Bagheli/Baghel Khandi Parallel Text Corpus:
Linguistic Features and Structures

BAGHELI/BAGHEL KHANDI 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

Bagheli/Baghel Khandi Parallel Text Corpus: Linguistic Features and Structures

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

Contributors: Priyanshu Kushwaha, Abhyuday Pratap Singh, Alankrita Singh

e-ISBN: 978-81-69099-24-0
CIIL Publication No.: 1680

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: Ankita Tiwari

TABLE OF CONTENTS

BAGHELI/BAGHEL KHANDI 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 BAGHELI/BAGHEL KHANDI: A BRIEF INTRODUCTION

The Bagheli/Baghel Khandi language is primarly spoken in the Baghelkhand region of Madhya Pradesh, particularly in the districts of Rewa, Satna, Sidhi, Shahdol, Umaria, and Anuppur. A considerable number of speakers are also found in adjoining areas of Chhattisgarh and Uttar Pradesh.[1] According to the Census of India, 2011, 26,79,129 individuals across India identified Bagheli/Baghel Khandi as their mother tongue. Bagheli, also referred to as Baghel Khandi, belongs to the Indo-Aryan language family and is classified as a dialect of Hindi as per Census of India.[2]

Grierson (1904: 122–182) classifies Bagheli/Baghel Khandi within the Indo-Aryan group, specifically under the Eastern Hindi subdivision. This classification is further supported by the comprehensive study conducted by Shukla (1972), which concurs with Grierson's categorization of the language. He states that the dialects Rimahim Powar, Kumhari Gahora, Judar, and Thirhari existed before the 12th century. Among these, Rimahim Powar and Kumhari were considered more prestigious than the others. He further explains that present-day Bagheli/Baghel Khandi has developed as a mixed form, resulting from the combination and interaction of these earlier dialects. Although Bagheli/Baghel Khandi shows variation based on region and caste, there is no clearly defined or dominant sub-dialect in its present form. However, the Rewa variety, commonly referred to as Rimahi, is generally accepted by speakers as the standard form of Bagheli/Baghel Khandi.[3]

1.4 BAGHELI/BAGHEL KHANDI: A BRIEF LINGUISTIC OVERVIEW

Bagheli/Baghel Khandi is officially classified as a separate language in the ISO 639-3 standard under the code *bfy*. It is genetically affiliated with the Indo-Aryan branch of the Indo-European language

[1] प्रो. सेवाराम त्रिपाठी, बघेली: अंतरंग और बहिरंग

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

[3] Shukla, who was a teacher in Rewa University, is a native of Rewa. Now retired, he is one of the main researchers on Bagheli

family and is further situated within the Eastern Central subgroup.[4] Bagheli/Baghel Khandi employs the Devanagari script as its standard writing system.

From a linguistic perspective, Bagheli/Baghel Khandi exhibits the inflectional morphological features characteristic of Eastern Hindi varieties. Nouns are inflected for gender, number, and case, with case relationships predominantly expressed through postpositions rather than through synthetic case endings. The language differentiates between masculine and feminine genders, and pluralization is typically achieved by the addition of suffixes. Grammatical functions such as genitive, locative, and instrumental are conveyed through specific postpositional markers. Bagheli/Baghel Khandi conforms to the typological characteristics of Central Indo-Aryan languages by adopting a canonical Subject–Object–Verb (SOV) word order in declarative sentences. In this structure, the verb generally occupies the clause-final position, and the subject typically precedes the object. Nevertheless, deviations from this order may arise due to pragmatic considerations such as topicalization or emphasis. Despite such flexibility, morphological markers help preserve syntactic clarity and overall comprehensibility.[5]

Bagheli/Baghel Khandi constitutes an important resource for the development of parallel text corpora alongside other Indian mother tongues. Such corpora can function as significant research tools for examining syntactic divergence, morphological complexity, code-mixing, and script variation. These areas of investigation are especially relevant in multilingual settings like India, where language contact and structural diversity play a central role in communication and linguistic change.

1.5 SUMMARY OF BAGHELI/BAGHEL KHANDI PARALLEL CORPUS

The Bagheli/Baghel Khandi 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, 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,

[4] https://iso639-3.sil.org/code/bfy
[5] https://grokipedia.com/page/Bagheli_language

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 Bagheli/Baghel Khandi dataset comprises 5,332 sentences/phrases, systematically structured according to 159 grammatical categories. The Bagheli/Baghel Khandi portion of the corpus contains 32,164 words and a total of 1,42,480 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] Binoy Koshy. (2022). A Sociolinguistic Study of Bagheli Speakers in Madhya Pradesh. Journal of Languages Survey Report. SIL International. 2766-9327

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

[3] https://www.ethnologue.com/language/bfy/

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

[5] Prof. Sewaram Tripathi. Bagheli: Antarang Aur Bahirang

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