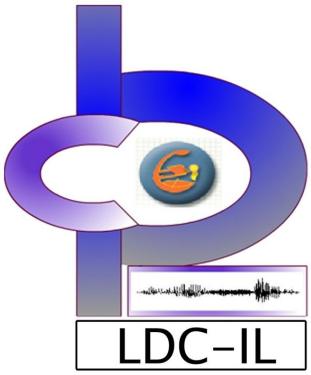
Kurmali Thar Parallel Text Corpus: Linguistic Features and Structures

KURMALI THAR 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:
Dr. Shahnawaz Alam
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

Kurmali Thar Parallel Text Corpus: Linguistic Features and Structures

Authors: Dr. Shahnawaz Alam, Dr. Rejitha K. S., Dr. Narayan Choudhary, Prof. Shailendra Mohan

Contributors: Dr. Hardeo Narayan Singh, Bijraj Mahto, Gyaneshwar Singh

e-ISBN: 978-81-69099-50-9
CIIL Publication No.: 1619

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: Shahnawaz Alam.

Table of Contents

KURMALI THAR PARALLEL TEXT CORPUS

Shahnawaz Alam

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 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 KURMALI THAR: A BRIEF INTRODUCTION

Kurmali Thar, also known as Kurmali or Kudmali, is an Indo-Aryan language spoken primarily in the eastern Indian states of Jharkhand, West Bengal, and Odisha. It is widely used among the Kurmi community and neighboring rural populations, functioning as an important medium of everyday communication and cultural expression. Linguistically, Kurmali Thar belongs to the Eastern Indo-Aryan group and shows structural affinities with Bengali, Odia, and Hindi. Its phonological and morphological systems reflect features typical of Indo-Aryan languages, including gender distinctions, number marking, case relations, and postpositional constructions. The language generally follows a Subject–Object–Verb (SOV) word order, with flexibility for pragmatic emphasis.

Kurmali Thar is primarily transmitted through oral traditions such as folk songs, tales, and community narratives. Although it is written in scripts such as Devanagari, Bengali, or Odia depending on the region, it lacks a fully standardized orthography. This variation in script usage reflects regional influences and contributes to challenges in formal documentation.

Socio-linguistically, Kurmali Thar plays a significant role in maintaining regional identity and cultural continuity. While dominant languages like Hindi, Bengali, or Odia are used in education and administration, Kurmali remains vital in informal and community domains. Recent academic and community-driven efforts emphasize documentation, literacy development, and digital preservation to strengthen its status within India’s multilingual landscape.

1.4 KURMALI THAR PARALLEL CORPUS

Kurmali Thar is a morphologically rich Indo-Aryan language characterized by distinct grammatical and phonological features shaped by prolonged regional contact and socio-historical developments. It exhibits significant morphological variation through inflectional mechanisms such as gender distinction, number marking, case relations, and the use of postpositions. Its verbal system encodes tense, aspect, mood, and agreement, aligning typologically with other Eastern Indo-Aryan languages. Kurmali Thar also demonstrates morphophonemic alternations, particularly in pluralisation, verb inflection, and lexical borrowing from neighbouring languages such as Bengali, Odia, and Hindi. The language generally follows a Subject–Object–Verb (SOV) word order, although pragmatic and discourse-driven variations are possible. It is written in multiple scripts, including Devanagari, Bengali, and Odia, depending on the region, and lacks a fully standardized orthography, reflecting its predominantly oral tradition and limited formal codification.

Kurmali Thar holds considerable potential for the development of parallel text corpora with other Indian languages, particularly Hindi and English. A Kurmali Thar parallel corpus would enable systematic research on syntactic divergence, dialectal diversity, lexical borrowing, and contact-induced linguistic change within India's multilingual environment. As a regionally significant yet under-resourced language, the creation of such a corpus would facilitate the study of cross-linguistic influence, translation strategies, and structural correspondences. Furthermore, it would support the advancement of natural language processing applications, machine translation systems, speech technologies, and bilingual lexicographic resources. Consequently, the development of a Kurmali Thar parallel corpus is essential not only for theoretical linguistic research but also for strengthening inclusive language technologies and preserving regional linguistic heritage.

This Kurmali Thar parallel corpus constitutes a systematically translated dataset derived from an original Hindi corpus, ensuring structured cross-linguistic alignment and semantic correspondence.

1.5 SUMMARY OF KURMALI THAR PARALLEL CORPUS

The Kurmali Thar parallel text corpus encompasses English alongside 146 other mother tongues of India. These mother tongues are Anal, Angami, Apatani, Are, Assamese, Awadhi, Bagheli/Baghel Khandi, Bagri Rajasthani, Bagri, 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, 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, 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.

There are 5,332 sentences/phrases in Kurmali Thar which are systematically structured based on 159 grammatical categories. Kurmali Thar has 28,825 word count and 148514 character count. The corpus is aligned with 146 mother tongues and English, comprising 4,404,845 words (over 4.4 million tokens) and 23,374,289 characters (approximately 23.3 million).

1.6 REFERENCES

[1] Rejitha K. S. and Narayan Kumar Choudhary. (ed.). (2025). LDC-IL Corpus Insights. CIIL, Mysore. 978-93- 48633-33-0.
[2] Census of India (2011). Language Data and Regional Language Distribution.
[3] George Abraham Grierson. Linguistic Survey of India.