Mundari Parallel Text Corpus:

Linguistic Features and Structures

MUNDARI 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:

Prangshu Manjul
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

Mundari Parallel Text Corpus: Linguistic Features and Structures

Authors: Prangshu Manjul, Dr. Rejitha K. S., Dr. Narayan Choudhary, Prof. Shailendra Mohan
Contributors: Dev Kumar Munda, Amar Topno, Hercules Singh Munda

e-ISBN: 978-81-69099-56-1
CIIL Publication No.: 1594

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: Prangshu Manjul

TABLE OF CONTENTS

MUNDARI PARALLEL TEXT CORPUS

Prangshu Manjul

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 MUNDARI: A BRIEF INTRODUCTION

Mundari (ISO 639-3: unr), a language from the Munda group of the Austroasiatic language family, reported to be “Vulnerable” by the UNESCO, spoken by approximately 1,128,228 people (2011 Census), is found in the Indian states of Jharkhand, Bihar, Chhattisgarh, Odisha, West Bengal, Madhya Pradesh, Assam and Tripura. In Odisha, Mundari speakers are located in Sundergarh, Mayurbhanj and Keonjhar districts. Significant speakers of Mundari language also reside in Bangladesh. The speakers of this language are fluent in the respective official languages of the above-mentioned states. Mundari is a non-scheduled language.

Rohidas Singh Nag created the “Mundari Bani” script, also known as Nag Mundari in the late 20th Century. The first book written in Mundari Bani was published in 2004, in handwritten form. In 2022, this script was encoded as a Unicode block called “Nag Mundari” (Unicode 15.0). Mundari language is also written in Devanagari, Odia, Bengali and Roman scripts.

Hasadaʔ, Naguri, Tamaria, and Keraʔ are the primary varieties of Mundari. Keraʔ variant speakers are mainly concentrated in and around Ranchi city, the capital of the Indian state, Jharkhand.

The first Mundari language grammar was published by Jesuit priest John-Baptist Hoffmann in 1903.

1.4 MUNDARI PARALLEL CORPUS

Mundari is a non-tonal language. Keraʔ variety's core vocabulary is similar to the other three varieties but its verbal morphology is unique and noticeably different from the Hasadaʔ, Naguri, and Tamaria in relation to both derivation and inflection. Keraʔ and Hasadaʔ share the same vowels whereas the consonants in both the varieties are highly similar to their equivalents except for the checked palatal stop. Also both the varieties have three numbers: singular, dual and plural. In Keraʔ, word stress is nondistinctive. Post-positions is another feature of this language. Gender differences are not typically marked in Mundari. In a simple unmarked sentence, it has subject+ object+ verb (SOV) word order.

These features contribute to Mundari's uniqueness among the languages of India and highlight its importance in the study of linguistics.

Mundari is a valuable resource for developing parallel text corpora with other Indian mother tongues. This parallel text corpus can serve as a valuable research resource for studying syntactic divergence, morphological richness, code-mixing phenomena, and script variation, which are particularly relevant in multilingual contexts such as India.

1.5 SUMMARY OF MUNDARI PARALLEL CORPUS

The Mundari parallel text corpus connected with English and 146 mother tongues of India. Those 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, Kurmali Thar, Ladakhi, Lepcha, Liangmei, Limbu, Lotha, Lyngngam, Magadhi/Magahi, Maithili, Malayalam, Malvi, Manipuri, Mao, Mara, Maram, Marathi, Maring, Mech/Mechhia, Mewari, Mewati, Mising, Mishmi, Mizo, Mongsen, Monpa, 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 Mundari which are systematically structured based on 159 grammatical categories. Mundari has 31,470 word count and 166,724 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 REFERENCE

[1] Kera? Mundari in Living Tongues Institute For Endangered Languages. https://livingtongues.org/kera-mundari/#:~:text=Kera%C9%81%20Mundari%20is%20a%20variant%20of%20the,varieties%20in%20terms%20of%20derivation%20and%20inflection. accessed March 19, 2026.

[2] Kobayashi, Masato and Ganesh Murmu. (2008). Kera? Mundari. In Anderson, Gregory D.S. (ed.), *The Munda languages*, 165-194. London & New York: Routledge.

[3] Linguistic Survey of India-Jharkhand. (2023). People's Archive of Rural India. Language Division, Office of the Registrar General, Government of India. https://ruralindiaonline.org/en/library/resource/linguistic-survey-of-india--jharkhand/ accessed March 19, 2026.

[4] Munda in Dr. Ramdayal Munda Tribal Welfare Research Institute, Jharkhand. https://www.trijharkhand.in/en/munda# accessed March 19, 2026.

[5] Mundari Bani in Atlas of Endangered Alphabets. https://www.endangeredalphabets.net/mundari-bani/ accessed March 19, 2026.

[6] Nedumaran, Muthu. (2023). *Font and Keyboard for Nag Mundari*. https://muthunedumaran.com/2023/03/16/nag-mundari/#:~:text=Until%20Unicode%2015%2C%20Nag%20Mundari,as%20garbled%20text%20in%20Latin. accessed March 19, 2026.

[7] Ota, AB and SC Mohanty and Soumalin Mohanty. (2016). *Mundari*. Scheduled Castes & Scheduled Tribes Research and Training Institute (SCSTRTI), Bhubaneswar.

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

[9] The Revd. J. Hoffmann. (1903). *Mundari Grammar*. Bengal Secretariat Press, Calcutta.

[10] Wolf-Sonkin, Lawrence and Biswajit Mandal. (2021). *Proposal to Encode the Mundari Bani Script in the Universal Character Set*.