

সনা ফুদোংবা
খুং চংদ্রিবা মনিপুৰী মপী ব্ৰাহ্মকোন

A Gold Standard Manipuri Raw Text Corpus



**A GOLD STANDARD
MANIPURI RAW TEXT CORPUS**

Manipuri Raw Text Corpus

Authors:

Amom Nandaraj Meetei

Yumnam Premila Chanu

Rajesha N.

Manasa G.

Dr. Narayan Choudhary

Dr. L. Ramamoorthy



34

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

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

Manipuri Raw Text Corpus

This Documentation is a part of LDC-IL's "A Gold Standard Manipuri Raw Text Corpus".

For Further contacts: oic@ldcil.org

All rights reserved

© Central Institute of Indian Languages, Mysore-2018

Authors:

Amom Nandaraj Meetei

Yumnam Premila Chanu

Rajesh N.

Manasa G.

Dr. Narayan Choudhary

Dr. L. Ramamoorthy

Last updated: December 12, 2018

ISBN 978-81-7343-245-3 Electronic publication of Corpus

Publisher:

Linguistic Data Consortium for Indian Languages

Central Institute of Indian Languages

Mysore, India-570006

Contents

1	LDC-IL Raw Text Corpora; An Overview	1
2	LDC-IL Approach of Sampling.....	3
2.1	Sampling Approach for Books.....	3
2.2	Sampling Approach for Magazines	4
2.3	Sampling Approach for newspaper.....	4
3	LDC-IL Text Corpus categorization	5
3.1	Aesthetics.....	5
3.2	Commerce.....	5
3.3	Mass Media.....	6
3.4	Official Document	6
3.5	Science and Technology.....	6
3.6	Social Sciences	7
4	LDC-IL Text Data encoding and format	7
5	LDC-IL Text Corpus Metadata	8
6	LDC-IL Text Corpus and Naming conventions.....	10
7	Proof Reading.....	11
8	Copyright.....	12
9	Manipuri Raw Text Corpus.....	13
9.1	Introduction	13
9.2	Peculiarities of Manipuri text.....	14
9.2.1	The Scripts used in writing the language	14
9.3	Principles of Data Sampling	15
9.4	Field Works Undertaken	15
9.5	Data Inputting	17
9.6	Validation and Normalization Workshops	17
9.7	Proofreading	17
9.8	Copyright Consents	18
10	Transliterations in LDC-IL Manipuri text corpus	19

11	Overview of Represented Domains	20
11.1	Official Document	20
11.2	Aesthetics.....	21
11.3	Commerce.....	22
11.4	Mass Media.....	22
11.5	Science and Technology.....	23
11.6	Social Sciences	24
	References	24

Table of Figures

Table 1: Subcategories of the Category Aesthetics	5
Table 2: Subcategories of the Category Commerce	5
Table 3: Subcategories of the Category Mass Media	6
Table 4: Subcategories of the Category Official Documents	6
Table 5: Subcategories of the Category Science and Technology	7
Table 6: Subcategories of the Category Social Sciences	7
Table 7: Metadata Legends for LDC-IL Text Data.....	8
Table 8: LDC-IL notations for Indian Scheduled languages.....	10
Table-9: Representation of the Domains in Manipuri Text Corpus	20
Table 10: Official Document Category Representation	20
Table 11: Aesthetics Category Representation.....	21
Table 12: Commerce Category Representation.....	22
Table 13: Mass Media Category Representation.....	22
Table 14: Science and Technology Representation	23
Table 15: Social Sciences Representation.....	24

1 LDC-IL RAW TEXT CORPORA; AN OVERVIEW

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 acquisition¹.”

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

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.

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 10th page and if the book has 200+ pages we collect every 20th 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.

2.2 SAMPLING APPROACH FOR MAGAZINES

In case of magazine texts are 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.

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 writers would 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.

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.

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 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 are 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’s Literature	Literature-Text Books (School)	Literature-Travelogues	Literature-Science Fiction
				Literature-Short Stories

Table 1: Subcategories of the Category Aesthetics

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 2: Subcategories of the Category Commerce

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 3: Subcategories of the Category Mass Media

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 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 4: Subcategories of the Category Official Documents

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 5: Subcategories of the Category Science and Technology

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 communal context 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 Science	Political Science	Public Administration	Health and Family Welfare	Sociology
Epigraphy					Linguistics

Table 6: Subcategories of the Category Social Sciences

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.

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	Project Description	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	Sampling Description	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	Text Type	Is mostly blank however sometimes it is used to provide the broad topic of the news items e.g. whether it is a political news or editorial or sports news etc.
11	Headline	This information is a verifiable proof for the corpus. This is normally the heading of the chapter of the selected sample. Gives the fine tuned information of the topic present in particular file.
12	Author	Specifies the name of the author.
13	Editor	Specifies the name of the editor.
14	Translator	Specifies the name of the translator.
15	Words	Specifies the total number of words in the file.
16	Letters	Specifies the total number of UTF8 characters in the file.
17	Publishing Place	Specifies the place where the title was published.
18	Publisher	Specifies the name of the publisher.
19	Published Year	Specifies the 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 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>
	<editor>		</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>			
<p>			
<p>			
</text> </body> </Doc>			

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 639 Language 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	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	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

Table 8: LDC-IL notations for Indian Scheduled languages

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 suffixed and enclosed in braces.

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

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.

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.

9 MANIPURI RAW TEXT CORPUS

9.1 INTRODUCTION

Manipuri (also locally known Meeteilon by the native speakers) is the lingua franc 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 unmemorable. 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.

9.2 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.

9.2.1 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’
সাতিন/ছাটিন/ছাটিণ্/শাতিন/সাতীন/শাতীন	pronounced as /satin/ ‘umbrella’
সাতিন্/ছাটিন্/ছাটিণ্/শাতিন্/সাতীন্/শাতীন্	
গারি/গাড়ি/গাড়ী	pronounced as /gari/ ‘cart/van’
রিতু/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 □ 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 □□ /sən/. In a similar way, we will have the following representational forms for the items shown above.

সন/শন/ ষন/ ষণ/সন্/শন্/ষণ্	□□ /sən/ ‘cow’
সাতিন/ছাটিন/ছাটিণ্/শাতিন/সাতীন/শাতীন	□□□□□ /satin/ ‘umbrella’
সাতিন্/ছাটিন্/ছাটিণ্/শাতিন্/সাতীন্/শাতীন্	
গারি/গাড়ি/গাড়ী	□□□□/gari/ ‘cart/van’
রিতু/রীতু/ঋতু	□□□□/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.

9.3 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.

9.4 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.

9.5 DATA INPUTTING

All the texts have been typed in Unicode using the In Script 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.

9.6 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.

9.7 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:

1. Manipuri Corpus Normalization, 2010
2. Short Term Goal Oriented Project- Manipuri Language Text Cleaning Workshop:
7th Oct. – 11th Dec. 2014
3. 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 its 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.

9.8 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.

10 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	□		□		□								
Bengali	া		ি	ী	ু	ূ	্	ে	ৈ	ো	ৌ	ং	ঃ
Meetei Mayek		□	□		□			□	□	□	□	□	
Roman	a	A	i	I	U	U	x	E	ai	O	au	M	H
Consonants							Unreleased Consonants						
Bengali	ক	খ	গ	ঘ	ঙ	ক	ক্		ঙ্				
Meetei Mayek	□	□	□	□	□	□	□		□				
Roman	ka	kha	ga	gha	ng'a	ka	k		ng'				
Bengali	চ	ছ	জ	ঝ	ঞ	চ	ম্		প্				
Meetei Mayek	□		□			□	□		□				
Roman	ca	cha	Ja	jha	nj'a	ca	m		p				
Bengali	ট	ঠ	ড	ঢ	ণ	ট	ন্	ত্	ল্				
Roman	Ta	Tha	Da	Dha	Na	Ta	□	□	□				
							n	t	l				
Bengali	ত	থ	দ	ধ	ন	ত							
Meetei Mayek	□	□	□	□	□	□							
Roman	ta	tha	Da	dha	na	ta							
Bengali	প	ফ	ব	ভ	ম	প							
Meetei Mayek	□	□	□	□	□	□							
Roman	pa	pha	Ba	bha	ma	pa							
Bengali	য	র	ল	শ	স	ষ	হ	ড়	ঢ়	য়	ৎ		
Meetei Mayek		□	□		□		□			□			
Roman	ya	ra	la	sha	Sa	sa	ha	D'a	Dh'a	Ya	t		
Numerals (Bengali and Meetei Mayek to Hindu-Arabic)													
Bengali	০	১	২	৩	৪	৫	৬	৭	৮	৯			
Meetei Mayek	□	□	□	□	□	□	□	□	□	□			
Hindu-Arabic	0	1	2	3	4	5	6	7	8	9			

11 OVERVIEW OF REPRESENTED DOMAINS

LDC-IL Manipuri Text Corpus size is 61,48,220 words and 4,31,27,842 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
Official Document	442950	7.20%
Aesthetics	3775535	61.39%
Commerce	18450	0.30%
Mass Media	775662	12.63%
Science and Technology	304545	4.96%
Social Sciences	831078	13.52%
Total	6148220	100

Table-9: 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.

11.1 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%	7.18%
03	Police Documents	344	0.08%	0.01%
	Total	442950	100%	7.20%

Table 10: Official Document Category Representation

11.2 AESTHETICS

The Aesthetics category of Manipuri text corpus covers 29 sub-categories bearing a total of 37,75,535 words along with the overall percentage of 61.39%. 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	5.25%	3.23%
03	Cinema	17473	0.46%	0.28%
04	Culture	321048	8.50%	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	Humour	2541	0.07%	0.04%
14	Literary Texts	9186	0.24%	0.15%
15	Literature-Children's Literature	2502	0.07%	0.04%
16	Literature-Criticism	809311	21.44%	13.16%
17	Literature-Diaries	36784	0.97%	0.60%
18	Literature-Epics	1398	0.04%	0.02%
19	Literature-Essays	319538	8.46%	5.20%
20	Literature-Letters	1559	0.04%	0.03%
21	Literature-Novels	354659	9.39%	5.77%
22	Literature-Plays	319728	8.47%	5.20%
23	Literature-Poetry	7668	0.20%	0.12%
24	Literature-Science Fiction	1061	0.03%	0.02%
25	Literature-Short Stories	852839	22.59%	13.87%
26	Literature-Speeches	58061	1.54%	0.94%
27	Literature-Text Books (School)	38644	1.02%	0.63%
28	Literature-Travelogues	173173	4.59%	2.82%
29	Mythology	22366	0.59%	0.36%
	Total	3775535	100%	61.39%

Table 11: Aesthetics Category Representation

11.3 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: Commerce Category Representation

11.4 MASS MEDIA

The Mass Media category of Manipuri text corpus covers 13 sub-categories bearing a total of 7,75,662 words along with the overall percentage of 12.63%. 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	36.62%	4.62%
05	General News	31352	4.04%	0.51%
06	Interviews	6016	0.78%	0.10%
07	Letters	92855	11.97%	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	6.10%	0.77%
12	Speeches	1067	0.14%	0.02%
13	Sports News	234563	30.24%	3.82%
	Total	775662	100%	12.63%

Table 13: Mass Media Category Representation

11.5 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 14: Science and Technology Representation

11.6 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 15: Social Sciences Representation

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.