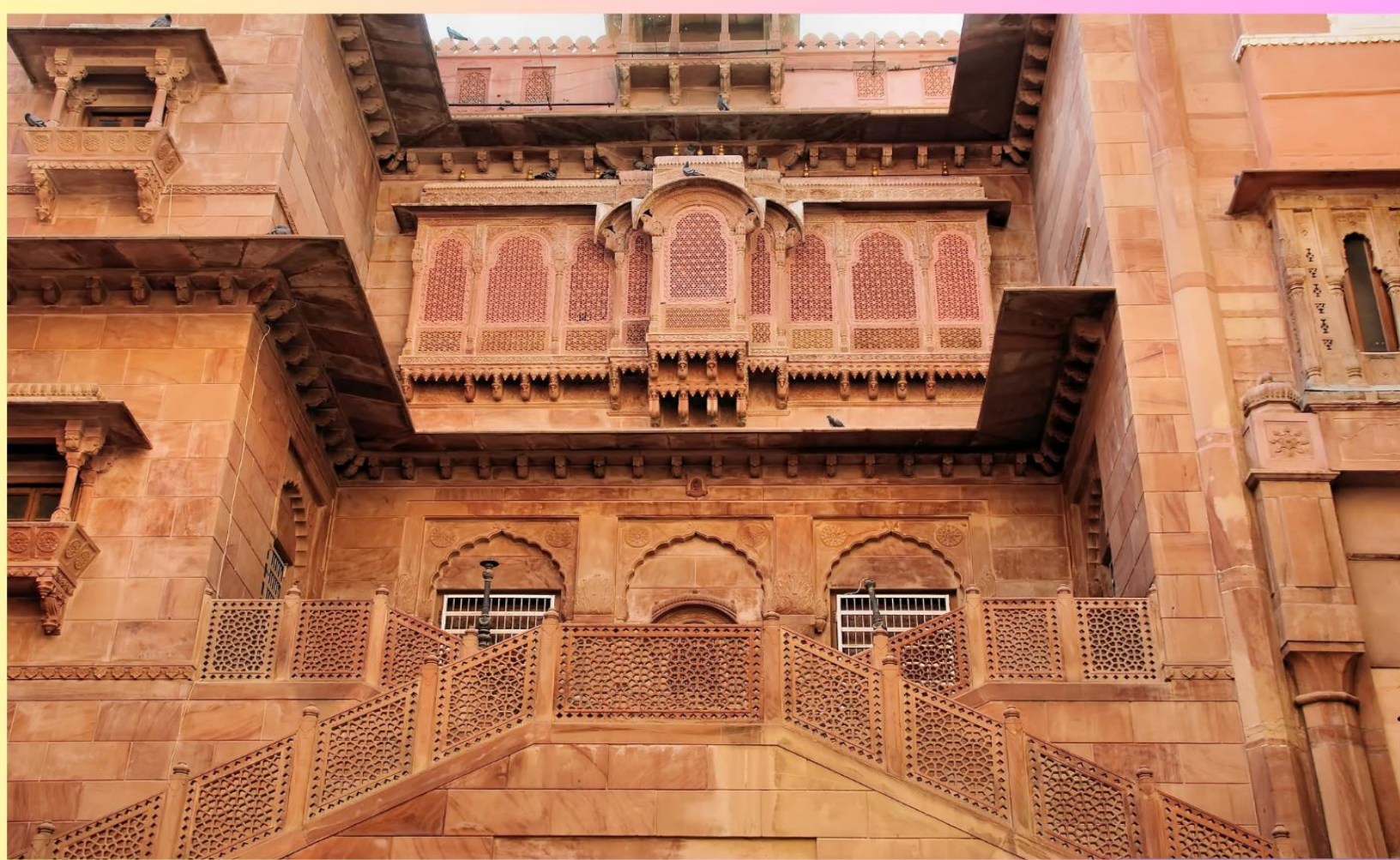


Dhundhari

Parallel Text Corpus: Linguistic Features and Structures



DHUNDHARI 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:

***Dr. Vijayalaxmi F Patil,
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

Dhundhari Parallel Text Corpus: Linguistic Features and Structures

Authors: Dr. Vijayalaxmi F Patil, Dr. Rejitha K. S., Dr. Narayan Choudhary, Prof. Shailendra Mohan

Contributors: Ramji Lal Bairwa, Puran Mal Bairwa

e-ISBN: 978-81-69099-18-9

CIIL Publication No.: 1659

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: Dr. Vijayalaxmi F Patil

Table of Contents

Dhundhari Parallel Text Corpus	1
1.1 LDC-IL Parallel Text Corpus.....	1
1.2 Challenges in Creating A Parallel Corpus.....	2
1.3 Dhundhari: A Brief Introduction.....	2
1.4 Dhundhari parallel Corpus	3
1.5 Summary of Dhundhari parallel Corpus	4
1.6 Reference.....	4

DHUNDHARI PARALLEL TEXT CORPUS

Dr. Vijayalaxmi F Patil

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 system, 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 DHUNDHARI: A BRIEF INTRODUCTION

The Dhundhar region, though not geographically vast, holds significant historical, cultural, and linguistic importance in Rajasthan. Covering approximately 40,349 square kilometers, it is located in the central-eastern part of the state. The region primarily includes Jaipur, Dausa, Tonk, and parts of Ajmer districts, and is partially surrounded by the Aravalli mountain range. The name “Dhundhar” is derived from the Dhund River, which flows through the area and gives the region its identity. From a linguistic perspective, Dhundhari is regarded as a rich and vibrant language. It is often considered a central variety among the languages of Rajasthan due to its wide usage and mutual intelligibility across the state.

Dhundhari belongs to the Indo-Aryan branch of languages and is an important member of the Rajasthani language group. It is predominantly spoken in the Dhundhar region of Rajasthan. Jaipur serves as the principal urban and cultural center of this region, where Dhundhari continues to be actively used in daily communication, folk traditions, oral narratives, and cultural expressions. Despite the increasing influence of Standard Hindi in education and administration, Dhundhari has retained its vitality in social and cultural life.

Historically, Dhundhari evolved from Shauraseni Prakrit, a language widely spoken in northern India during the early centuries CE. Over time, Shauraseni developed into various Apabhramsha forms, which between the 6th and 12th centuries CE gave rise to modern Rajasthani languages, including Dhundhari. According to the 2011 Census of India, 1,476,446 reported Dhundhari as their mother tongue. Most speakers are concentrated in the districts of Jaipur, Dausa, Tonk, and Sawai Madhopur. These figures indicate that Dhundhari remains actively used in everyday life, cultural activities, and oral traditions.

The linguistic features of Dhundhari reflect its Indo-Aryan origin and its historical development from Shauraseni Prakrit and Apabhramsha. Its phonological system, suffix-based morphology, honorific verb forms, and rich oral tradition distinguish it within the Rajasthani language group. Dhundhari includes both short and long vowels, and it preserves certain phonetic features from the Middle Indo-Aryan period. The use of retroflex, dental, and aspirated consonants is characteristic of the language. In rural varieties, the pronunciation of the sounds “r” and “l” exhibits distinctive regional patterns. Morphologically, Dhundhari expresses case, number, and gender through suffixation. Nouns and pronouns change form according to masculine and feminine gender as well as singular and plural number. Verbs inflect for tense, person, number, and gender, and the language exhibits ergative constructions in the past tense.

The typical word order in Dhundhari follows the Subject–Object–Verb (SOV) pattern. Honorific usage is reflected in changes to pronouns and verb forms based on social respect and hierarchy. The vocabulary of Dhundhari primarily derives from the Rajasthani linguistic base, with significant lexical influence from Sanskrit and Hindi. Its cultural richness is reflected in its idioms, proverbs, folk songs, and oral narratives. Traditionally, Dhundhari has been sustained through oral transmission, where folk poetry, storytelling, and songs have preserved its archaic grammatical forms and vocabulary.

In conclusion, Dhundhari is not merely a dialect but a powerful expression of the cultural identity, historical heritage, and social life of the Dhundhar region.

1.4 DHUNDHARI PARALLEL CORPUS

Dhundhari is a highly agglutinative language with rich linguistic features. It exhibits complex morphophonemic alternations, extensive case marking, productive compounding, and relatively flexible word order. Dhundhari language is written in Devanagari script derived from the Brahmi tradition. It has an extensive set of consonant and vowel symbols that precisely capture its phonemic distinctions. It generally follows Subject-Object-Verb pattern but it is linguistically rich and structurally complex language. These features contribute to Dhundhari’s uniqueness among the languages of India and highlight its importance in the study of linguistics.

Dhundhari 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 DHUNDHARI PARALLEL CORPUS

The Dhundhari 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, 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, Kurmal 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.

There are 5,332 sentences/phrases in Dhundhari which are systematically structured based on 159 grammatical categories. Dhundhari has 32,067 word count and 145,546 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] Rejitha K. S. and Narayan Kumar Choudhary. (ed.). (2025). LDC-IL Corpus Insights. CIIL, Mysore. 978-93- 48633-33-0.