ಪ್ರಮಾಣಿತ ಕನ್ನಡ ಸಹಜ ಪಠ್ಯ ದತ್ತಾಂಶ

A Gold Standard
Kannada Raw Text Corpus

Kannada Raw Text Corpus

Authors:
Vijayalaxmi F. Patil
Chetan S Baji
Rajesha N.
Manasa G.
Dr. Narayan Choudhary
Dr. L. Ramamoorthy

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

Kannada Raw Text Corpus

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

For Further contacts: oic@ldcil.org

Authors:
Vijayalaxmi F. Patil
Chetan S Baji
Rajesha N.
Manasa G.
Dr. Narayan Choudhary
Dr. L. Ramamoorthy

Last updated: December 12, 2018

ISBN 978-81-7343-226-2 Electronic publication of Corpus

Publisher:
Linguistic Data Consortium for Indian Languages
Central Institute of Indian Languages
Mysore, India-570006

Contents

Table of Figures

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 acquisition[1]."

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>			
</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 KANNADA RAW TEXT CORPUS

9.1 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.

9.2 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 half letters (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 marks 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 ‘ಋ’ 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 agglutinativness. Kannada uses period as a sentence boundary maker just like in English and many other languages.

9.3 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.

9.4 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 9: 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 Investigator/Linguists working on the data collection had to deal with and get going.

9.5 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.

9.6 NORMALIZATION WORKSHOPS

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 Dec2012, 5th January 2015 – 6th March 2015, 1st June – 31st July 2015.

9.7 PROOF READING

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.

9.8 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.

9.9 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.

10 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.

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

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

Symbols

ಂ	ಁ	ಂ	ಃ	಼	ಽ
M'	m'	M	H	J	G

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

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.

11 OBSERVATIONS WORTH MENTIONING IN KANNADA CORPUS

11.1 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.

11.2 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 ‘ೱ’ 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 ೱ as in ೱಕ್ಟ್. To keep it true to text and LDC-IL Corpus has retained.

11.3 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
ರ್ಯಾಂಕ್	ರ + ್ + ZWJ + ಯ + ಾ + ಂ + ಕ + ್
(Accepted form)	
ರ್ಯಾಂಕ್	ರ + ್ + ಯ + ಾ + ಂ + ಕ + ್

(Non-accepted form)

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.

12 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 10: 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 11: 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.

12.1 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 12: Representation of Aesthetics

12.2 COMMERCE

The Commerce domain of Kannada text corpus covers 7 subdomains bearing a total of 2,07,053words 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 13: Representation of Commerce

12.3 MASS MEDIA

The Mass Media domain of 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 14: Representation of Mass Media

12.4 OFFICIAL DOCUMENT

The Official Document domain of Kannada text corpus covers a single sub-domain bearing 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 15: Representation of Official Document

12.5 SCIENCE AND TECHNOLOGY

The Science and Technology domain of Kannada text corpus cover 31 sub-domain bearing 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 16: Representation of Science and Technology

12.6 SOCIAL SCIENCE

The Social Science domain of Kannada text corpus cover 17 sub-domain bearing 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 17: 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