MEWATI PARALLEL TEXT CORPUS:
LINGUISTIC FEATURES AND STRUCTURES

MEWATI PARALLEL TEXT CORPUS: LINGUISTIC FEATURES AND STRUCTURES

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

Mewati Parallel Text Corpus: Linguistic Features and Structures

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

Contributors: Hanif Khan and Ammar Khan

e-ISBN: 978-81-69099-61-5
CIIL Publication No.:1600

First published: AD 2026 March :: Phalguna 1947 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

MEWATI 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 MEWATI: A BRIEF INTRODUCTION

Mewati, with no official script of its own, is considered a dialect of Hindi in the sense that it is linguistically defined as a Rajasthani language variety (Gusain, 2003, p. 1) and According to the Census of India (2001), the Government of India classifies Rajasthani as a dialect of Hindi for administrative and census purposes. Within this classification, Rajasthani comprises several major dialects, including Bagri, Wagri, Shekhawati, Mewati, Marwari, Mewari, Harauti, and Dhundhari. Mewati (ISO 639-3: wtm)[1] is an Indo-Aryan language grouped under Hindi as per Census. It is spoken mainly by the Meo ethnic community in the Mewat region, which extends across Alwar and Bharatpur districts of Rajasthan, Nuh district of Haryana, and adjoining areas of Uttar Pradesh and the Delhi National Capital Region. The 2011 Census of India recorded 8,56,643 individuals reporting Mewati as their mother tongue.

Mewati is largely sustained through oral tradition, with a strong heritage of folklore, ballads and epic tales passed down across generations. These traditions hold significant importance in the cultural and religious practices of the Meo community. Although the language is mainly spoken rather than written, it is recorded in both the Devanagari and Perso-Arabic scripts. The use of two scripts reflects the mixed cultural and linguistic environment of its speakers.[2]

[1] https://iso639-3.sil.org/code/wtm
[2] https://grokipedia.com/page/Mewati_language

1.4 MEWATI: A BRIEF LINGUISTIC OVERVIEW

Mewati marks three grammatical cases in nouns: the direct case, which is unmarked and used for subjects and objects; the oblique case, which appears before postpositions; and the vocative case, used for direct address. Grammatical relationships such as dative, genitive, and locative are mainly expressed through postpositions attached to the oblique form of the noun. Nouns, pronouns, and adjectives in Mewati inflect for two genders (masculine and feminine) and two numbers (singular and plural). Adjectives agree with the nouns they modify in both gender and number. The basic sentence structure in Mewati follows a Subject–Object–Verb (SOV) order. The language shows a considerable degree of mutual intelligibility with Hindi–Urdu, especially at the lexical level.

Dialectal variation in Mewati reflects geographical location and language contact. Northern varieties, particularly those spoken around the Nuh district, tend to use postpositions more frequently in locative constructions, while southern varieties retain older Rajasthani-type forms that function similarly to prepositions. The core vocabulary of Mewati is primarily Indo-Aryan in origin, shared with other Rajasthani varieties and historically derived from Prakrit and Sanskrit. At the same time, the language has adopted many loanwords from Arabic, Persian, Urdu, and English due to historical contact, Islamic cultural influence in the Mewat region, and modern developments.

Different semantic domains show distinct influences. Terms related to agriculture and kinship largely preserve indigenous Indo-Aryan and Rajasthani elements, including words such as beto (‘son’) and chori (‘girl’). In contrast, religious vocabulary is strongly influenced by Arabic and Persian sources, reflecting the Muslim heritage of the Meo community. This pattern illustrates Mewati’s development as a contact language within a multilingual setting. Mewati is written in both the Devanagari and Perso-Arabic scripts, reflecting the cultural and regional background of its speakers.[3]

[3] https://grokipedia.com/page/Mewati_language

1.5 SUMMARY OF MEWATI PARALLEL CORPUS

The Mewati 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, Chirr, Chungli, Churahi, Coorgi/Kodagu, Deori, Dhundhari, Dimasa, Dogri, Gangte, Garhwali, Garo, Gojri/Gujjari/Gujar, Gujarati, Gujari, 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, Miri/Mising, Mishmi, Mizo, Mongsen, Monpa, Mundari, Muwasi, Nawait, Nepali, Nimadi, Nissi, Nocte, Odia, Pahari, Paite, Palmuha, Pania, Pawari/Powari, Phom, Pnar/Synteng, Pochury, Poula, Punjabi, Purkhi, Rai, Rajasthani, 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 Mewati dataset comprises 5,332 sentences/phrases, systematically structured according to 159 grammatical categories. The Mewati portion of the corpus contains 31,268 words and a total of 1,44,110 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] Lakhan Gusain. (2003). Mewati. Lincom Europa

[2] Prerna Bakshi. (August, 2013). Language Policy, Politics and Ideology in Mewati: Comparative Case Studies of Mewati in Two School Types [Master's Thesis]. Department of Linguistics, University of Sydney.

[3] https://censusindia.gov.in/nada/index.php/catalog/42561

[4] https://grokipedia.com/page/Mewati_language

[5] https://iso639-3.sil.org/code/wtm

[6] Rejitha K. S. and Narayan Kumar Choudhary. (ed.). (2025). LDC-IL Corpus Insights. CIIL, Mysore. 978-93- 48633-33-0.