Chhattisgarhi Parallel Text Corpus:
Linguistic Features and Structures

CHHATTISGARHI 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

Chhattisgarhi Parallel Text Corpus: Linguistic Features and Structures

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

Contributors: Jayant Kumar Sahu, Sanjeev Tiwari and Dindayal Sahu

e-ISBN: 978-81-69099-69-1
CIIL Publication No.: 1665

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

CHHATTISGARHI 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 CHHATTISGARHI: A BRIEF INTRODUCTION

Chhattisgarhi (ISO 639-3 code: hne)[1] is an indigenous Indo-Aryan language primarily spoken in Chhattisgarh and nearby regions of Odisha and Madhya Pradesh. As part of the Eastern Hindi subgroup. Chhattisgarhi developed as a distinct vernacular in the historical region of Dakshina Kosala, corresponding to present-day Chhattisgarh, during the transition from Middle Indo-Aryan Prakrits to New Indo-Aryan languages, approximately between the 10th and 12th centuries CE. Its fundamental structure originates from Sanskrit via Prakrit intermediaries, featuring phonological simplifications and grammatical changes typical of this stage of linguistic evolution.

Chhattisgarhi is written in the Devanagari script and functions as the primary vernacular in major districts including Raipur, Bilaspur, Durg, Raigarh, Rajnandgaon, and Surguja. According to the 2011 Census of India, 1,62,45,190 people identified Chhattisgarhi as their mother tongue. [2]Since 2007, the language has held official status alongside Hindi in Chhattisgarh, being used in both administrative and educational contexts.

Chhattisgarhi exhibits considerable dialectal diversity, with major variants including central (Kedri), eastern (Utti), northern (Baidyuni), southern (influenced by Halbi), and western (Khairagadi), reflecting both regional and sociolinguistic differences. Although the Government of India classifies Chhattisgarhi as an eastern dialect of Hindi for census and administrative

[1] https://iso639-3.sil.org/code/hne
[2] https://censusindia.gov.in/nada/index.php/catalog/42561

purposes, international linguistic authorities, such as Ethnologue, recognize it as a separate language based on its lexical, phonological, and functional distinctiveness.[3]

1.4 CHHATTISGARHI: A BRIEF LINGUISTIC OVERVIEW

Chhattisgarhi predominantly follows a Subject–Object–Verb (SOV) word order, with the verb occurring in clause-final position in declarative sentences. This syntactic pattern aligns with the typical structure of Eastern Indo-Aryan languages. Grammatical relationships are mainly indicated through case markers and postpositions rather than fixed word order. As a result, although SOV is the unmarked and preferred structure, some flexibility in constituent arrangement is possible, particularly in spoken discourse, without causing ambiguity. Chhattisgarhi nouns inflect for two grammatical genders- masculine and feminine, and for two numbers- singular and plural. Inflectional patterns are determined by the noun's inherent gender and its stem form, which govern agreement within the clause.

From a lexical perspective, Chhattigarhi language reflects extensive contact with neighboring tribal communities. Borrowings from Dravidian languages, notably Gondi, occur in domains such as agriculture and environmental terminology. Interaction with Munda (Austroasiatic) languages has similarly introduced vocabulary related to ritual life and local ecology. These influences demonstrate the historically layered linguistic environment of central India. Overall, Chhattisgarhi combines an Indo-Aryan structural base with lexical contributions from non-Indo-European sources.

1.5 SUMMARY OF CHHATTISGARHI PARALLEL CORPUS

The Chhattisgarhi 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

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

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, 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 Chhattisgarhi dataset comprises 5,332 sentences/phrases, systematically structured according to 159 grammatical categories. The Chhattisgarhi portion of the corpus contains 31,973 words and a total of 1,39,986 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] https://censusindia.gov.in/nada/index.php/catalog/42458

[2] https://grokipedia.com/page/Chhattisgarhi_language

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

[4] Vinay Kumar Pathak,& Vinod Kumar Verma. (2018). Chhattisgarhi ka Sampurna Vyakaran. Bilaspur: Vadanya Publication

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