# मानक छत्तीसगढ़ी मूल पाठ कॉर्पस

## खंड - II

# A GOLD STANDARD CHHATTISGARHI RAW TEXT CORPUS

## VOLUME - II

# A GOLD STANDARD CHHATTISGARHI RAW TEXT CORPUS VOL. II


34

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

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

CENTRAL INSTITUTE OF INDIAN LANGUAGES

Manasagangotri, Mysuru, Karnataka, India, 570006
www.ciil.org

A Gold Standard Chhattisgarhi Raw Text Corpus Vol. II

Ankita Tiwari, Dr. Satyaendra Kumar Awasthi, Shantanu Kumar, Rajesha N., Manasa G., Dr. Narayan Kumar Choudhary, Prof. Shailendra Mohan

*e-ISBN: 978-93-48633-16-3*
*CIIL Publication No.: 1509*

*First published: AD 2025 March* :: *Phalguna 1946 Śaka*



*Publisher*: Prof. Shailendra Mohan, Director, CIIL

*Printing & Publication Unit*
Aleendra Brahma, Officer-in-Charge, Printing Press & Publication Unit
B Manju Bhargavi, Unit-in-Charge, Printing Press & Publication Unit
M.N. Chandrashekar, Compositor, Printing Press & Publication Unit

*Cover design:* Ankita Tiwari

# TABLE OF CONTENTS



# FIGURES



# TABLES

# 1 A GOLD STANDARD CHHATTISGARHI RAW TEXT CORPUS VOL. II

*Ankita Tiwari, Narayan Kumar Choudhary*

## 1.1 INTRODUCTION

The Linguistic Data Consortium for Indian Languages (LDC-IL) launched an initiative to conduct a fieldwork for the collection of Chhattisgarhi speech data in January 2023. Over time, this initiative expanded to include the compilation of textual resources for building the Chhattisgarhi Raw Text Corpus and released that dataset in January 2024. This initiative has set a significant precedent for other mother tongues as well.

As part of this ongoing initiative, LDC-IL has developed a new Chhattisgarhi dataset containing a substantial number of tokens from multiple domains. Whereas the earlier dataset concentrated largely on the aesthetic domain, the new dataset seeks to provide a more inclusive representation by incorporating data from additional areas, such as news, website and many more, to reduce the prior domain bias. This continuous commitment reflects LDC-IL's commitment and dedication to the preservation and promotion of mother tongues and enhancing the resources to develop language technologies for native languages.

The primary objective of Volume II of the Chhattisgarhi Raw Text Corpus is to collect a vast amount of written material from various domains to preserve the diversity and richness of Chhattisgarhi culture and traditions. And this second volume tries to capture the complexities and unique characteristics of the language by meeting two essential criteria: a substantial volume and an authentic representation of multiple domains.

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 selected in entirety to be a part of the text corpus, avoiding the process of data sampling.

Mr. Jayant Sahu, a distinguished author and owner/publisher of the Chhattisgarhi magazine Anjor, along with Mr. Sanjeev Tiwari, an advocate and writer, and Mr. Deen Dayal Sahu, a renowned writer from Chhattisgarh, played a significant role in the collection and development of the Chhattisgarhi text corpus (Vol. II) while working remotely as Resource Persons for LDC-IL. Their contributions were noteworthy in compiling Chhattisgarhi textual materials for the corpus. Additionally, they actively facilitated the LDC-IL team's engagement with writers, publishing houses, and individual authors through email and phone communication. Notably, Mr. Tiwari made a remarkable contribution by providing a substantial collection of Chhattisgarhi content published on his website. Their dedicated efforts were crucial in acquiring PDFs of books and magazines, thereby enhancing the corpus development process.

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 categorized 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.

Among the candidates who applied for the freelance language expert position at CIIL, those who met the eligibility criteria were interviewed to assess their proficiency in Chhattisgarhi. Candidates who successfully cleared the initial screening were then given sample tasks to assess their proficiency and accuracy when using the digitization platform. The performance of each candidate was evaluated by an expert on four levels; 1) spelling, 2) grammar 3) punctuation and 4) spacing accuracy. The candidates passing the test were selected for the task of digitization of books. The entire dataset was digitized by the selected candidates using the same tool against monetary remuneration. 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 finalizing the data.

A dedicated and diverse team of freelance language experts from different regions of Chhattisgarh, including Mr. Jayant Sahu, Mr. Sanjeev Tiwari, Ms. Hasina, Mr. Sanju Gupta, Mr. Sanjay Gangwal, Mr. Pramod Kumar, Ms. Pushplata, and Mr. Sandeep Dubey played a vital role in the Chhattisgarhi text digitization project. These experts worked remotely to digitize Chhattisgarhi texts and meticulously review the digitized content.

## 1.2 CHHATTISGARHI RAW TEXT CORPUS VOL. II

The LDC-IL Chhattisgarhi Raw Text Corpus comprises a total of 22,19,592 words, systematically gathered from 33 books, 22 monthly magazines and a website: https://gurturgoth.com/. The following table gives a summary of the Chhattisgarhi Raw Text Corpus.

| # | Category | Sub-Category | Word Count | Percentage (Within Sub-Domain) | Overall Percentage |
|---|---|---|---|---|---|
| 1 | Aesthetics | Cinema | 149 | 0.02% | 0.01% |
| 2 | Aesthetics | Culture | 6085 | 0.57% | 0.28% |
| 3 | Aesthetics | Folk Tales | 57440 | 5.34% | 2.59% |
| 4 | Aesthetics | Literary Texts | 94202 | 8.76% | 4.25% |
| 5 | Aesthetics | Literature-Children's Literature | 12893 | 1.20% | 0.59% |
| 6 | Aesthetics | Literature-Criticism | 36065 | 3.36% | 1.63% |
| 7 | Aesthetics | Literature-Novels | 115627 | 10.75% | 5.21% |
| 8 | Aesthetics | Literature-Plays | 16897 | 1.58% | 0.77% |
| 9 | Aesthetics | Literature-Short Stories | 408728 | 37.99% | 18.42% |
| 10 | Aesthetics | Literature-Text Books (School) | 7593 | 0.71% | 0.35% |
| 11 | Aesthetics | Literature-Travelogues | 561 | 0.06% | 0.03% |

| # | Category | Sub-Category | Word Count | Percentage (Within Sub-Domain) | Overall Percentage |
|---|---|---|---|---|---|
| 12 | Aesthetics | Mythology | 319613 | 29.71% | 14.39% |
| 13 | Mass Media | Cinema News | 4950 | 0.45% | 0.23% |
| 14 | Mass Media | Classifieds | 85 | 0.01% | 0.01% |
| 15 | Mass Media | Discussions | 1103 | 0.09% | 0.05% |
| 16 | Mass Media | Editorial | 97920 | 8.85% | 4.42% |
| 17 | Mass Media | General News | 931254 | 84.14% | 41.96% |
| 18 | Mass Media | Health | 29226 | 2.65% | 1.32% |
| 19 | Mass Media | Interviews | 2502 | 0.23% | 0.12% |
| 20 | Mass Media | Letters | 1744 | 0.16% | 0.08% |
| 21 | Mass Media | Political | 2713 | 0.25% | 0.13% |
| 22 | Mass Media | Religious/Spiritual News | 21931 | 1.99% | 0.99% |
| 23 | Mass Media | Social | 1360 | 0.13% | 0.07% |
| 24 | Mass Media | Sports News | 12022 | 1.09% | 0.55% |
| 25 | Science and Technology | Agriculture | 16585 | 97.11% | 0.75% |
| 26 | Science and Technology | Ayurveda | 494 | 2.89% | 0.03% |
| 27 | Social Sciences | History | 1116 | 5.63% | 0.06% |
| 28 | Social Sciences | Linguistics | 18734 | 94.38% | 0.85% |

Table 1: Representation of Sub-categories in Chhattisgarhi Raw Text Corpus Vol. II

This corpus is classified into four primary domains and further subdivided into 28 distinct sub-categories. The distribution of words across these domains is as follows: the aesthetic domain contains 10,75,853 words, the mass media domain includes 11,06,810 words, Science and technology domain includes 17,079 and the social sciences domain accounts for 19,850 words. Collectively, the corpus consists of 22,19,592 tokens. A comprehensive representation of this distribution is illustrated in the accompanying chart.

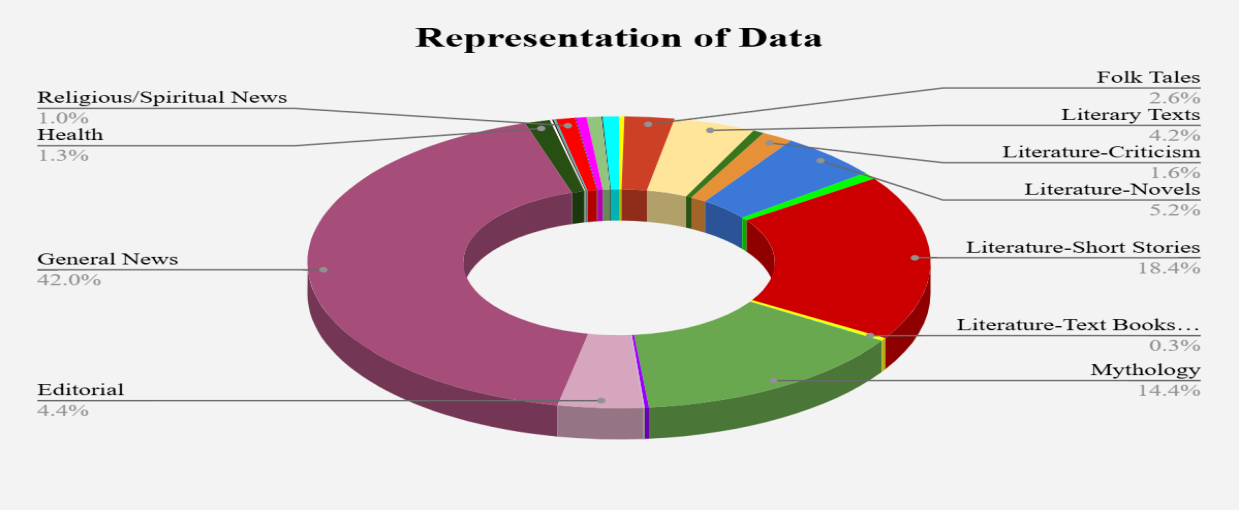


Figure 1: Representation of the Sub-categories in Chhattisgarhi Raw Text Corpus Vol. II

## 1.3 REFERENCES


1. Narayan Choudhary. LDC-IL: The Indian repository of resources for language technology. Lang Resources & Evaluation 55, 855–867 (2021). https://doi.org/10.1007/s10579-020-09523-3

2. 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

3. Ankita Tiwari, S.K. Awasthi , N. Rajesha, G. Manasa, D. Srikanth, N. K. Choudhary, Shailendra Mohan, 2023. A Gold Standard Chhattisgarhi Raw Text Corpus. Central Institute of Indian Languages, Mysore. ISBN: 978-81-19411-64-1: https://data.ldcil.org/chhattisgarhi-raw-text-corpus?tag=Chhattisgarhi