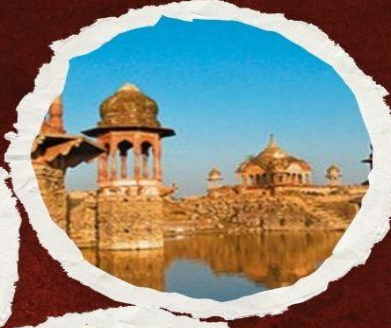




RAJASTHANI/MARWARI PARALLEL TEXT CORPUS: LINGUISTIC FEATURES AND STRUCTURES



RAJASTHANI/MARWARI PARALLEL TEXT CORPUS: LINGUISTIC FEATURES AND STRUCTURES



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.

Authors:

Ankita Tiwari

Dr. Rejitha K. S

Dr. Narayan Choudhary

Prof. Shailendra Mohan

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

भारतीय भाषा संस्थान

उच्चतर शिक्षा विभाग, शिक्षा मंत्रालय, भारत सरकार, मानसगंगोत्री, हुन्सूर रोड, मैसूरु-570006, कर्नाटक
CENTRAL INSTITUTE OF INDIAN LANGUAGES

Department of Higher Education, Ministry of Education, Government of India
Manasagangotri, Hunsur Road, Mysuru-570006, Karnataka

www.ciil.gov.in :: e-✉ ciil.publication@gmail.com :: ☎ +91-821-234-5182

Rajasthani/Marwari Parallel Text Corpus: Linguistic Features and Structures

Authors: Ankita Tiwari, Dr. Rejitha K. S., Dr. Narayan Choudhary, Prof. Shailendra Mohan

Contributors: Dr. Madan Gopal Ladha, Dr. Neeraj Daiya, Dr. Sharada Krishna

e-ISBN: 978-81-69175-62-3

CIIL Publication No.: 1742

First published: AD 2026 March :: Phalguna 1947 Śaka

© Central Institute of Indian Languages, Mysuru 2026

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

RAJASTHANI/MARWARI PARALLEL TEXT CORPUS	5
1.1 LDC-IL PARALLEL TEXT CORPUS.....	5
1.2 CHALLENGES IN CREATING A PARALLEL CORPUS.....	6
1.3 RAJASTHANI/MARWARI: A BRIEF INTRODUCTION	6
1.4 RAJASTHANI/MARWARI: A BRIEF LINGUISTIC OVERVIEW.....	8
1.5 SUMMARY OF RAJASTHANI/MARWARI PARALLEL CORPUS	9
1.6 REFERENCE	11

TABLES

TABLE 1: NUMBER OF SPEAKERS AND ISO 639-3 LANGUAGE CODE (CENSUS 2011).	8
---	---

RAJASTHANI/MARWARI PARALLEL TEXT CORPUS

Ankita Tiwari

1.1 LDC-IL PARALLEL TEXT CORPUS

LDC-IL has developed parallel corpus in Indian mother tongues. India has 270 mother tongues as per 2011 census. Most of these mother tongues have no written text readily available and they are mostly oral languages. Most of them do not have any script of their own or use one or more than one dominant script used in that region. Following the requirements of the NEP-2020, LDC-IL initiated to collect sentences from different mother tongues that could give an idea about the basic structure of these 270 mother tongues.

A parallel corpus is a collection of texts in two or more languages that are carefully aligned sentence wise in pairs and placed either side by side or matched with a common sentence/phrase identification number. It is widely used in linguistics, translation studies, and Natural Language Processing. Researchers use it to compare vocabulary, grammar, and meaning across languages, while computational systems rely on it to train statistical and neural machine translation models. Parallel corpora also support language learning by providing real translation examples and enable cross-language information retrieval, allowing users to access information in multiple languages efficiently.

LDC-IL developed a framework to accommodate all the linguistic features of mother tongues. The sentences were collected from Indian scheduled languages especially from descriptive grammar books. The sentences cover the general grammatical and structural types like adjectival phrases, adverbial phrases, compounding, Interrogative sentences, Imperative sentences, different speech types, statements, structural question, coordination, subordination and so on. Moreover, it includes linguistic features such as use of adposition, anaphora, derivational morphology, inflectional morphology and language specific features like emphasis, equative, possession, reciprocation, reflexive expression and so on. For more details on the source collection for this parallel corpus, please refer to the [LDC-IL Corpus Insights \(2025\)](#).

This parallel corpus can be used in AI development by enabling the creation of Indic multilingual foundation models, develop cross-lingual knowledge graphs and create low-resource language adapters. By providing high-quality linguistic data, it strengthens natural language understanding and generation. As India rapidly advances in the digital age, it is essential to preserve and promote undocumented mother tongues by bringing them onto digital platforms, ensuring they are sustained and able to thrive for future generations.

Low-resource languages have a significant impact on both AI development and digital world. When AI systems are trained mostly on high-resource languages, they become biased toward those linguistic and cultural contexts. Including low-resource languages helps build stronger multilingual foundational models. Exposure to diverse grammar structures, vocabulary, and linguistic patterns improves cross-lingual learning and generalization. Language carries cultural knowledge. Incorporating low-resource languages enriches AI systems with diverse worldviews, idioms, and traditional knowledge systems. If a language is not supported digitally, its speakers may be excluded from online education, e-governance, healthcare services, and digital commerce. When digital tools support local languages, more people can participate in the economy. Local languages often contain indigenous knowledge about agriculture, medicine, environment, and culture and therefore bringing them into digital ecosystem safeguards the linguistic treasures that language might contain.

Supporting low-resource languages makes AI more inclusive, intelligent, and representative of linguistic diversity. In India, where hundreds of mother tongues co-exist, investing in low-resource language technologies is essential for equitable digital growth.

1.2 CHALLENGES IN CREATING A PARALLEL CORPUS

Creating a Parallel Corpus for mother tongues in India involves several linguistic, technical, and socio-cultural challenges. The major difficulty is the lack of standardized orthography in many regional and tribal mother tongues. Most of the mother tongues use the prominent script used in their region. Some mother tongues may use multiple scripts or even a newly created writing systems, which complicates the corpus development. Another challenge is the scarcity of written and digitized resources. Many mother tongues are used for daily communication only.

Dialectal variation is another serious issue, as a single mother tongue may have numerous regional varieties with differences in vocabulary, pronunciation, and grammar. This makes it difficult to ensure consistency and representativeness in the corpus. Accurate translation is a concerned matter. Proficient bilingual speakers may be limited for low-resource languages. Maintaining semantic equivalence across mother tongues with different cultural contexts can be challenging. Sometimes equivalent words are not available due to the socio-cultural differences. These factors affect large-scale parallel corpus development in Indian mother tongues. The creation of reliable and balanced parallel corpora for Indian mother tongues is a complex and resource-intensive task.

1.3 RAJASTHANI/MARWARI: A BRIEF INTRODUCTION

The state of Rajasthan contains a variety of linguistic forms that are deeply connected to its cultural and historical background. Rajasthani/Marwari plays a pivotal role in preserving regional traditions, folk literature, and cultural heritage. As per Census of India (2011), Rajasthani/Marwari is grouped under Hindi. The Census records approximately 5,19,68,635 (including all the major varieties) speakers of Rajasthani/Marwari across the country. These varieties includes Marwari, Mewari, Mewati, Dhundhari, Harauti, Bagri, Wagdi and Malvi etc.

According to the Census, Rajasthani and Marwari are recorded as separate dialects. However, many language experts from Rajasthan argue that Rajasthani and Marwari are essentially the same, and that “Rajasthani” functions as an umbrella term encompassing the various speech varieties spoken across the state, such as Marwari, Mewari, Mewati, Dhundhari, Harauti, Bagri, Wagdi and Malvi etc. Although these varieties share common historical roots, they reflect regional and historical divisions, and full mutual intelligibility does not exist across all of them.

The table below presents the major Rajasthani tongues, along with their respective ISO Language codes and the number of speakers recorded across India.

#	Tongue	ISO 639-3 Language Code	Number of Speakers ¹ (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: Number of Speakers and ISO 639-3 Language Code (Census 2011)

1.4 RAJASTHANI/MARWARI: A BRIEF LINGUISTIC OVERVIEW

Rajasthani/Marwari follow Subject–Object–Verb (SOV) word order, consistent with the head-final structure typical of Indo-Aryan languages. In declarative sentences, the verb generally appears at the end of the clause. While this pattern is standard, some variation may occur in interrogative or emphatic contexts.

Grammatical relations are mainly expressed through postpositions, which function as case markers. These postpositions indicate semantic roles such as agent, recipient, and location, rather than relying only on inflectional endings.

The morphological structure of Rajasthani/Marwari language reflects typical Western Indo-Aryan patterns. Nominal forms carry markers for gender, number, and case, while verbal forms

¹ <https://censusindia.gov.in/nada/index.php/catalog/42561>

encode tense, aspect, and agreement features including person and number. In certain constructions, particularly in the perfective aspect, split ergative alignment is observed.

The core vocabulary of Rajasthani/Marwari largely consists of *tadbhava* words forms that developed from Sanskrit through gradual phonological and morphological changes in earlier linguistic stages. These inherited forms make up the stable base of everyday vocabulary, including terms related to numerals, kinship, body parts, and natural elements. Grierson, in his 1912 *Linguistic Survey of India*, systematically described the historical development of Rajasthani/Marwari and traced its vocabulary to earlier Prakrit sources. He noted that Rajasthani/Marwari preserved a significant portion of inherited lexical material, which supports mutual intelligibility among them, while at the same time showing clear divergence from Standard Hindi.

Devanagari is the principal script used for writing Rajasthani/Marwari. Earlier, it was also written in the *Mahajani* script, also known as *Modiya*, which was widely used for commercial purposes before the adoption of Devanagari.²

1.5 SUMMARY OF RAJASTHANI/MARWARI PARALLEL CORPUS

The Rajasthani/Marwari Parallel Text Corpus has been developed using Hindi as the source language. This Corpus connected with English and 146 mother tongues of India. Those mother tongues are Anal, Angami, Apatani, Are, Assamese, Awadhi, Bagri Rajasthani, Bagri, Bagheli/Baghel Khandi, Balti, Bengali, Bhadrawahi, Bharmauri/Gaddi, Bhojpuri, Bilaspuri Kahluri, Bodo, Brajbhasha, Bundeli/Bundel khandi, Chakru/Chokri, Chambeali/Chamrali, Chang, Chhattisgarhi, Chirri, Chungli, Churahi, Coorgi/Kodagu, Deori, Dhundhari, Dimasa, Dogri, Gangte, Garhwali, Garo, Gojri/Gujjari/Gujar, Gujarati, Gujar, Halabi, Handuri, Hara/Harauti, Haryanvi, Hindi Multani, Hindi, Irula/Irular Mozhi, Kabui, Kachchhi, Kangri, Kannada, Karbi/Mikir, Kashmiri, Khandeshi, Khari Boli, Khasi, Khezha, Khiemnungan, Khortha/Khotta, Kisan, Kodava, Kokbarak, Kolami, Koli, Kom, Konda, Konkani, Konyak, Koya, Kudubi/Kudumbi, Kuki, Kurmali Thar, Ladakhi, Lepcha, Liangmei, Limbu, Lotha, Lyngngam, Magadhi/Magahi, Maithili, Malayalam, Malvi, Manipuri, Mao, Mara, Maram, Marathi, Maring, Mech/Mechhia, Mewari, Mewati, Miri/Mising, Mishmi, Mizo, Mongsen,

² https://grokipedia.com/page/Rajasthani_languages

Monpa, Mundari, Muwasi, Nawait, Nepali, Nimadi, Nissi, Nocte, Odia, Pahari, Paite, Palmuha, Pania, Pawari/Powari, Phom, Pnar/Synteng, Pochury, Poula, Punjabi, Purkhi, Rai, Reang, Rengma, Rongmei, Sadan/Sadri, Sambalpuri, Sangtam, Sanskrit, Santali, Saurashtra/Saurashtri, Sema, Shina, Sindhi, Sirmauri, Sugali, Surjapuri, Talgalo, Tamil, Tangkhul, Telugu, Thado, Tibetan, Tikhir, Tripuri, Tulu, Urdu, Vaiphei, Wagdi, Wancho, Yerava, Yerukala/Yerukula, Yimchungre, Zeliang, Zemi, Zou.

The Rajasthani/Marwari dataset comprises 5,332 sentences/phrases, systematically structured according to 159 grammatical categories. The Rajasthani/Marwari portion of the corpus contains 31,096 words and a total of 1,43,551 characters. Moreover, the corpus is aligned with 146 Indian mother tongues in addition to English. Altogether, the multilingual parallel corpus encompasses 4,404,845 words (over 4.4 million tokens) and 23,374,289 characters (approximately 23.3 million).

1.6 REFERENCE

- [1] Saartje Verbeke. (2010). Postpositions and agreement in Rajasthani languages: A comparative analysis of alignment features in Marwari and Harauti. De Gruyter: https://www.academia.edu/115933326/Postpositions_and_agreement_in_Rajasthani_languages_A_comparative_analysis_of_alignment_features_in_Marwari_and_Harauti
- [2] <https://censusindia.gov.in/nada/index.php/catalog/42561>
- [3] https://grokipedia.com/page/Rajasthani_languages
- [4] Lakhan Gusain. (2004). Marwari. Lincom Europa.
- [5] Sajayan Chako and Liahya Nagwazah. (2012). SIL International.
- [6] Rejitha K. S. and Narayan Kumar Choudhary. (ed.). (2025). LDC-IL Corpus Insights. CIIL, Mysore. 978-93- 48633-33-0.