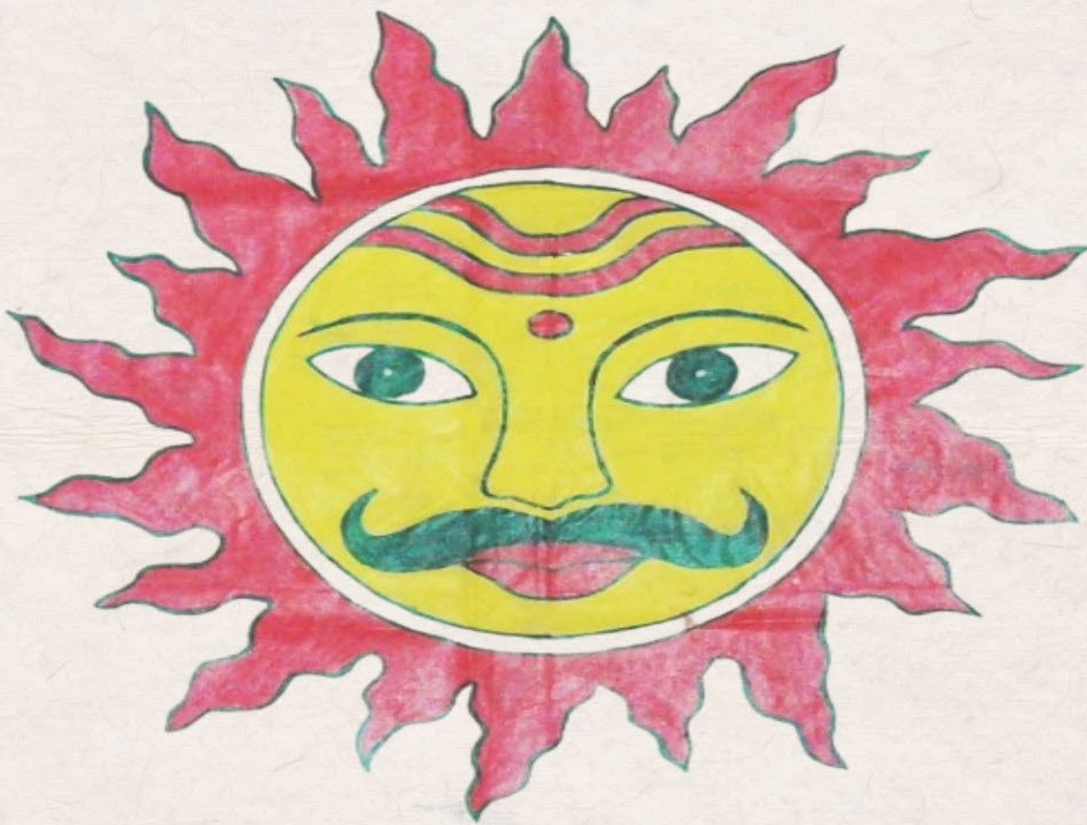


**मानक मैथिली
मूल पाठ कॉर्पस
खंड - II**



**A GOLD STANDARD
MAITHILI
RAW TEXT CORPUS
Volume - II**

A GOLD STANDARD MAITHILI 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 Maithili Raw Text Corpus Vol. II

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

e-ISBN: 978-93-48633-01-9

CIIL Publication No.: 1512

First published: AD 2025 March :: Phalguna 1946 Śaka

© Central Institute of Indian Languages, Mysuru 2025

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

Table of Contents	iv
Figures.....	iv
Tables.....	iv
1 Gold Standard Maithili Raw Text Corpus Vol. II	1
1.1 Introduction	1
1.2 Maithili Raw Text Corpus Vol. II	2
1.3 References	3

FIGURES

Figure 1: Representation of the Domains in Maithili Raw Text Corpus Vol. II	2
---	---

TABLES

Table 1: Representation of the Domains in Maithili Raw Text Corpus Vol. II.....	2
Table 2: Representation of Sub domains in Maithili Raw Text Corpus Vol. II	3

1 A GOLD STANDARD MAITHILI RAW TEXT CORPUS VOL. II

1.1 INTRODUCTION

In 2019, the LDC-IL released a Gold Standard Maithili Raw Text Corpus [1], which was developed from available contemporary texts. The corpus creation followed the pre-defined criteria which are mentioned at [2]. The corpus contained 5,316,552 Words and 29,658,942 characters drawn from 499 different titles, including the extracts from Magazine and newspapers that can be referred at [3]. It provides comprehensive representations of major domains such as aesthetics and social sciences, among others. The Maithili raw text corpus, in its original form, appears to be relatively small, limiting its effectiveness for comprehensive linguistic analysis and natural language processing tasks. To address this limitation, the Linguistic Data Consortium for Indian Languages (LDC-IL) has made significant enhancements to the corpus by incorporating additional domains of text. This expansion not only increases the overall size of the corpus but also diversifies its content, making it more representative of the variety of language use across different contexts.

Maithili is spoken by around 35 million people across India (Census 2011)¹. As per the linguistic survey made by GA Grierson [4], the Maithili language is spoken widely in the Bihar and Jharkhand states of India. It has several varieties [5]. A few studies mentioned above discuss that the variety of Maithili spoken in the regions of Bhagalpur and the neighboring district is called Angika by the native speakers [6]. The region carries a great cultural and historical significance and is widely used by millions of people across districts. Though the region is primarily claimed as the Angika-speaking region by several works [6] [7] and a few other individual research works, still, based on the government documents and the Census report of 2011[8] by the Govt of India, the area is considered as the language variety of Maithili. Hence, the area is referred to as the Maithili-speaking region in this document hereafter.

LDCIL selects specific chapters from each book to ensure a balanced representation of domains and variety within the corpus for languages with sufficient data. However, due to the limited availability of literary resources in this particular variety of Maithili, the entire content of all accessible materials was included, avoiding the process of data sampling.

In November 2022, a team of six resource persons- Shantanu Kumar, Rupesh Kumar Pandey, Akanksha Tiwari, Jyoti Kumari, Nikhil Kumar, and Mukesh Kumar was sent to various districts in Bihar and Jharkhand, including Bhagalpur, Banka, Jamui, Munger, Lakhisarai, Khagaria, Godda, and Deoghar, to conduct field work. The team visited villages, universities, colleges, and individual authors across these regions to gather texts in books, magazines, and journals. The team also contacted several authors via email and phone and successfully obtained books through India Post. During the fieldwork of the data collection, there was a huge support by the local authors and individuals in every aspect of the project. There were a few influential literary people who generously contributed directly or indirectly in collecting the text and speech corpus from the remote areas of the locality or from the places wherever possible.

¹ <https://censusindia.gov.in/nada/index.php/catalog/42458>

1.2 MAITHILI RAW TEXT CORPUS VOL. II

A Gold Standard Maithili Raw Text Corpus Vol. II is a comprehensive collection of Maithili language texts, comprising 8,11,680 words from 38 books and 16 magazines. This corpus includes extracts from books and magazines, providing a diverse range of linguistic data. It serves as a valuable resource for linguistic research, language processing applications, and the preservation of the Maithili language. The corpus has been meticulously compiled and is available for access through the Linguistic Data Consortium for Indian Languages (LDC-IL). Researchers and developers can utilize this resource to enhance their understanding and applications related to the Maithili language.

The representation of the three major domains covered has been shown in the table below:

Domain	Word count	Percentage
Aesthetics	596841	73.53%
Mass Media	184957	22.79%
Social Sciences	29882	3.68%
Total	8,11,680	100%

Table 1: Representation of the Domains in Maithili Raw Text Corpus Vol. II

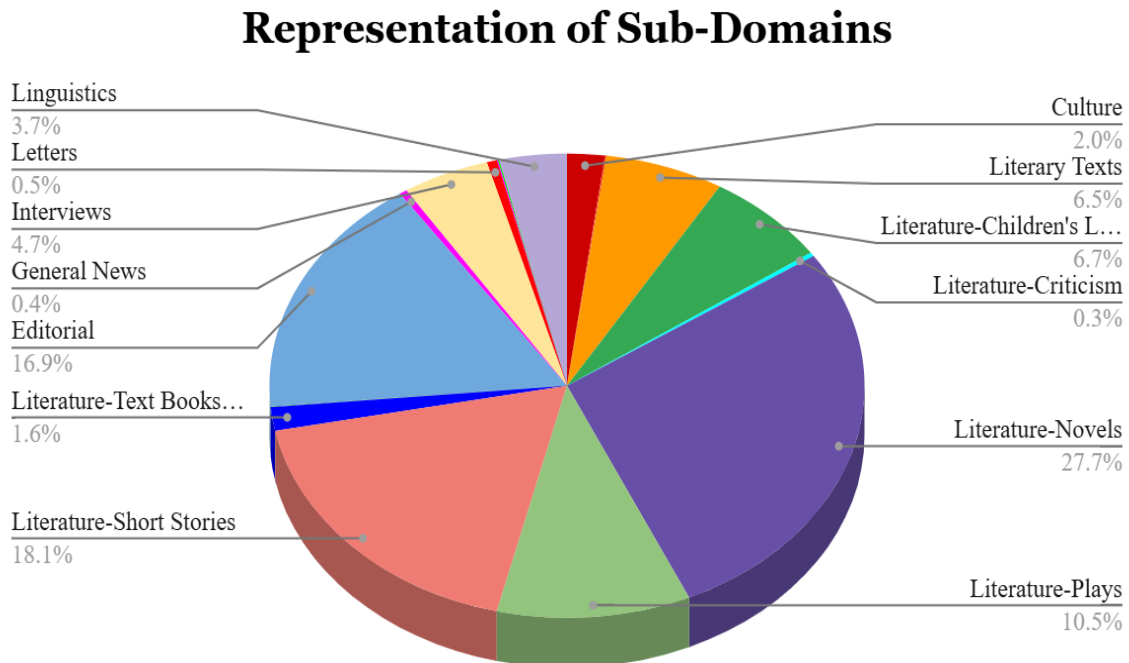


Figure 1: Representation of the Domains in Maithili Raw Text Corpus Vol. II

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.

Category	Sub Category	Word Count	Percentage (within Sub domain)	Overall Percentage
Aesthetics	Biographies	28915	3.71%	2.85%
Aesthetics	Culture	12930	1.66%	1.28%
Aesthetics	Fine Arts-Dance	288	0.04%	0.03%
Aesthetics	Folk Tales	7329	0.94%	0.72%
Aesthetics	Culture	16,540	2.78%	2.04%
Aesthetics	Folk Tales	608	0.10%	0.08%
Aesthetics	Literary Texts	52,499	8.80%	6.47%
Aesthetics	Literature-Children's Literature	54,426	9.12%	6.71%
Aesthetics	Literature-Criticism	2,192	0.37%	0.28%
Aesthetics	Literature-Novels	224,967	37.70%	27.72%
Aesthetics	Literature-Plays	85,393	14.31%	10.53%
Aesthetics	Literature-Short Stories	146,847	24.61%	18.10%
Aesthetics	Literature-Text Books (School)	13,369	2.24%	1.65%
Mass Media	Cinema News	321	0.18%	0.04%
Mass Media	Editorial	137,438	74.31%	16.94%
Mass Media	General News	3,372	1.83%	0.42%
Mass Media	Interviews	38,446	20.79%	4.74%
Mass Media	Letters	4,339	2.35%	0.54%
Mass Media	Obituary	565	0.31%	0.07%
Mass Media	Social	476	0.26%	0.06%
Social Sciences	Linguistics	29,882	100	3.69%

Table 2: Representation of Sub domains in Maithili Raw Text Corpus Vol. II

1.3 REFERENCES

1. Ramamoorthy, L., Narayan Choudhary & Dinesh Mishra. 2019. A Gold Standard Maithili Raw Text Corpus. Central Institute of Indian Languages, Mysore.
2. Choudhary, Narayan & L. Ramamoorthy. 2019. "LDC-IL Raw Text Corpora: An Overview" in Linguistic Resources for AI/NLP in Indian Languages, Central Institute of Indian Languages, Mysore. pp. 1-10.
3. Dinesh Mishra et.al. 2019. "Kashmiri Raw Text Corpus" in Linguistic Resources for AI/NLP in Indian Languages, Central Institute of Indian Languages, Mysore. pp. 73-81.
4. Grierson, GA. 1909, An introduction to the Maithili dialect of the Bihari language as spoken in North Bihar (2 ed.). Calcutta: Asiatic Society of Bengal. pp. xi-xiii.
5. Brass, P. R. (2005). Language, Religion, and Politics in North India. iUniverse, Lincoln, NE
6. Sharma, RM. 2007, Angika Bhasha Ka Dhwaniyaigyanik Adhyayan (In Hindi) (II ed.). Shri Tara Geeta Printing Press, Bhagalpur, pp. 130-133.
7. Jha, S. 1958, "The formation of the Maithili language." PhD diss., Luzac London. pp. 5-7.