

मानक राजस्थानी मूल पाठ कॉर्पस



A GOLD STANDARD
RAJASTHANI
RAW TEXT CORPUS



A GOLD STANDARD RAJASTHANI RAW TEXT CORPUS



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 Rajasthani Raw Text Corpus

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

e-ISBN: 978-93-48633-93-4

CIIL Publication No.: 1508

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 A Gold Standard Rajasthani Raw Text Corpus	1

FIGURES

Figure 1: Representation of the Sub-categories in Rajasthani Raw Text Corpus.....	4
---	---

TABLES

Table 1: ISO Codes for Mother Tongue	1
Table 3:Representation of Sub-Categories in the Rajasthani Raw Text Corpus	3

1 A GOLD STANDARD RAJASTHANI RAW TEXT CORPUS

Ankita Tiwari, Narayan Kumar Choudhary

1.1 INTRODUCTION

The Linguistic Data Consortium for Indian Languages (LDC-IL) is actively collecting written materials from various mother tongues. What began with Chhattisgarhi has now extended to Rajasthani. This continuous effort highlights LDC-IL's dedication and commitment to preserving and promoting mother tongues; and paving the way for their advancement in language technology.

The LDC-IL has chosen to release this dataset as the Rajasthani Raw Text Corpus, encompassing its recognized linguistic varieties, which are elaborated upon in the subsequent paragraph. Although the Rajasthani is listed as one of the mother tongues in the 2011 Census, it also appears to serve as a broad term encompassing the diverse languages spoken in Rajasthan, which may be collectively referred to as Rajasthani. This corpus aims to capture its linguistic diversity comprehensively.

A corpus effectively captures the linguistic nuances and unique characteristics of a language when it meets two essential criteria: a significant volume and an authentic representation across various domains. A language's text corpus serves as a valuable resource for scientific exploration, providing reliable evidence of its features and enabling a deeper understanding of its distinct properties.

The Rajasthan state hosts a wide array of linguistic forms that hold deep cultural and historical significance. More than just a means of communication, Rajasthani serves as a vital medium for preserving and transmitting the rich heritage, traditions, and folklore of Rajasthan. Rajasthani is recognized by the Government of India as a Western Indo-Aryan variant of Hindi, which is predominantly spoken within Rajasthan. According to the Census of India, 2011 approximately 51,968,635 individuals across India identified Rajasthani (including all the varieties) as their spoken language.¹ Rajasthani may be considered an umbrella term which includes a diverse linguistic group encompassing multiple mother tongues spoken across various regions of Rajasthan. A list of mother tongues with ISO codes listed below in the Table 3.

#	Mother Tongue	ISO 639-3 Language Code	Number of Speakers (Across India)
1	Rajasthani	(ISO 639-3: raj)	2,58,06,344
2	Marwari	(ISO 639-3: rwr)	78,31,749
3	Mewari	(ISO 639-3: mtr)	42,12,262
4	Wagdi	(ISO 639-3: wbr)	33,93,991
5	Hara/Harauti	(ISO 639-3: hoj)	29,44,356
6	Dhundhari	(ISO 639-3: dhd)	14,76,446
7	Bagri Rajasthani	(ISO 639-3: bgq)	2,34,227
8	Malvi	(ISO 639-3: mup)	52,12,617
9	Mewati	(ISO 639-3: wtm)	8,56,643

Table 1: ISO Codes for Mother Tongue

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

1.2 EXTRACTION OF TEXTUAL MATERIALS

For languages with abundant data, LDC-IL selects a few chapters from each book to ensure consistency across domains and linguistic variations within the corpus. However, in the case of Rajasthani, the limited availability of literary text resources necessitated the inclusion of most available materials in their entirety, eliminating the need for data sampling. Nonetheless, selective sampling was applied to certain books containing extensive tables and data, with only the relevant textual content being extracted. This process was carried out with careful consideration to ensure that paragraphs and sentences retained meaningful context.

1.3 DATA COLLECTION

Dr. Neeraj Kumar Daiya, a distinguished author from Rajasthan and a Resource Person for LDC-IL, contributed significantly to the collection of Rajasthani textual materials for the development of this corpus while working remotely on a Marwari translation project. He played a pivotal role in facilitating the LDC-IL team's outreach to universities, colleges, publishing houses, and individual authors through emails and phone calls, successfully assisting in the acquisition of PDFs of books and magazines.

1.4 PROCESSING OF 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 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.

1.5 SELECTION OF LANGUAGE EXPERTS AND DIGITIZATION PROCESS

Among the candidates who applied for the freelance language expert position at the CIIL, those who met the eligibility criteria were interviewed to assess their proficiency in Rajasthani. 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 various regions of Rajasthan, including Mr. Kishan Kumar Dodia, Mr. Satvir, Dr. Neeraj Daiya, Ms. Mamta Kumari, Mr. Puran Mal Bairwa, Mr. Ram Lal Bairwa, and Mr. Hoshang Panchal, played a crucial role in the Rajasthani text digitization project. The entire process of digitizing Rajasthani texts and reviewing the digitized materials was carried out remotely by these language experts.

The Rajasthani Raw Text Corpus constitutes a valuable resource for systematically documenting colloquial expressions, idiomatic usage, regional lexicon, and grammatical structures that are fundamental to the development of linguistic processing frameworks. This corpus functions as an extensive repository, preserving the essential linguistic characteristics of Rajasthani textual materials, thereby facilitating scholarly research and computational applications in language processing.

1.6 RAJASTHANI RAW TEXT CORPUS

The total volume of LDC-IL Rajasthani Raw Text Corpus is 11,99,502 words, systematically gathered from 12 books and 62 magazines. The following table gives a summary of the Rajasthani Raw Text Corpus.

#	Category	Sub-Category	Wordcount	Percentage (Within Sub-Domain)	Overall Percentage
1	Aesthetics	Biographies	1259	0.20%	0.11%
2	Aesthetics	Culture	50628	7.91%	4.23%
3	Aesthetics	Folk Tales	24253	3.79%	2.03%
4	Aesthetics	Literary Texts	201927	31.56%	16.84%
5	Aesthetics	Literature-Children's Literature	7721	1.21%	0.65%
6	Aesthetics	Literature-Criticism	10354	1.62%	0.87%
7	Aesthetics	Literature-Diaries	18842	2.95%	1.58%
8	Aesthetics	Literature-Letters	237	0.04%	0.02%
9	Aesthetics	Literature-Novels	1773	0.28%	0.15%
10	Aesthetics	Literature-Plays	28171	4.40%	2.35%
11	Aesthetics	Literature-Poetry	908	0.15%	0.08%
12	Aesthetics	Literature-Short Stories	158738	24.78%	13.24%
13	Aesthetics	Literature-Text Books (School)	120591	18.82%	10.06%
14	Aesthetics	Literature-Travelogues	13857	2.17%	1.16%
15	Aesthetics	Mythology	1535	0.24%	0.13%
16	Mass Media	Cinema News	24569	5.49%	2.05%
17	Mass Media	Classifieds	1687	0.38%	0.15%
18	Mass Media	Editorial	305484	68.19%	25.47%
19	Mass Media	General News	41215	9.20%	3.44%
20	Mass Media	Health	1056	0.24%	0.09%
21	Mass Media	Interviews	53944	12.05%	4.50%
22	Mass Media	Letters	16679	3.73%	1.40%
23	Mass Media	Obituary	137	0.04%	0.02%
24	Mass Media	Religious/Spiritual News	1660	0.38%	0.14%
25	Mass Media	Sports News	831	0.19%	0.07%
26	Mass Media	Weather	730	0.17%	0.07%
27	Social Sciences	History	110716	100	9.24%

Table 2:Representation of Sub-Categories in the Rajasthani Raw Text Corpus

The LDC-IL Rajasthani Raw Text Corpus comprises a total of 11,99,502 words. This corpus is classified into three primary domains and further subdivided into 27 distinct sub-categories. The distribution of words across these domains is as follows: the aesthetic domain contains 6,40,794 words, the mass media domain includes 4,47,992 words, and the social sciences domain accounts

for 1,10,716 words. Collectively, the corpus consists of 11,99,502 tokens. A comprehensive representation of this distribution is illustrated in the accompanying chart.

Representation of Data

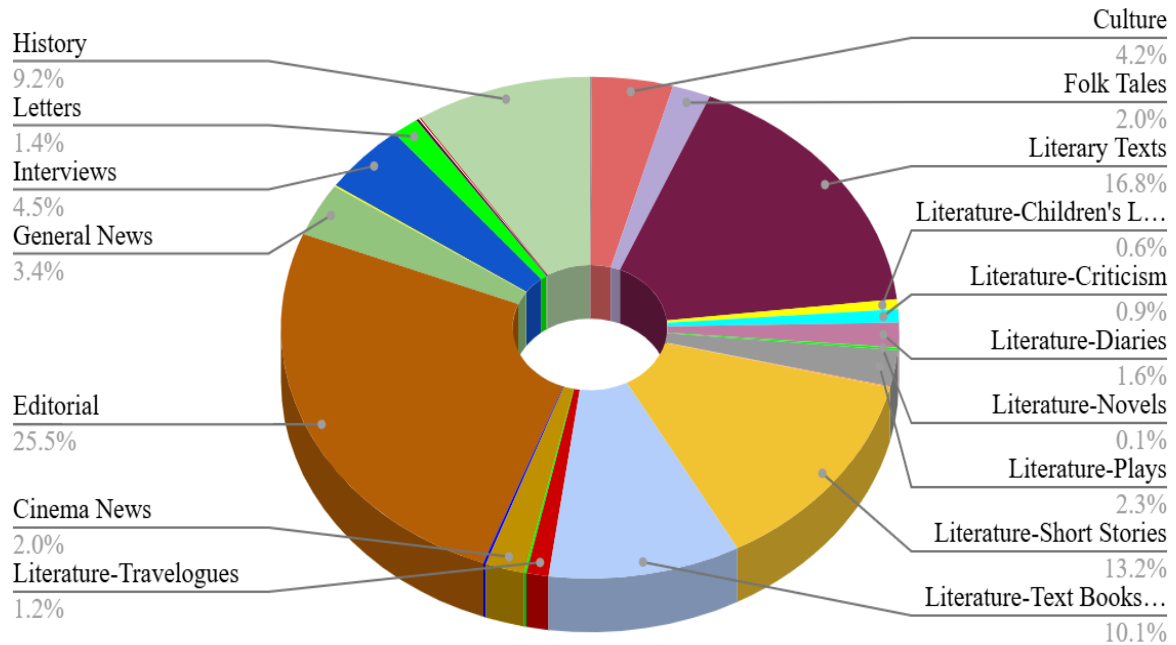


Figure 1: Representation of the Sub-categories in Rajasthani Raw Text Corpus

1.7 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. Ethnologue. (n.d.). hne|ISO-639-3. Retrieved 09 15, 2023, from SIL: <https://iso639-3.sil.org/code/hne>
3. 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>