मानक छत्तीसगढ़ी मूल पाठ कॉर्पस

A GOLD STANDARD CHHATTISGARHI RAW TEXT CORPUS

A GOLD STANDARD CHHATTISGARHI RAW TEXT CORPUS

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
Mysuru, India-570006

CENTRAL INSTITUTE OF INDIAN LANGUAGES
Manasagangotri, Mysuru, Karnataka, India, 570006
www.ciil.org

Title: A Gold Standard Chhattisgarhi Raw Text Corpus
Authors: Ankita Tiwari, Satyaendra Kumar Awasthi, Rajesha N., Manasa G., Srikanth D., Stephen Fernandes, Nithin S., Narayan Kumar Choudhary, Shailendra Mohan

ISBN: 978-81-19411-64-1
CIIL Publication No.: 1437

First published: AD 2023September
Bhadrapada1945 Śaka

Publisher: Prof. Shailendra Mohan, Director, CIIL

Production Team
Head, Publication Unit: Umarani Pappuswamy
Officer-in-Charge, Publication Unit: Aleendra Brahma
Artist: H. Manohara
Staff-in-charge: R. Nandeesh
Compositor: M.N. Chandrashekar
Cover design: Ankita Tiwari

TABLE OF CONTENTS

FIGURES

TABLES

CHHATTISGARHI RAW TEXT CORPUS

Ankita Tiwari, Satyaendra Kumar Awasthi, Narayan KumarChoudhary

1.1 RAW TEXT CORPUS: AN OVERVIEW

A corpus is a vast collection of appropriate representation of expressions in a language, either in written or spoken form. The electronic text corpus is a collection of texts from various languages that have been chosen for linguistic research as a source of data. Textual materials best reflect a language or linguistic variant according to external standards. Text Corpus is one of the key resources for language technology.

A corpus serves as a fundamental resource for numerous modules within a variety of applications, such as grammar checkers and spell checkers commonly used in word processing software. Indian languages often present formidable challenges in the realm of Natural Language Processing (NLP) and Artificial Intelligence (AI). The field of language technology frequently encounters obstacles due to the scarcity of language resources. To foster the development of language technology for mass-application tools, there is a consistent demand for the long-term availability of substantial linguistic data. However, it is imperative that this data is collected, organised, and stored in a manner that accommodates the diverse needs of tech developers.

Over the years, many efforts have been made to generate text corpus in Indian languages and various organisations, including government agencies, academic institutions, and private bodies, have made their contributions. The text corpus generated under the LDC-IL ambit is intended to bridge the gap by offering more and more electronic data for the NLP/AI and language technology community. Therefore, promoting technology and increasing its availability could mark a turning point for wider use of Indian languages through technology.

1.2 INTRODUCTION TO CHHATTISGARHI

LDC-IL extended its efforts to promote non-scheduled languages apart from the 22 scheduled languages, commencing with the collection of written material in Chhattisgarhi which has the ISO639-3 Notation of hne 639-3[1]. The prime goal of LDC-IL is to gather the written material to generate a quality text corpus to emphasise the significance of mother tongues.

A corpus effectively showcases the linguistic intricacies and distinctive features of a language when it achieves two critical criteria: substantial size and a genuine representation of diverse domains. Text corpus of a language is an asset to scientifically explore the language and locate reliable evidence for features of the language, thus helping us comprehend its explicit attributes.
Chhattisgarhi, a tongue of approximately 17 million people, carries profound cultural and historical significance within the region of Chhattisgarh. It goes beyond being merely a means of communication; Chhattisgarhi serves as a powerful conduit for the people of this region to convey their time-honoured traditions and captivating stories. Recognized by the Indian

[1] https://iso639-3.sil.org/code/hne

government as an eastern variant of Hindi, Chhattisgarhi primarily finds its voice in the state of Chhattisgarh. According to the 2011 Census, an estimated 16.2 million individuals were identified as Chhattisgarhi speakers.[2]

Since the year 2007, Government of Chhattisgarh has bestowed official status upon Chhattisgarhi, acknowledging its importance in the state's cultural tapestry. Additionally, Chhattisgarhi shares remarkable linguistic and cultural affinities with other languages such as Bagheli, Bhojpuri, Magahi, Jharkhandi (Nagpuriya and/or Sadari Korwa), Odia, Dravidian, Marathi, Bundeli, and Bhojpuri. This linguistic diversity reflects the rich mosaic of cultures and traditions that thrive in this vibrant corner of India. (Pathak & Verma 2018)

The Chhattisgarhi Raw Text Corpus endows an unrivalled window in documenting the colloquialisms, idioms, regional vocabularies, and grammar that are essential to establishing frameworks for linguistic processing. The Chhattisgarhi Raw Text Corpus is an extensive repository encapsulating the viable linguistic elements of Chhattisgarhi textual materials.

1.3 CHHATTISGARHI RAW TEXT CORPUS CATEGORIZATION

The LDC-IL text corpus illustrates language use as it manifests in day-to-day life rather than providing speculative or idealised instances. In order to fulfill these criteria, it is necessary to collect an enormous volume of data from various domains. Otherwise, frequently occurring contents might come from a certain lexicon or style. A large volume of data provides somewhat accurate results of what occurs frequently and what occurs rarely in a language.

In terms of form, function, content, and features, each corpus text source differs from the other. The Chhattisgarhi Raw Text Corpus captures the data from the Aesthetic and Official Document domains. The texts are extracted from various literary works that capture a vast range of literary terms. The textual materials are collected from the standard textbooks that are descriptive in nature. These textbooks demonstrate the style of Chhattisgarhi of a particular time period. Apart from standard textbooks, we also collected texts from creative writings to generate the Chhattisgarhi Raw Text Corpus. It comprises novels, plays, short stories, children's story books, mythology, culture, essays, biographies, folk tales, general news and editorials etc.

#	Category	Sub Category	Word Count	Character Count
1	Aesthetics	Culture	105195	512709
2	Aesthetics	Folk Tales	54682	240359
3	Aesthetics	Literary Texts	313203	1480820
4	Aesthetics	Literature-Children's Literature	7960	36682
5	Aesthetics	Literature-Novels	409982	1909127
6	Aesthetics	Literature-Plays	150523	712267
7	Aesthetics	Literature-Short Stories	389140	1823611
8	Aesthetics	Mythology	4982	25278
9	Mass Media	Editorial	38266	175216
10	Mass Media	General News	563	2633

Table 1: Domains and their sub-categorisation of Chhattisgarhi

[2] https://censusindia.gov.in/nada/index.php/catalog/42458

1.4 LDC-IL TEXT CORPUS METADATA

The metadata information is the integral part of the entire data for linguistic analysis. The collected data is structured in accordance with its information called metadata that involves the data's category, sub-category, text title, author, source, publisher, year of publication, page numbers, etc. This information is quite helpful for users to access data in the database or repository. Metadata provides the detailed information; how the data was initially created and what it contains. The following table shows a brief description of LDC-IL metadata information.

#	Legend	Description
1	Filename	Represented by the "docID" tag in the XML files. This is a unique file number across the datasets.
2	Project Description	The LDC-IL team collected texts from the field in the form of books, newspapers, magazines, and journals for the Chhattisgarhi data by visiting universities, colleges, publishing companies, and individual authors.
3	Sampling Description	The corpus may employ this information for verifiable proof. It will provide details about related book page numbers selected for the corpus. (In the case of Chhattisgarhi, due to the limited availability of text resources, all the resources made available to us were selected in entirety to be a part of the text 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 articles are provided here.
8	Volume	Specifies volume number of the title, if any.
9	Issue	Specifies issue number of 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 UTF-8 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 publication 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 digitised/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 in which the text is written.

Table 2: Metadata Legends for LDC-IL Text Data[3]

[3] https://data.ldcil.org/upload/pubs/LDCIL_Release_Documentation.pdf

1.5 CHHATTISGARHI TEXT CORPUS NAMING CONVENTIONS

Collected Chhattisgarhi hardcopies in Devanagari scriptare digitised in to XML format, adhering to Unicode standards. Each corpus file has a unique filename, indexing it. These files are usually excerpts from specific book titles, named per language and source conventions. Naming convention uses notations and a five-digit number for distinct filenames.

For example, the text taken from the Chhattisgarhi book for LDC-IL Chhattisgarhi Text Corpus always starts with 'CH' followed by 5-digit numbers which is continuous, whereas text collected fromChhattisgarhi Magazine will start with 'CHM' followed by 5 digit numbers. If the source is from a Newspaper then 'CHN' notation will be followed where as if the News is taken from the Web source 'CHNW' will be used as a 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 a roman small letter suffixed and enclosed in braces.

Such file names could be 'CH00001(a)', 'CH00002(b)', 'CH00001(c)', 'CH00001(d)' etc.

1.6 METHODOLOGY

This section outlines the sequence of activities, beginning with data collection, followed by data processing methodologies, culminating in the creation of the final corpus compilation.

1.6.1 DATA SAMPLING

For languages in which there is ample data, LDCIL selects a few chapters from each book as part of the corpus to maintain uniformity in terms of domains and variety. In the case of Chhattisgarhi, due to the limited availability of literary text resources, all the resources made available to us were collected in entirety to be a part of the text corpus, avoiding the process of data sampling.

1.6.2 DATA COLLECTION

A team of 6 resource persons namely, Dr Satyendra Awasthi, Ankita Tiwari, Shantanu Jha, Saurabh Varik, Rupesh Pandey and Dr Srishti Singh was sent to Raipur and Bilaspur Districts of Chhattisgarh to collect data from the field in January 2023. The team visited the Universities, Colleges, Publication houses and individual authors, and collected texts in the form of books, Newspapers, Magazines and Journals from the field. The team also contacted some authors over e-mails and phone calls and managed to receive books from them through India Post.

1.7 DATA PROCESSING

This subsection describes each step involved in the processing of the collected data.

1.7.1 PROCESSING RAW DATA

The books underwent a scanning process where each page's image was assigned a distinct identifier. These images were subsequently uploaded onto the LDC-IL Data portal. The complete dataset was then categorised into multiple tasks based on character count, typically around 30,000 characters per task. These tasks were subsequently imported into the LDC-IL digitization platform, which possesses the capability to perform Optical Character Recognition (OCR) and extract the text content from each file.

1.7.2 SELECTION OF LANGUAGE EXPERTS AND DIGITIZATION PROCESS

The Chhattisgarhi text corpus is digitized and review is done by freelance language expert empanlledat CIILby evaluating the performance of each freelancerfor expertise on four levels; 1) spelling 2) grammar 3) punctuation and 4) spacing accuracy.The entire dataset was digitised by theempanelledlanguage expertsusing the same CIIL digitization platform. Once the Digitization process was complete, each task went through a rigorous review process to check the authenticity of the work using the same platform, before finalising the data.

1.7.3 DATA WAREHOUSING AND CORPUS CREATION

Once the review process is complete, the reviewed data is exported from the digitisation platform and the entire data is obtained in CSV format. The data is then warehoused using the in-house tool called Kanaja, along with the metadata information such as Book ID, page number, author information etc.

1.7.4 COPYRIGHT CONSENT

Copyright consent was sought from the copyright holders of the Chhattisgarhi texts during the process of data collection. Some of the copyright holders gave their consent later on via email. This copyright consent allows the Central Institute of Indian Languages (CIIL) to use electronic excerpts from the source text or book, excluding images or diagrams, at no cost. CIIL can use these extracts for language technology development related to the Chhattisgarhi, citing the source when distributing the text for research or commercial purposes. CIIL cannot sell the extracted content separately without prior consent but may use it commercially as part of a larger corpus for language technology development. The copyright owner retains all rights; CIIL has the consent to use its electronic form of text.

1.8 OVERVIEW OF REPRESENTED DOMAINS

The size of LDC-IL Chhattisgarhi Raw Text Corpus is 14,74,496 words corresponding to 69,18,702 characters, collected from 51 books.

1.8.1 Aesthetics and Mass Media

The Aesthetics category of Chhattisgarhi text corpus covers 8 sub-categories bearing a total of 14,35,667 words and the Mass Media category of Chhattisgarhi text corpus covers 2 sub-categories bearing a total of 38,829 words. The representational details are given in the chart.

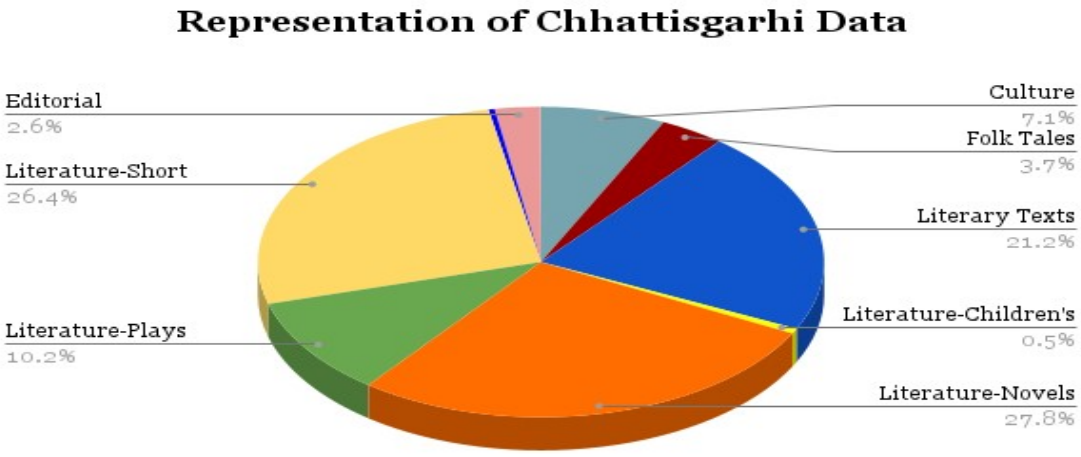

Figure 1: Representation of Chhattisgarhi Data

1.9 REFERENCES

1 Narayan Choudhary. 2019. Linguistic Resources for AI/NLP in Indian Languages.Central Institute of Indian Languages, Mysuru.https://data.ldcil.org/upload/pubs/LDCIL_Release_Documentation.pdf
2 Ethnologue.(n.d.).hne|ISO-639-3. Retrieved 09 15, 2023, from SIL:
3 https://iso639-3.sil.org/code/hne
4 Vinay Kumar Pathak,& Vinod Kumar Verma. (2018). Chhattisgarhi ka Sampurna Vyakaran. Bilaspur: Vadanya Publication.
5 Office of the Registrar General & Census Commissioner, India (ORGI).(2022, 04 22).Language. Retrieved from Census of India 2011: https://censusindia.gov.in/nada/index.php/catalog/42458