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 Science	Political Science	Public Administration	Health and Family Welfare	Sociology
Epigraphy					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.