten character recognition is the challenging task in the pattern recognition field. Marathi first appeared in writing during the 11th century in the form of inscriptions on stones and copper plates. Marathi was written in Modi script a cursive script minimizes the lifting of pen from paper while writing. Most writings of the Maratha

Empire are in Modi script. However, Persian based scripts were also used for court documentation.

With the advent of large scale printing, Modi script fell into disuse, as it proved very difficult for typesetting. Currently, due to the availability of Modi fonts and the enthusiasm of the younger speakers, and getting slots in Unicode slot the script is getting revived. Now, Marathi is written in the Devanagari script, a set consists of 16 vowels. The phoneme inventory of Marathi is similar to that of many other Indo-Aryan languages.

Many government and semi-government organizations exist which work for the regulation, promotion, and enrichment of the Marathi language. These are either initiated or funded by the government of Maharashtra. Marathi Language Day (Marathi Dina, Marathi Diwasa : is celebrated on 27 February every year across the Indian states of Maharashtra and Goa. This day is regulated by the State Government. Earlier Marathi suffered from weak support by computer operating systems and Internet services, as have other Indian languages. But recently, with the introduction of language localization projects and new technologies, various software and Internet applications have been introduced.

. LDC-IL divided the Marathi speaking areas into these four regions and collected speech data from each. After determining the regions for fieldwork, the dataset is prepared from which the prompt sheets were generated

Places from which LDC-IL Marathi Speech Data is collected in Each Region is listed in the table below:

Marathi Regional Dialect	District
GOA	South Goa
MARATHWADA	Ahmednagar
MARATHWADA	Aurangabad
MARATHWADA	Beed
MARATHWADA	Hingoli
MARATHWADA	Jalna
MARATHWADA	Latur
MARATHWADA	Nanded
MARATHWADA	Nashik
MARATHWADA	Osmanabad
MARATHWADA	Parbhani
PUNERI	Parbhani
PUNERI	Pune
VIDHARBH	Yavatmal

Table 29-1: Districts from LDC-IL collected Marathi Speech Data

29.2 DATASET PREPARATION FOR MARATHI

For the selected Regions, Marathwada, Puneri, Vidharbh and Goa LDC-IL prepared the following dataset by which the prompt sheets were prepared. The prompt sheets were in Devanagari Script.

Content Type	Count
Created Text	8
Date	2
Command and Control Words	265
Most Frequent Words	1000
Form and Function Words	542
Phonetically Balanced Words	386
Person Name	440
Place Name	447
Sentences	385

Table 29-2: Representation of Marathi Content Type

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 part of the dataset prepared as follows.

Content Type	Content that Each typical prompt sheet had	Content selection type
News Text	1 Text	Distinct Text
Created 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 29-3: Representation of Marathi Prompt Sheet

The full set of

16. Phonetically Balanced Vocabulary
17. Form and Function Words
18. 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 four phases for different regions as follows:

Filed Work	Investigator name
July 2010 to August 2010	Apoorva Betkekar
January 2010 to July 2010	Gajanan
June 2018	Bhageshree
June 2018	Godavari

Table 29-4: Data Collection period and investigator details of Marathi Speech Data

29.3 TRANSLITERATIONS IN LDC-IL MARATHI READ CORPUS

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

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

LDC-IL Transliteration Schema

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

Vowels and Vowel Signs																	
अ	आ	इ	ई	उ	ऊ	ऋ	ॠ	ऌ	ॡ	ऍ	ऎ	ए	ऐ	ऑ	ऒ	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	ऴ
ya	ra	la	va	sha	Sa	sa	ha	La	Za

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

29.4 SUMMARY OF THE CORPUS

In the sections below, we provide the tabular details of the different content types of the Marathi 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 89:17:25 (hh:mm:ss) comprising 58544 audio segments.

29.4.1 Summary of the Audio Segments

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

LDC-IL Marathi 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	302	17	97	39	16	91	42
Created Text-T2	302	17	97	39	16	91	42
Sentence-S	7555	422	2436	978	400	2270	1049
Date-D	604	34	194	78	32	182	84
Command and Control Words-W1	9068	509	2925	1168	479	2729	1258
Person Name-W2	6058	339	1961	781	319	1820	838
Place Name-W2	3037	170	980	392	160	914	421
Most Frequent Word-Part-W3A	9104	510	2943	1177	480	2733	1261
Most Frequent Word-FullSet-W3B	10987	996	3997	997	999	2999	999
Phonetically Balanced-W4	4609	380	1923	384	385	1152	385
Form and Function Word-W5	6918	541	3248	538	540	1511	540

Table 29-5: Representation of Audio Segments of Marathi Raw Speech Data

29.4.2 Duration of the Marathi Raw Speech Data

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

LDC-IL Marathi 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	**22:26:06**	1:14:27	7:34:40	3:05:48	1:07:00	6:10:46	3:13:25
Created Text-T2	**13:37:34**	0:46:59	4:17:29	1:51:27	0:44:41	3:59:43	1:57:15
Sentence-S	**06:49:58**	0:22:15	2:11:00	0:55:09	0:21:04	2:01:30	0:59:00
Date-D	**00:39:57**	0:02:09	0:13:30	0:05:03	0:01:54	0:12:02	0:05:19
Command and Control Words-W1	**07:50:10**	0:24:10	2:30:23	0:59:32	0:22:28	2:23:50	1:09:47
Person Name-W2	**07:44:56**	0:23:44	2:28:41	1:02:18	0:22:45	2:18:47	1:08:41
Place Name-W2	**02:49:32**	0:08:38	0:53:46	0:21:46	0:08:17	0:51:47	0:25:18
Most Frequent Word-Part-W3A	**07:22:57**	0:22:02	2:20:51	0:56:25	0:21:34	2:16:40	1:05:25
Most Frequent Word-FullSet-W3B	**09:53:28**	0:46:24	3:40:07	0:46:22	0:45:07	3:13:55	0:41:33
Phonetically Balanced-W4	**04:10:47**	0:18:44	1:45:43	0:18:43	0:18:56	1:12:24	0:16:17
Form and Function Word-W5	**05:52:00**	0:26:26	2:46:29	0:25:30	0:26:00	1:27:04	0:20:31

Table 29-6: Representation of Marathi Raw Speech Data Duration

29.5 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.

29.5.1 Contemporary Text (News)

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

Age Group	Total Audio Segments	Gender-wise Distribution of text		Region-wise Distribution							
				MARATHWADA		PUNERI		VIDHARBH		GOA	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 to 20	35	19	16	9	8	8	8	0	0	0	0
21 to 50	186	95	91	46	45	49	46	1	0	1	0
50+	81	39	42	21	21	18	21	0	0	0	0
Total	**302**	153	149	76	74	75	75	1	0	1	0

Table 29-7: Representation of Marathi Contemporary text (News)

29.6 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 datasets available. The random sets are given below:

29.6.1 Creative Text-T2

One randomly selected text of literature out of 8 texts from the prepared Marathi 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 of text		Region-wise Distribution							
				MARATHWADA		PUNERI		VIDHARBH		GOA	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 to 20	33	17	16	9	8	8	8	0	0	0	0
21 to 50	188	97	91	46	45	49	46	1	0	1	0
50+	81	39	42	21	21	18	21	0	0	0	0
Total	**302**	153	149	76	74	75	75	1	0	1	0

Table 29-8: Representation of Marathi Creative Text

29.6.2 Sentences

The sentences content type contains a list of sentences that is a representation of almost all the phonemes occurring in Marathi. 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 of Sentences		Region-wise Distribution							
				MARATHWADA		PUNERI		VIDHARBH		GOA	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16to20	822	422	400	225	200	197	200	0	0	0	0
21 to 50	4706	2436	2270	1149	1126	1214	1144	49	0	24	0
50+	2027	978	1049	524	524	454	525	0	0	0	0
Total	**7555**	3836	3719	1898	1850	1865	1869	49	0	24	0

Table 29-9: Representation of Marathi Sentences

29.6.3 Date Format

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

Age Group	Total Audio Segments	Gender-wise Distribution of date format		Region-wise Distribution							
				MARATHWADA		PUNERI		VIDHARBH		GOA	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16to20	66	34	32	18	16	16	16	0	0	0	0
21 to 50	376	194	182	92	90	98	92	2	0	2	0
50+	162	78	84	42	42	36	42	0	0	0	0

Total	**604**	306	298	152	148	150	150	2	0	2	0

Table 29-10: Representation of Marathi Date formats

29.6.4 Command and Control Words

The command and control words content type contains a list of 265 words that is a representation of almost all the command and control words occurring in Marathi. 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 of words		Region-wise Distribution							
				MARATHWADA		PUNERI		VIDHARBH		GOA	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16to20	988	509	479	270	239	239	240	0	0	0	0
21 to 50	5654	2925	2729	1378	1349	1463	1380	55	0	29	0
50+	2426	1168	1258	630	629	538	629	0	0	0	0
Total	**9068**	4602	4466	2278	2217	2240	2249	55	0	29	0

Table 29-11: Representation of Marathi Command and Control Words

29.6.5 Person Name

The person name contains a list of 440 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 of words		Region-wise Distribution							
				MARATHWADA		PUNERI		VIDHARBH		GOA	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16to20	658	339	319	180	160	159	159	0	0	0	0
21 to 50	3781	1961	1820	920	900	982	920	34	0	25	0
50+	1619	781	838	420	420	361	418	0	0	0	0
Total	**6058**	3081	2977	1520	1480	1502	1497	34	0	25	0

Table 29-12: Representation of Marathi Person Names

29.6.6 Place Name

The place name contains a list of 447 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 of words		Region-wise Distribution							
				MARATHWADA		PUNERI		VIDHARBH		GOA	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16to20	330	170	160	90	80	80	80	0	0	0	0

21 to 50	1894	980	914	461	452	490	462	15	0	14	0
50+	813	392	421	209	211	183	210	0	0	0	0
Total	**3037**	1542	1495	760	743	753	752	15	0	14	0

Table 29-13: Representation of Marathi Place Names

29.6.7 Most Frequent Word-Part

The most frequent words-part contains a list of 1000 most frequent words occurring in Marathi. 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 of words		Region-wise Distribution							
				MARATHWADA		PUNERI		VIDHARBH		GOA	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16to20	1082	602	480	270	240	240	240	0	0	0	0
21 to 50	5584	2851	2733	1381	1353	1470	1380	59	0	33	0
50+	2438	1177	1261	630	630	547	631	0	0	0	0
Total	**9104**	4630	4474	2281	2223	2257	2251	59	0	33	0

Table 29-14: Representation of Marathi Most Frequent Words-Part

29.7 FULL SETS

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

29.7.1 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. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution of words		Region-wise Distribution							
				MARATHWADA		PUNERI		VIDHARBH		GOA	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 to 20	1995	996	999	0	0	996	999	0	0	0	0
21 to 50	6996	3997	2999	0	0	3997	2999	0	0	0	0
50+	1996	997	999	0	0	997	999	0	0	0	0
Total	**10987**	5990	4997	0	0	5990	4997	0	0	0	0

Table 29-15: Representation of Marathi Most Frequent Words -Full

29.7.2 Phonetically Balanced Vocabulary

The phonetically balanced vocabulary contain a list of words where almost all the phonemes of Marathi language has occurred in all the possible positions of a word. In full set all the 386 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 of words		Region-wise Distribution							
				MARATHWADA		PUNERI		VIDHARBH		GOA	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 to 20	765	380	385	0	0	380	385	0	0	0	0
21 to 50	3075	1923	1152	0	0	1538	1152	0	0	385	0
50+	769	384	385	0	0	384	385	0	0	0	0
Total	**4609**	2687	1922	0	0	2302	1922	0	0	385	0

Table 29-16: Representation of Marathi Phonetically Balanced Vocabulary

29.7.3 Form and Function Word

The form and function words contains a list of 542 words that is a representation of almost all the form and function words occurring in Marathi. 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 of words		Region-wise Distribution							
				MARATHWADA		PUNERI		VIDHARBH		GOA	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 to 20	1081	541	540	0	0	541	540	0	0	0	0
21 to 50	4759	3248	1511	0	0	2168	1511	539	0	541	0
50+	1078	538	540	0	0	538	540	0	0	0	0
Total	**6918**	4327	2591	0	0	3247	2591	539	0	541	0

Table 29-17: Representation of Marathi Form and Function Word

29.8 NATIVE SPEAKERS DISTRIBUTIONS

The following table shows the distributions of Marathi Native Speakers across the regional dialects.

Region-wise Distribution of Native Speakers											
Age Group	Total Native Speakers	Gender-wise Distribution of Native Speakers		Dialects							
				MARATHWADA		PUNERI		VIDHARBH		GOA	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 to 20	33	17	16	9	8	8	8	0	0	0	0
21 to 50	193	100	93	46	45	52	48	1	0	1	0
50+	81	39	42	21	21	18	21	0	0	0	0
Total	**307**	156	151	76	74	78	77	1	0	1	0

Table 29-18: Representation of Marathi Native Speakers Distributions

29.9 MOTHER TONGUE DISTRIBUTIONS OF THE NATIVE SPEAKERS

The following table shows the distribution of mother tongue of the native speakers in LDC-IL speech data.

Mother Tongue of the	Geographical Dialect Distribution of LDC-IL Marathi Speech Corpus	Total speaker

native speaker	Marathwada	Puneri	Vidharbh	Goa	
Marathi	149	155	1	-	305
Konkani	-	-	-	1	1
Hindi	1	-	-	-	1
Total	**150**	**155**	**1**	**1**	**307**

Table 29-19: Representation of Mother Tongue Distributions of Marathi Native Speakers.

30 NEPALI RAW SPEECH CORPUS

Rupesh Rai, Umesh Chamling Rai, Rajesha N, Manasa G, Narayan Choudhary, L Ramamoorthy

30.1 INTRODUCTION

Nepali is the principal and administrative language of Darjeeling and Sikkim. Nepali is written in Devanagari Script, from left to right direction. It also called Nagari. Nagari script has roots in the ancient Brāhmī script family. It has long been used traditionally by religiously educated people in South Asia. The Devanagari script is used for over 120 languages, and those are Nepali, Hindi, Marathi, Bhojpuri, Maithili etc. It closely related to the Nandinagari script commonly found in numerous ancient manuscripts of South India. The script is also used to write several minority languages of Nepali community such as Magar, Bhujel, Thami etc.

Nepali text corpus is collected from various libraries of Darjeeling, Sikkim, Assam, Uttranchal. Mostly from Kurseong, Mirik, Kalimpong, Silgadhi, Gangtok Guwahati, Almora, Mussoorie. The greater part of the text has been taken from Darjeeling General Library, Sonada Library, Mirik Public Library, Kalimpong City Library, NERLC (North-East Regional Language Centre, Guwahati) Library. LDC-IL tried to cover the entire category in its standard list. Some categories like novel, short stories has huge amount of books but some categories like physics, chemistry, economics, agriculture has very less amount of books. Literary texts are easily available in Nepali but getting scientific text is very difficult. Some categories like epigraphy, finance, oceanology text are too rare in Nepali.

LDC-IL divided the Nepali speaking areas into these three regions and collected speech data from each. After determining the regions for fieldwork, the dataset is prepared from which the prompt sheets were generated.

Places from which LDC-IL Nepali Speech Data is collected in Each Region is listed in the table below:

Region→	Darjeeling	Assam(North-East)	Uttranchal
Places →	4. Darjeeling 5. Dooars 6. Silgadhi	4. Guwahati 5. Udalguri	4. Deheradun 5. Pithoraghar

Table 30-1: Dialects and Places Covered for Nepali Speech Data

30.2 DATASET PREPARATION FOR NEPALI

For the selected Regions, Darjeeling, Dooars, Silgadhi, Guwahati, Udalguri, Deheradun and Pithoraghar LDC-IL prepared the following dataset by which the prompt sheets were prepared.

Content Type	Count
Creative Text	6
Date	3
Command and Control Words	74
Most Frequent Words	1290
Person Name	510
Place Name	324
Sentences	200

Table 30-2: Representation of Content Type

Distinct News Items were prepared to get the audio recording of contemporary text. Each prompt sheet had a distinct news item and selected 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*
		*randomly selected by machine

Table 30-3: Representation of Prompt Sheet

The full set of

1. Phonetically Balanced Vocabulary of 416 Words
2. Form and Function Words of 186 words
3. Most Frequent Wordlist 1278

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.

Region/	Year of data collection	Resource Person
Darjeeling-Assam	2009	Samar Sinha
Deheradun-Pithoraghar	2010	Jeena Rai
Darjeeling-Dooars-Silgadhi	2010	Umesh Chamling

Table 30-4: Fieldwork Details of Nepali Speech Data Collection

30.3 TRANSLITERATIONS IN LDC-IL NEPALI READ CORPUS

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

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

LDC-IL Transliteration Schema Nepali-Devanagari characters to Roman and Nepali Numerals to Hindu-Arabic										
Vowels and Vowel Signs										
अ	आ	इ	ई	उ	ऊ	ऋ	ए	ऐ	ओ	औ
	ा	ि	ी	ु	ू	ृ	े	ै	ो	ौ
A	A	i	I	u	U	x	E	ai	O	au
Consonants								Ayogavaha		
क	ख	ग	घ	ङ				ँ	ं	ः
Ka	kha	ga	gha	ng'a				M’	M	H
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						
य	र	ल	व	श	ष	स	ह			
Ya	ra	la	va	sha	Sa	sa	ha			
Numerals (Nepali-Devanagari)										
०	१	२	३	४	५	६	७	८	९	
0	1	2	3	4	5	6	7	8	9	

30.4 SUMMARY OF THE CORPUS

In the sections below, we provide the tabular details of the different content types of the Nepali 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 87:14:44 (hh:mm:ss) comprising 48975 audio segments.

30.4.1 Summary of the Audio Segments

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

LDC-IL Nepali 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	343	35	122	27	25	94	40
Creative Text-T2	341	35	123	27	24	93	39
Sentence-S	8583	873	3097	669	625	2295	1024
Date-D	1029	102	370	81	75	281	120
Command and Control Words-W1	10308	1050	3718	807	749	2757	1227
Person Name-W2	6878	699	2479	541	500	1839	820
Place Name-W2	3398	349	1206	269	249	918	407
Most Frequent Word-Part-W3A	10292	1050	3724	809	750	2730	1229
Most Frequent Word-FullSet-W3B	2994	0	997	0	0	1997	0
Phonetically Balanced-W4	3321	415	416	0	414	829	1247
Form and Function Word-W5	1488	186	186	0	186	372	558

Table 30-5: Representation of Audio Segments of Nepali Raw Speech Data

30.4.2 Duration of the nepali Raw Speech Data

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

LDC-IL Nepali 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	14:33:19	1:32:56	5:00:56	0:55:18	1:12:57	4:19:21	1:31:51
Creative Text-T2	19:46:34	2:21:33	7:00:18	1:23:42	1:37:47	5:20:42	2:02:32
Sentence-S	13:45:34	1:30:12	5:03:48	1:04:38	0:59:05	3:33:58	1:33:53
Date-D	0:57:20	0:05:23	0:20:28	0:05:09	0:03:53	0:15:29	0:06:58
Command and Control Words-W1	8:44:19	0:49:09	3:08:49	0:40:47	0:38:18	2:25:08	1:02:08
Person Name -W2	9:15:04	0:55:27	3:22:09	0:46:29	0:39:25	2:25:53	1:05:41
Place Name-W2	3:20:06	0:19:11	1:12:02	0:15:53	0:14:41	0:55:05	0:23:14
Most Frequent Word-Part-W3A	8:51:06	0:49:06	3:12:14	0:40:46	0:39:23	2:26:24	1:03:13
Most Frequent Word-FullSet-W3B	3:41:39	0:00:00	00:50:16	0:00:0	0:00:0	2:51:23	0:00:0
Phonetically Balanced-W4	3:00:08	0:19:02	0:20:15	0:00:0	0:16:25	1:02:05	1:02:21
Form and Function Word-W5	1:19:35	0:08:54	0:09:28	0:00:0	0:07:41	0:26:26	0:27:06

Table 30-6: Representation of Nepali Raw Speech Data Duration

30.5 DISTINCT SET

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

30.5.1 Contemporary Text (News) -T1

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

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution					
				Darjeelinge		Dehraduni		Assamiya	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	60	35	25	12	12	10	8	13	5
21 To 50	216	122	94	58	32	40	30	24	32
50+	67	27	40	9	16	17	13	1	11
Total	343	184	159	79	60	67	51	38	48

Table 30-7: Representation of Nepali Contemporary text (News)

30.6 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 datasets available. The random sets are given below:

30.6.1 Creative Text-T2

One randomly selected text of literature out of 6 texts from the prepared Nepali 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					
				Darjeelinge		Dehraduni		Assamiya	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	59	35	24	12	11	10	8	13	5
21 To 50	216	123	93	59	31	41	30	23	32
50+	66	27	39	9	15	17	13	1	11
Total	341	185	156	80	57	68	51	37	48

Table 30-8: Representation of Nepali Creative Text

30.6.2 Date Format-D

The answer of 3 questions is collected from each speaker to get the Nepali date format of the informants. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution					
				Darjeelinge		Dehraduni		Assamiya	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	177	102	75	36	36	30	24	36	15
21 To 50	651	370	281	175	95	123	90	72	96
50+	201	81	120	27	45	51	42	3	33
Total	1029	553	476	238	176	204	156	111	144

Table 30-9: Representation of Nepali Date format

30.6.3 Sentences-S

The sentences content type contains a list of sentences that is a representation of almost all the phonemes occurring in Nepali. 25 Randomly selected Sentences is 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					
				Darjeelinge		Dehraduni		Assamiya	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	1498	873	625	300	300	250	200	323	125
21 To 50	5392	3097	2295	1474	801	1023	744	600	750
50+	1693	669	1024	224	400	420	349	25	275
Total	8583	4639	3944	1998	1501	1693	1293	948	1150

Table 30-10: Representation of Nepali Sentences

30.6.4 Command And Control Words-W1

The command and control words content type contains a list of 74 words that is a representation of almost all the command and control words occurring in Nepali. 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		Region-wise Distribution					
				Darjeelinge		Dehraduni		Assamiya	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	1799	1050	749	360	360	300	240	390	149
21 To 50	6475	3718	2757	1769	960	1229	901	720	896
50+	2034	807	1227	269	480	508	417	30	330
Total	10308	5575	4733	2398	1800	2037	1558	1140	1375

Table 30-11: Representation of Nepali Command and Control words

30.6.5 Person Name-W2

The person name contains a list of 510 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		Region-wise Distribution					
				Darjeelinge		Dehraduni		Assamiya	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	1199	699	500	240	240	199	160	260	100
21 To 50	4318	2479	1839	1183	640	816	598	480	601
50+	1361	541	820	180	320	341	280	20	220
Total	6878	3719	3159	1603	1200	1356	1038	760	921

Table 30-12: Representation of Nepali Person Names

30.6.6 Place Name-W2

The place name contains a list of 324 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		Region-wise Distribution					
				Darjeelinge		Dehraduni		Assamiya	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	598	349	249	119	120	100	79	130	50
21 To 50	2124	1206	918	586	319	390	299	230	300
50+	676	269	407	90	160	169	137	10	110
Total	3398	1824	1574	795	599	659	515	370	460

Table 30-13: Representation of Nepali Place Names

1.1. MOST FREQUENT WORD-PART-W3A

The most frequent words-part contains a list of 1290 most frequent words occurring in Nepali. 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		Region-wise Distribution					
				Darjeelinge		Dehraduni		Assamiya	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	1800	1050	750	360	360	300	240	390	150
21 To 50	6454	3724	2730	1770	961	1231	899	723	870
50+	2038	809	1229	270	480	509	419	30	330
Total	10292	5583	4709	2400	1801	2040	1558	1143	1350

Table 30-14: Representation of Nepali Most Frequent Words-Part-W3A

30.7 FULL SET

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

30.7.1 Most Frequent Word-Full-W3B

The Most Frequent Words contain a list of 1278 most frequent words. In full set all the 1000 words are recorded from the informant. Each word is uttered three times. The distribution of data is as follows:

Age Group	Total Audio Segments	Assamiya		Darjeelinge
		Female	Male	Male
21 To 50	2994	997	997	1000

Table 30-15: Representation of Nepali Most Frequent Word-Full

1.2. PHONETICALLY BALANCED VOCABULARY-W4

The Phonetically Balanced words contain a list of words where almost all the phonemes of Nepali language has occurred in all the possible positions of a word. In full set all the 416 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 of words		Region-wise Distribution			
				Darjeelinge		Assamiya	
		Female	Male	Female	Male	Female	Male
16 to 20	829	415	414	0	0	415	414
21 to 50	1245	416	829	0	413	416	416

50+	1247	0	1247	0	415	0	832
Total	3321	831	2490	0	828	831	1662

Table 30-16: Representation of Nepali Phonetically Balanced Vocabulary

30.7.2 Form And Function Word-W5

The Form and Function Words contain a list of 186 words which is a representation of almost all the form and function words occurring in Nepali. 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 of words		Region-wise Distribution			
				Darjeelinge		Assamiya	
		Female	Male	Female	Male	Female	Male
16 to 20	372	186	186	0	0	186	186
21 to 50	558	186	372	0	186	186	186
50+	558	0	558	0	186	0	372
Total	1488	372	1116	0	372	372	744

Table 30-17: Representation of Nepali Form And Function Word

30.8 NATIVE SPEAKERS DISTRIBUTIONS

The following table shows the distributions of Nepali Native Speakers across the regions

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution					
				Darjeelinge		Dehraduni		Assamiya	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	61	36	25	12	12	10	8	14	5
21 To 50	219	124	95	59	32	41	30	24	33
50+	70	27	43	9	16	17	14	1	13
Total	350	187	163	80	60	68	52	39	51

Table 30-18: Representation of Nepali Native Speakers Distributions

31 PUNJABI RAW SPEECH CORPUS

Poonam Dhillon, Rajesha N, Manasa G, Narayan Choudhary, L. Ramamoorthy

31.1 INTRODUCTION

Punjabi is an Indo-Aryan language spoken by inhabitants of the historical Punjab region (northwestern India and in Pakistan)

Punjabi is one of the Indo-Aryan Language. Punjabi is a tonal language it has three tones, high-falling, low-rising, and level (neutral). As we know Punjabi is not spoken only in India it is also a language of Pakistan called Shahmukhi Punjabi. Here we are talking about only Indian Gurmukhi Punjabi. Punjabi language has four different dialects, spoken in the different sub-regions of Punjab. In Gurmukhi Punjabi Majhi is the prestige dialect which is spoken in Majha region of the Punjab mainly in Amritsar, Gurdaspur, Taran Taran and Pathankot.

Malwai dialect is spoken in the eastern part of Indian Punjab. Main areas are Ludhiana, Moga, Sangrur, Barnala, Faridkot, Patiala, Fatehgarh Sahib, Mansa, Muktsar, Ambala, Bathinda, Ganganagar and Malerkotla.

Doabi: "Do Aabi" mean "the land between two rivers" and this dialect is spoken between the rivers of Beas and Sutlej. It includes Jalandhar, Nawanshahr, Kapurthala and Hoshiarpur districts

Puadhi is spoken between the Satluj and Ghaggar rivers. The Puadhi dialect is spoken over a large area in present Punjab as well as Haryana. In Punjab, Kharar, Kurali, Ropar, Nurpurbedi, Morinda, Pail, Rajpura, and Samrala are the areas where the Puadhi language is spoken and the area itself is claimed as including from Pinjore, Kalka to Bangar area in Hisar district which includes even Nabha and Patiala in it. In Puadhi dialect we don’t find tone.

LDC-IL divided the Punjabi speaking areas into these four regions and collected speech data from Malwai, Doabi and Puadhi regions. After determining the regions for fieldwork, the prompt sheets were prepared for each region from master dataset.

31.2 DATASET PREPARATION FOR PUNJABI

For the selected Regions, 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	60
Most Frequent Words	1000
Person Name	396
Place Name	107
Sentences	202

Table 31-1: Representation of Content Type

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 in each typical prompt sheet	Content selection type
Contemporary Text (News Text)	1 Text	Distinct Text
Created 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 31-2: Representation of Prompt Sheet

The full set of

1. Phonetically Balanced Vocabulary of 775 Words
2. Form and Function Words of 589 words
3. 1000 Most Frequent Wordlist

were also carried to the field to get recorded by selected individuals.

Once all these preparations are made, the investigator started collecting the data.
Places from which LDC-IL Punjabi Speech Data is collected in Each Region.

Region	Malwa	Doaba	Puadh
Places	6. Patiala 7. Bathinda 8. Fatehgarh Sahib	9. Jalandhar 10. Nawanshahr 11. Kapurthala	12. Ropar 13. Kharar 14. Kurali 15. Mohalli

Table 31-3: Filed work details

31.3 TRANSLITERATIONS IN LDC-IL PUNJABI READ CORPUS

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

The LDC-IL transliteration scheme of Punjabi (in Gurumukhi scripts) to Roman is given below.

LDC-IL Transliteration Schema

Gurmukhi characters to Roman and Gurmukhi Numerals to Hindu-Arabic

Vowels

ਅ	ਆ	ਇ	ਈ	ਉ	ਊ	ਏ	ਐ	ਓ	ਔ
	ਾ	ਿ	ੀ	ੁ	ੂ	ੇ	ੈ	ੋ	ੌ
a	A	i	I	u	U	E	ai	O	au

Consonants

ਕ	ਖ	ਗ	ਘ	ਙ
ka	kha	ga	gha	ng'a
ਚ	ਛ	ਜ	ਝ	ਞ
ca	cha	ja	jha	nj'a
ਟ	ਠ	ਡ	ਢ	ਣ
Ta	Tha	Da	Dha	Na
ਤ	ਥ	ਦ	ਧ	ਨ
ta	tha	da	dha	na
ਪ	ਫ	ਬ	ਭ	ਮ
pa	pha	ba	bha	ma

Symbols

ੱ	ੰ	ਂ	ਃ
Null	m'	M	H

ਯ	ਰ	ਲ	ਵ	ੜ	ਸ਼	ਖ਼	ਗ਼	ਜ਼	ਫ਼	ਲ਼
ya	ra	la	va	Ra	sha	Kh'a	g'a	j'a	ph'a	La

Numerals (Punjabi to Hindu-Arabic)

੦	੧	੨	੩	੪	੫	੬	੭	੮	੯
0	1	2	3	4	5	6	7	8	9

The greyed out characters are obsolete. They may rarely present in the current LDC-IL corpus.

31.4 SUMMARY OF THE CORPORA

In the sections below, we provide the tabular details of the different content types of the Punjabi 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.

31.4.1 Summary of the Utterances

The table below shows the total number of utterances and their distribution in the Punjabi speech dataset.

LDC-IL Punjabi 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	448	27	133	65	23	134	66
Creative Text-T2	446	26	132	65	22	134	67
Date-D	887	54	262	128	46	263	134
Sentence-S	11168	673	3293	1625	550	3353	1674
Command and Control Words-W1	13274	795	3920	1923	683	3964	1989
Person Names-W2	8949	540	2642	1295	460	2671	1341
Place Name-W2	4473	269	1318	650	230	1336	670
Most Frequent Word-Part-W3A	8889	537	2644	1292	481	2614	1321
Most Frequent Word-FullSet-W3B	3988	990	1000	0	0	1998	0
Phonetically Balanced-W4	13939	2322	2323	2323	2321	2325	2325
Form and Function Word-W5	9779	1155	1720	1728	1713	1732	1731

Table 31-4: Representation of Audio Segments of Punjabi Raw Speech Data

31.4.2 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 Punjabi 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	27:07:41	01:46:05	07:34:12	03:59:07	01:24:06	08:12:09	04:12:02
Creative Text-T2	19:29:15	01:03:26	05:29:17	03:02:21	00:55:05	05:44:27	03:14:39
Date-D	00:27:53	00:01:29	00:08:00	00:04:28	00:01:15	00:07:51	00:04:50
Sentence-S	08:58:33	00:30:22	02:34:53	01:25:09	00:25:45	02:36:52	01:25:32
Command and Control Words-W1	07:49:16	00:25:08	02:16:56	01:11:24	00:23:01	02:15:09	01:17:38
Person Names-W2	10:28:40	00:34:42	02:59:48	01:43:20	00:28:38	02:57:07	01:45:05
Place Name-W2	03:17:02	00:11:03	00:57:37	00:31:02	00:09:22	00:55:55	00:32:03
Most Frequent Word-Part-W3A	05:21:56	00:16:10	01:35:04	00:49:27	00:16:15	01:30:46	00:54:14
Most Frequent Word-FullSet-W3B	02:52:44	00:36:46	00:45:45	00:00:00	00:00:00	01:30:13	00:00:00
Phonetically Balanced-W4	08:56:04	01:29:31	01:52:33	01:32:00	01:19:14	01:19:02	01:23:44
Form and Function Word-W5	06:24:07	00:45:32	01:23:09	01:06:35	01:03:09	01:00:16	01:05:26

Table 31-5: Representation of Punjabi Raw Speech Data Duration

31.5 DISTINCT SET

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

31.5.1 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	Gender-wise Distribution		Region-wise Distribution					
				MALWAI		PUADHI		DOABI	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	50	23	27	9	7	9	7	9	9
21 To 50	267	134	133	44	45	45	44	44	45
50+	131	66	65	22	23	22	23	21	20
Total	448	223	225	75	75	76	74	74	74

Table 31-6: Representation of Punjabi Contemporary text (News)

31.6 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 datasets available. The random sets are given below

31.6.1 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					
				MALWAI		PUADHI		DOABI	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	48	22	26	9	6	8	7	9	9
21 To 50	266	134	132	43	45	45	44	44	45
50+	132	67	65	22	23	22	23	21	21
Total	446	223	223	74	74	75	74	74	75

Table 31-7: Representation of Punjabi Creative Text

31.6.2 The Date

Answer to questioner of two 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					
				MALWAI		PUADHI		DOABI	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	100	46	54	18	14	18	14	18	18
21 To 50	525	263	262	88	90	86	86	88	87
50+	262	134	128	42	46	44	46	42	42
Total	887	443	444	148	150	148	146	148	147

Table 31-8: Representation of Punjabi Date Format

31.6.3 Sentences

The sentences content type contains a list of sentences that is a representation of almost all the phonemes occurring in Punjabi. 25 Randomly selected sentences are recorded from a list of 202 sentences. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender wise Distribution		Region-wise Distribution					
				MALWAI		PUADHI		DOABI	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	1223	550	673	225	150	224	175	224	225
21 To 50	6646	3353	3293	1092	1125	1101	1100	1100	1128
50+	3299	1674	1625	550	574	550	575	525	525
Total	11168	5577	5591	1867	1849	1875	1850	1849	1878

Table 31-9: Representation of Punjabi sentences

31.6.4 Command and Control Words

The command and control words content type contains a list of 60 words that is a representation of almost all the command and control words occurring in Punjabi. 30 randomly selected words are recorded from a list of words. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender wise Distribution		Region-wise Distribution					
				MALWAI		PUADHI		DOABI	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	1478	683	795	256	203	269	210	270	270
21 To 50	7884	3964	3920	1281	1296	1319	1319	1320	1349
50+	3912	1989	1923	634	669	660	690	629	630
Total	13274	6636	6638	2171	2168	2248	2219	2219	2249

Table 31-10 Representation of Punjabi command and control words

31.6.5 Person Names

The person name contains a list of 396 names 20 randomly selected names is recorded from a list of names. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender wise Distribution		Region-wise Distribution					
				MALWAI		PUADHI		DOABI	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	1000	460	540	180	140	180	140	180	180
21 To 50	5313	2671	2642	882	901	881	880	879	890
50+	2636	1341	1295	435	460	440	461	420	420
Total	8949	4472	4477	1497	1501	1501	1481	1479	1490

Table 31-11 Representation of Punjabi Person Names

31.6.6 Place Names

The place name contains a list of 107 popular regional place names. 10 randomly selected names are recorded from a list of names. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender wise Distribution		Region-wise Distribution					
				MALWAI		PUADHI		DOABI	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	499	230	269	90	70	89	70	90	90
21 To 50	2654	1336	1318	440	450	438	440	440	446
50+	1320	670	650	220	230	220	230	210	210
Total	4473	2236	2237	750	750	747	740	740	746

Table 31-12: Representation of Punjabi Place Names

31.6.7 Most Frequent Words

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

Age Group	Total Audio Segments	Gender wise Distribution		Region-wise Distribution					
				MALWAI		PUADHI		DOABI	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	1018	481	537	0	0	268	210	269	271
21 To 50	5258	2614	2644	0	0	1324	1322	1320	1292
50+	2613	1321	1292	0	0	660	691	632	630
Total	8889	4416	4473	0	0	2252	2223	2221	2193

Table 31-13: Representation of Punjabi Most Frequent Words-Part

31.7 FULL SET

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

31.7.1 Most Frequent Words

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

Age Group	Total Audio Segments	Gender wise Distribution		Region-wise Distribution					
				MALWAI		PUADHI		DOABI	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	990	0	990	0	0	0	0	990	
21 To 50	2998	1998	1000	0	0	1000	1000	0	998
Total	3988	1998	1990	0	0	1000	1000	990	998

Table 31-14: Representation of Punjabi Most Frequent Words-Full

31.7.2 Phonetically Balanced Vocabulary

The phonetically balanced vocabulary contain a list of words where almost all the phones of Punjabi language have occurred in all the possible positions of a word. In full set all the 775 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		Region-wise Distribution					
				MALWAI		PUADHI		DOABI	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	4643	2321	2322	775	773	775	775	772	773
21 To 50	4648	2325	2323	775	775	775	775	773	775
50+	4648	2325	2323	774	775	775	775	774	775
Total	13939	6971	6968	2324	2323	2325	2325	2319	2323

Table 31-15: Representation of Punjabi Phonetically Balanced Vocabulary

31.7.3 Form and Function Words

The form and function words content type contains a list of 589 words that is a representation of almost all the form and function words occurring in Punjabi. 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		Region-wise Distribution					
				MALWAI		PUADHI		DOABI	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	3433	1713	1720	588	588	567	562	565	563
21 To 50	2887	1732	1155	588	589	567	566	0	577
50+	3459	1731	1728	584	589	566	566	578	576
Total	9779	5176	4603	1760	1766	1700	1694	1143	1716

Table 31-16: Representation of Punjabi Form and Function Words

31.8 NATIVE SPEAKERS DISTRIBUTIONS

The distribution of native speakers across the regional dialect in LDC-IL Punjabi Speech corpus is as follows:

Region-wise Distribution of Native Speakers									
Age Group	Total Native Speakers	Gender-wise Distribution of Native Speakers		Dialects					
				MALWAI		PUADHI		DOABI	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	56	30	26	10	8	10	8	10	10
21 To 50	273	136	137	45	46	46	45	45	46
50+	138	68	70	23	24	23	24	22	22
Total	467	234	233	78	78	79	77	77	78

Table 31-17: Distribution of Punjabi Native Speakers

32 TELUGU RAW SPEECH CORPUS

Kavitha Lenin, Rajesha N, Manasa G, Narayan Choudhary, L. Ramamoorthy

32.1 INTRODUCTION

Telugu is the principal and official language of Andhra Pradesh and Telangana. It was also referred to as `Tenugu' in the past. Telugu language is the largest member of the Dravidian language family. Telugu, which is primarily spoken in southeastern India, it is the official language of the states of Andhra Pradesh and Telangana. Among the Dravidian languages, Telugu is spoken by the largest population. Based on 2011 census after Hindi and Bengali, Telugu is the third most frequently spoken of all the Indian languages. Telugu also has official language status in the Yanam district of the union territory of Puducherry.

Telugu language has four major dialects namely (i) Northern Telugu dialect spoken in Telangana region (10 districts) (ii) Southern Telugu dialect spoken in Rayalaseema region (4 districts), Nellore and Prakasam districts (iii) Eastern Telugu dialect spoken in Visakhapatnam, Vijayanagaram and Srikakulam districts and (iv) Central Coastal Telugu dialect which is considered as modern Standard Telugu dialect spoken in Guntur, Krishna, East and West Godavari. Its vocabulary is very much influenced by Sanskrit. In the course of time, some Sanskrit expressions used in Telugu got so naturalized that people regarded them as pure Telugu words. With the advent of the Muslim rule, several Persian and Arabic words entered into the Telugu language. Telugu script is originated from Brahmi script. The Brahmi script is used by Mauryan kings. The Bhattiprolu script is a variant of the Brahmi script which has been found in old inscriptions. The Bhattiprolu Brahmi script evolved to become the Telugu script by 5th century. Being a member of Dravidian family Telugu is agglutinative in nature.

Despite having a common language, Telugu Speaking areas have vast cultural and socio-economic differences. The Telugu speaking areas were divided into 3 geographical regions based on historical rulers, geographical features, regions of neighboring influential languages etc. These 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 Telugu spoken in these areas. For example, the Telangana region is highly influenced by Urdu as it was the part of the erstwhile Nizam Princely State of Hyderabad. Rayalaseema region was a part of Madras Presidency and has Kannada and Tamil Speaking areas in neighborhood. Historically Coastal Andhra part of many royal dynasties and became a part of Madras Presidency.

LDC-IL divided the Telugu speaking areas into these three regions and collected speech data from each.

32.2 DATASET PREPARATION FOR TELUGU

For the selected Regions, Telangana, Rayalaseema and Coastal Andhra, 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	228
Most Frequent Words	1402
Person Name	104
Place Name	254
Sentences	427

Table 32-1: Representation of Content Type

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. The distribution of content type in prompt sheet is as follows.

Content Type	Content that Each typical prompt sheet had	Content selection type
Contemporary Text (News Text)	1 Text	Distinct Text
Created 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 32-2: Representation of Prompt Sheet

The full set of Most Frequent Words was also recorded by selected individuals.

Once all these preparations were made, the investigator started collecting the data.

32.3 TRANSLITERATIONS IN LDC-IL READ CORPUS

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

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

LDC-IL Transliteration Scheme
Telugu characters to Roman and Telugu Numerals to Hindu-Arabic

Vowels and Vowel Signs[1]																
అ	ఆ	ఇ	ఈ	ఉ	ఊ	ఋ	ౠ	ఌ	ఎ	ఏ	ఐ	ఒ	ఓ	ఔ	అం	అః
	ా	ి	ీ	ు	ూ	ృ	ౄ	ౢ	ె	ే	ై	ొ	ో	ౌ	ం	ః
a	A	i	I	u	U	x	X	q	e	E	ai	o	O	au	aM	aH

Consonants				
క	ఖ	గ	ఘ	ఙ
ka	kha	ga	gha	ng'a
చ	ఛ	జ	ఝ	ఞ
ca	cha	ja	jha	nj'a
ట	ఠ	డ	ఢ	ణ
Ta	Tha	Da	Dha	Na
త	థ	ద	ధ	న
ta	tha	da	dha	na
ప	ఫ	బ	భ	మ
pa	pha	ba	bha	ma

య	ర	ల	వ	ళ	శ	ష	స	హ	ఱ
Ya	ra	la	va	La	sha	Sa	sa	ha	ṛ

[1] The greyed out characters are old in use. They are not present in the current LDC-IL corpus.

Numerals (Telugu to Hindu–Arabic numeral system)									
౦	౧	౨	౩	౪	౫	౬	౭	౮	౯
0	1	2	3	4	5	6	7	8	9

32.4 SUMMARY OF THE CORPUS

In the sections below, we provide the tabular details of the different content types of the Telugu 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 data size is of total duration 23:14:21 (hh:mm:ss) comprising 10510 audio segments.

32.4.1 Summary of the Audio Segments

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

LDC-IL Telugu 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)	77	14	4	5	7	45	2
Creative Text-T2	77	15	4	5	7	44	2
Sentence-S	1828	349	100	125	175	1029	50
Date-D	142	26	6	10	14	82	4
Command and Control Words-W1	2170	419	119	150	208	1214	60
Person Name-W2	1438	280	80	100	140	798	40
Place Name-W2	707	140	40	50	68	389	20
Most Frequent Word-Part-W3A	2162	420	120	150	210	1202	60
Most Frequent Word-FullSet-W3B	1909	1909	0	0	0	0	0

Table 32-3: Representation of Telugu Audio Segments

32.4.2 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 Telugu 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)	8:28:19	1:16:22	0:20:55	0:33:27	0:36:32	5:21:57	0:19:06
Creative Text-T2	7:10:35	1:12:55	0:28:16	0:25:05	0:43:07	4:15:22	0:05:50
Sentence-S	1:39:00	0:16:02	0:09:44	0:04:54	0:07:11	0:59:14	0:01:55
Date-D	0:14:49	0:02:28	0:01:30	0:01:02	0:01:03	0:08:26	0:00:20
Command and Control Words-W1	1:43:49	0:18:37	0:05:01	0:07:20	0:09:40	01:00:40	0:02:31
Person Name -W2	1:09:31	0:12:53	0:03:30	0:04:43	0:06:41	0:40:04	0:01:40
Place Name-W2	0:33:24	0:06:20	0:01:41	0:02:30	0:03:17	0:18:47	0:00:49
Most Frequent Word-Part-W3A	1:33:31	0:16:40	0:06:28	0:06:20	0:08:52	0:52:51	0:02:20
Most Frequent Word-FullSet-W3B	0:41:23	0:41:23	0:00:00	0:00:00	0:00:00	0:00:00	0:00:00

Table 32-4: Representation of Telugu Speech Data

32.5 DISTINCT SET

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

32.5.1 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	Gender-wise Distribution		Region-wise Distribution					
				COASTAL ANDHRA		RAYALASEEMA		TELANGANA	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	21	14	7	12	6	0	0	2	1
21 To 50	49	4	45	1	22	1	5	2	18
50+	7	5	2	5	2	0	0	0	0
Total	**77**	**23**	**54**	**18**	**30**	**1**	**5**	**4**	**19**

Table 32-5: Representation of Telugu Contemporary Text (News)

32.6 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 datasets available. The random sets are given below:

32.6.1 Creative Text

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

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution					
				COASTAL ANDHRA		RAYALASEEMA		TELANGANA	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	22	15	7	13	6	0	0	2	1
21 To 50	48	4	44	1	22	1	5	2	17
50+	7	5	2	5	2	0	0	0	0
Total	**77**	**24**	**53**	**19**	**30**	**1**	**5**	**4**	**18**

Table 32-6: Representation of Telugu Creative Text

32.6.2 Date

The answer to 2 questions are recorded 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					
				COASTAL ANDHRA		RAYALASEEMA		TELANGANA	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	40	26	14	22	12	0	0	4	2
21 To 50	88	6	82	2	40	0	10	4	32
50+	14	10	4	10	4	0	0	0	0
Total	**142**	**42**	**100**	**34**	**56**	**0**	**10**	**8**	**34**

Table 32-7: Representation of Telugu Date

32.6.3 Sentences

The sentences content type contains a list of sentences that is a representation of almost all the phonemes occurring in Telugu. 25 Randomly selected Sentences is recorded from a list of 427 sentences. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution					
				COASTAL ANDHRA		RAYALASEEMA		TELANGANA	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	524	349	175	299	150	0	0	50	25
21 To 50	1129	100	1029	25	517	25	126	50	386
50+	175	125	50	125	50	0	0	0	0
Total	**1828**	**574**	**1254**	449	717	25	**126**	**100**	**411**

Table 32-8: Representation of Telugu Sentences

32.6.4 Command and Control Words

The command and control words content type contains a list of 228 words that is a representation of almost all the command and control words occurring in Telugu. 30 randomly selected words is recorded from a list of words. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution					
				COASTAL ANDHRA		RAYALASEEMA		TELANGANA	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	627	419	208	359	178	0	0	60	30
21 To 50	1333	119	1214	30	604	29	149	60	461
50+	210	150	60	150	60	0	0	0	0
Total	**2170**	**688**	**1482**	**539**	**842**	**29**	**149**	**120**	**491**

Table 32-9: Representation of Telugu Command and Control Words

32.6.5 Person Names

The person name contains a list of 104 popular Pan Indian and regional person names. 20 randomly selected names are recorded from this list. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution					
				COASTAL ANDHRA		RAYALASEEMA		TELANGANA	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	420	280	140	240	120	0	0	40	20
21 To 50	878	80	800	20	391	20	100	40	307
50+	140	100	40	100	40	0	0	0	0
Total	**1438**	**460**	**978**	**360**	**551**	**20**	**100**	**80**	**327**

Table 32-10: Representation of Telugu Person Names

32.6.6 Place Names

The place name contains a list of 254 popular Pan Indian and regional place names. 10 randomly selected names are recorded from this list. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution					
				COASTAL ANDHRA		RAYALASEEMA		TELANGANA	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	208	140	68	120	58	0	0	20	10
21 To 50	429	40	389	10	189	10	50	20	150
50+	70	50	20	50	20	0	0	0	0
Total	**707**	**230**	**477**	**180**	**267**	**10**	**50**	**40**	**160**

Table 32-11: Representation of Telugu Place Names

32.6.7 Most Frequent Words

The most frequent words-part contains a list of 1402 most frequent words of Telugu. 30 randomly selected words recorded from this list. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution					
				COASTAL ANDHRA		RAYALASEEMA		TELANGANA	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	630	420	210	360	180	0	0	60	30
21 To 50	1322	120	1202	30	588	30	150	60	464
50+	210	150	60	150	60	0	0	0	0
Total	**2162**	**690**	**1472**	**540**	**828**	**30**	**150**	**120**	**494**

Table 32-12: Representation of Telugu Most Frequent Words

32.7 FULL SET

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

32.7.1 Most Frequent Words

The Most frequent words dataset contains more than 1000 words and around 1000 words were collected from two female speakers of age group 16-20 of Coastal Andhra Region and the totally 1909 audio segments are present in the Corpus.

32.8 NATIVE SPEAKERS DISTRIBUTIONS

The following table shows the Region-wise distribution of Native Speakers across Telugu speaking areas.

Region-wise Distribution of Native Speakers									
Age Group	Total Native Speakers	Gender-wise Distribution of Native Speakers		Dialects					
				COASTAL ANDHRA		RAYALASEEMA		TELANGANA	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	22	15	7	13	6	0	0	2	1
21 To 50	51	4	47	1	24	1	5	2	18
50+	7	5	2	5	2	0	0	0	0
Total	**80**	**24**	**56**	**19**	**32**	**1**	**5**	**4**	**19**

Table 32-13: Representation of Telugu Native Speakers Distributions

33 URDU RAW SPEECH CORPUS

Mansoor Khan, Shahnawaz Alam, Bi. Bi. Mariyam, Rajesha N, Manasa G, Narayan Choudhary, L. Ramamoorthy

33.1 INTRODUCTION

The word 'Urdu' is derived from the Turkish word '*Ordu'* which means ARMY. Urdu was also called *Hindavi*, *Rekhta:Urdu-e- Mu'alla, Hindustani* etc. (Chatterji, S.K. 1960: 197). Urdu was called by the name of *Urdu-e-Mu'alla:* or royal military bazaar outside the Delhi palace of the Mughals. Urdu when used by men, especially employed for poetry, was called *Rexta:* (i. e. 'scattered' or 'mixed'). Persian words were 'scattered through it'. Now it is undoubted fact that the name '*Rexta:'* was the oldest and earlier name of Urdu.

The birth of Urdu language was the direct result of the synthesis between the invading armies of Mahmood of Ghazni with the civilian population of the Indian cities. The word Urdu itself means *Lashkar*, derived from the Turkish language meaning armies… In the south of India it flourished under the name of *Dakhni* and south west as *Gujari* while in Delhi its name changed from *Hindi* to *Hindavi* and *Hindustani*. The people of Delhi called it *Dehlvi* or *Zubān-e-Dehlvi* or *Urdu-e-Muallā*. There are various theories where exactly it was born. One theory is that it originated in basti Nizamuddin of Nizamuddin Aulia and Amir Khusru. Another theory is that it was born in the seminary of Baba ShaikhFarīd at PākPattan in the 13th century and its old name is *Multani* or *Old Lahori*. When Babar came to India, he did not find anything exclusively Hindu or exclusively Muslim. He talked of the Hindustani way of life.

According to Suniti Kumar Chatterji, (1969:103-4)

Urdu is one of the modern Indo-Aryan languages of India. It evolved from *Shaurseni Apabhramsha* through the dialects prevalent in and around Delhi at the end of the 12th century A.D., when the Muslims comprising Turks, Iranians and Afghans entered Delhi from the Punjab as the new settlers. Though it was a political incident in the history of India, it had greatly influenced the linguistic scene of Northern India. It had accelerated the process of the development of the New Indo-Aryan languages.

According to Masood Husain Khan, (1987: 234-62)

The historians are of the view that the cultural contacts of the Indian with the Arabs and the Iranians had been very old. These contacts were established long before the conquest of Mohammad Ghauri as a result of which the linguistic interaction between two communities began outside the Punjab and hectic political activities started in North India in A.D. 1193. The impact of this linguistic interaction and cohesion is well evidenced in the literary and other documents of those days, which assimilated quite a number of words from Arabic and Persian.

The people who have moved from the Punjab to Delhi in A.D. 1193 had brought with them at least four languages viz, Arabic, Persian, Turkish and an early form of Punjabi. The natives on the other hand had promoted the growth of indigenous dialects, which were the off shoots of western Hindi that had developed from *Shaurseni Apabhramsha*. When Delhi was made the capital the Muslim's sovereignty in India and when it assumed the importance of the military headquarters, the new settlers and the local people had frequent opportunities to get together. As a result of the political, social and cultural contacts between the two speech communities, there evolved a mix form of language known as *Rekhta* whose base was largely supplied by *Khaři Boli,* a dialect of western Hindi. It assimilated a large number of words from Persian and Arabic.

Besides lexical items, it also absorbed numerous expressions, phrases and clauses from Persian. Since it owed its existence to the indigenous dialects of India, the Muslim sovereigns called this language Hindi, i.e., the language of India. It was also known as *Hindavī*. In the course of its development, it assumed various names like *Zabān-e-Dehli*, *Zabān-e-Hindustān, Zabān-e-Urdū-e-mu'allā, Zabān-e-Urdu* and in later period simply *Urdu*.

LDC-IL divided the Urdu speaking areas into these three regions and collected speech data from each. After determining the regions for fieldwork, the dataset is prepared from which the prompt sheets were generated.

Places from which LDC-IL Urdu Speech Data is collected in each region is listed in the table below.

Region→	Uttar Pradesh	Madhya Pradesh	Uttar Pradesh
Places →	Aligarh	Bhopal	Lucknow

Table 33-1: Dialects and Places Covered for Urdu Speech Data.

33.2 DATASET PREPARATION FOR URDU

For the selected regions, Aligarh (Uttar Pradesh), Bhopal (Padhya Pradesh) and Lucknow (Uttar Pradesh). 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	141
Most Frequent Words	1000
Form and Function Words	370
Phonetically Balanced Words	775
Person Name	400
Place Name	100
Sentences	195

Table 33-2: 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 seleted part of the dataset prepared as follows.

Content Type	**Content that Each typical prompt sheet had**	**Content selection type**
News Text	1 Text	Distinct Text
Created 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 33-3: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 as follows.

Region/Place	Year of data collection	Resource Person
Aligarh (Uttar Pradesh)	2009	Rushda Irdees Khan
Bhopal (Madhya Pradesh)	2010	Rushda Irdees Khan
Lucknow (Uttar Pradesh)	2010	Mansoor Khan

Table 33-4: Three Phases of Speech Data Collection

33.3 SUMMARY OF THE CORPUS

In the sections below, we provide the tabular details of the different content types of the Urdu 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 99:18:21 (hh:mm:ss) comprising 88,708 audio segments.

33.3.1 Summary of the Audio Segments

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

LDC-IL Urdu 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	431	53	116	50	38	121	53
Creative Text-T2	433	53	116	51	37	122	54
Sentence-S	10646	1312	2895	1249	876	3015	1299
Date-D	846	104	227	100	71	240	104
Command and Control Words-W1	13580	1362	3578	1750	1390	3577	1923
Person Name-W2	6577	758	1810	795	542	1845	827
Place Name-W2	4273	528	1149	500	370	1196	530
Most Frequent Word-Part-W3A	12802	1585	3421	1500	1109	3617	1570
Most Frequent Word-FullSet-W3B	18927	1998	2979	3990	1997	2996	4967
Phonetically Balanced-W4	13646	1517	3017	2271	1527	3032	2282
Form and Function Word-W5	6547	731	1096	1460	729	1459	1072

Table 33-5: Urdu Audio Segments and their Distribution

33.3.2 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 Urdu 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	25:35:02	3:07:57	6:53:35	3:10:48	2:21:21	7:04:52	2:56:29
Creative Text-T2	19:40:11	2:20:22	5:05:52	2:16:11	1:47:15	5:21:37	2:48:54
Sentence-S	8:00:38	0:57:38	2:08:27	1:00:26	0:38:42	2:13:33	1:01:52
Date-D	0:43:37	0:05:19	0:11:39	0:05:05	0:03:29	0:12:13	0:05:52
Command and Control Words-W1	9:21:01	0:52:10	2:24:56	1:14:04	0:54:05	2:32:26	1:23:20
Person Name-W2	2:55:41	0:19:50	0:47:35	0:22:54	0:13:52	0:47:39	0:23:51
Place Name-W2	1:09:17	0:08:08	0:18:03	0:08:27	0:05:48	0:19:19	0:09:32
Most Frequent Word-Part-W3A	7:46:28	0:54:06	2:03:51	0:53:24	0:40:19	2:09:57	1:04:51
Most Frequent Word-FullSet-W3B	11:38:30	1:20:02	1:49:10	2:11:07	1:14:23	1:41:57	3:21:51
Phonetically Balanced-W4	8:13:20	0:48:08	1:52:05	1:10:52	0:51:54	2:07:27	1:22:54
Form and Function Word-W5	4:14:36	0:33:39	0:34:38	0:56:52	0:26:00	0:58:30	0:44:57

Table 33-6:Duration of the Collected Urdu Speech Data

33.4 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.

33.4.1 Contemporary Text (News) T-1

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 Text (One distinct text/speaker)	Gender-wise Distribution		Region-wise Distribution					
				Braj (Uttar Pradesh)		Bhopali (Madhya Pradesh)		Rekhta (Uttar Pradesh)	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	91	53	38	10	9	27	14	16	15
21 To 50	237	116	121	43	40	34	41	39	40
50+	103	50	53	21	20	9	13	20	20
Total	431	219	212	74	69	70	68	75	75

Table 33-7:Distribution of Urdu Contemporary Text (News) Data

33.5 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.

33.5.1 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 Text	Gender-wise Distribution		Region-wise Distribution					
				Braj (Uttar Pradesh)		Bhopali (Madhya Pradesh)		Rekhta (Uttar Pradesh)	
		Female	**Male**	**Female**	**Male**	**Female**	**Male**	**Female**	**Male**
16 To 20	90	53	37	10	9	27	13	16	15
21 To 50	238	116	122	43	41	34	41	39	40
50+	105	51	54	21	21	10	13	20	20
Total	433	220	213	74	71	71	67	75	75

Table 33-8:Distribution of Urdu Creative Text

33.5.2 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 questionnaire (Two questions per speaker)	Gender-wise Distribution		Region-wise Distribution					
				Braj (Uttar Pradesh)		Bhopali (Madhya Pradesh)		Rekhta (Uttar Pradesh)	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	175	104	71	20	18	52	25	32	28
21 To 50	467	227	240	84	82	67	78	76	80
50+	204	100	104	42	40	18	24	40	40
Total	846	431	415	146	140	137	127	148	148

Table 33-9:Distribution of Urdu Date Format

33.5.3 The Sentences-S

The Sentences contain a list of sentences that is a representation of almost all the phonemes occurring in Urdu. 25 Randomly selected sentences are recorded from a list of 195sentences. The distribution of data is as follows.

Age Group	Total Sentences	Gender wise Distribution		Region-wise Distribution					
				Braj (Uttar Pradesh)		Bhopali (Madhya Pradesh)		Rekhta (Uttar Pradesh)	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	2188	1312	876	250	224	662	277	400	375
21 To 50	5910	2895	3015	1076	998	844	1014	975	1003
50+	2548	1249	1299	524	500	225	299	500	500
Total	10646	5456	5190	1850	1722	1731	1590	1875	1878

Table 33-10:Distribution of Urdu Sentences

33.5.4 Command and Control Words-W1

The Command and Control Wordscontain a list of 141 words that is a representation of almost all the command and control words occurring in Urdu. 30 randomly selected words is recorded from a list of words. The distribution of data is as follows.

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution					
				Braj (Uttar Pradesh)		Bhopali (Madhya Pradesh)		Rekhta (Uttar Pradesh)	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	2752	1362	1390	441	409	300	390	621	591
21 To 50	7155	3578	3577	1549	1456	718	781	1311	1340
50+	3673	1750	1923	889	882	120	300	741	741
Total	13580	6690	6890	2879	2747	1138	1471	2673	2672

Table 33-11:Distribution of Command and Control Words

33.5.5 Person Names –W2

The Person Names contain a list of 400 popular Pan Indian and regional Person names. 20 randomly selected namesare recorded from the list. The distribution of data is as follows.

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution					
				Braj (Uttar Pradesh)		Bhopali (Madhya Pradesh)		Rekhta (Uttar Pradesh)	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	1300	758	542	199	178	351	169	208	195
21 To 50	3655	1810	1845	861	795	442	530	507	520
50+	1622	795	827	418	398	117	169	260	260
Total	6577	3363	3214	1478	1371	910	868	975	975

Table 33-12:Distribution of Urdu Person Names

33.5.6 Place Names-W2

The Place Names contain a list of 100 popular Pan Indian and regional Place names. 10 randomly selected namesare recorded from the list. The distribution of data is as follows.

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution					
				Braj (Uttar Pradesh)		Bhopali (Madhya Pradesh)		Rekhta (Uttar Pradesh)	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	898	528	370	100	90	270	130	158	150
21 To 50	2345	1149	1196	419	399	340	397	390	400
50+	1030	500	530	210	200	90	130	200	200
Total	4273	2177	2096	729	689	700	657	748	750

Table 33-13:Distribution of Urdu Place Names

33.5.7 Most Frequent Words-PART-W3A

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

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution					
				Braj (Uttar Pradesh)		Bhopali (Madhya Pradesh)		Rekhta (Uttar Pradesh)	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	2694	1585	1109	300	269	805	390	480	450
21 To 50	7038	3421	3617	1294	1196	956	1221	1171	1200
50+	3070	1500	1570	628	578	271	390	601	602
Total	12802	6506	6296	2222	2043	2032	2001	2252	2252

Table 33-14:Distribution of Urdu Most Frequent Words - Part

33.6 FULL SET

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 as below.

33.6.1 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		Region-wise Distribution					
				Braj (Uttar Pradesh)		Bhopali (Madhya Pradesh)		Rekhta (Uttar Pradesh)	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	3995	1998	1997	1998	1000	0000	0000	0000	0997
21 To 50	5975	2979	2996	0981	1998	0000	0998	1998	0000
50+	8957	3990	4967	0994	1974	1996	0994	1000	1999
Total	18927	8967	9960	3973	4972	1996	1992	2998	2996

Table 33-15:Distribution of Urdu Most Frequent Words - Full

33.6.2 The Phonetically Balanced Vocabulary-W4

The Phonetically Balanced Vocabulary contain a list of words where almost all the phones of Urdu language have occurred in all the possible positions of a word. In full set all the 773 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		Region-wise Distribution					
				Braj (Uttar Pradesh)		Bhopali (Madhya Pradesh)		Rekhta (Uttar Pradesh)	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	3044	1517	1527	0762	0772	0000	0000	0755	0755
21 To 50	6049	3017	3032	2262	1522	0000	0755	0755	0755
50+	4553	2271	2282	1517	1527	0000	0000	0754	0755
Total	13646	6805	6841	4541	3821	0000	755	2264	2265

Table 33-16:Distribution of Urdu Phonetically Balanced Vocabulary

33.6.3 The Form and Function Words-W5

The Form and Function Words contain a list of 370 words that is a representation of almost all the form and function words occurring in Urdu. 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		Region-wise Distribution					
				Braj (Uttar Pradesh)		Bhopali (Madhya Pradesh)		Rekhta (Uttar Pradesh)	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	1460	731	729	0366	0364	0000	0000	0365	0365
21 To 50	2555	1096	1459	0731	0730	0000	0364	0365	0365
50+	2532	1460	1072	1095	0707	0000	0000	0365	0365
Total	6547	3287	3260	2192	1801	0000	364	1095	1095

Table 33-17:Distribution of Urdu Form and Function words

33.7 NATIVE SPEAKERS DISTRIBUTIONS

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

Region-wise Distribution of Native Speakers									
Age Group	Total Native Speakers	Gender-wise Distribution of Native Speakers		Dialects					
				Braj (Uttar Pradesh)		Bhopali (Madhya Pradesh)		Rekhta (Uttar Pradesh)	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	105	60	45	15	13	27	14	18	18
21 To 50	263	128	135	51	49	34	44	43	42
50+	131	64	67	29	29	12	14	23	24
Total	499	252	247	95	91	73	72	84	84

Table 33-18:Distribution of Urdu Native Speakers

33.8 REFERENCES

1. Beg, Mirza Khalil Ahmad (1988). *Urdu Grammar: History and Structure*. New Delhi: Bahri Publications.
2. Chatterji, Suniti Kumar (1960).*Indo-Aryan and Hindi* (Second Edition). New Delhi: K.V. Sachdeva, Skylark Printers.
3. Faruqi, Shamsur Rahman (2001).*Early Urdu Literary Culture and History*. Delhi: Oxford University Press.
4. Khan, Iqtidar Hussain (1999).*A Contrastive and Comparative Study of Standard Urdu and Standard Hindi*. Aligarh: Aligarh Muslim University Press.
5. Khan, Zubair Shadab. (2013). *Urdu Language, Literature and Culture.* Arshia Publication Pvt. Ltd, Aligarh.