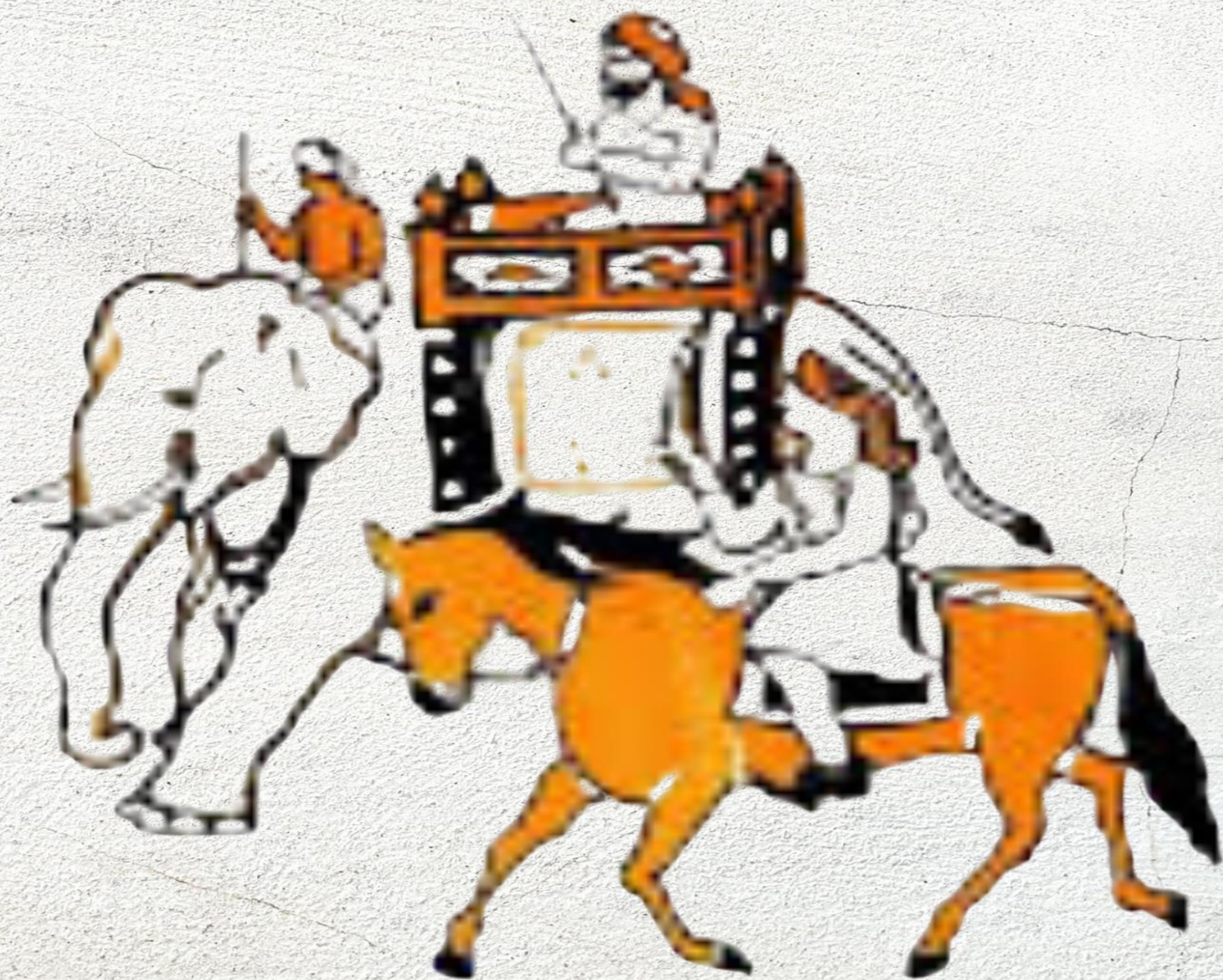


**BUNDELI/BUNDEL KHANDI PARALLEL TEXT
CORPUS: LINGUISTIC FEATURES AND
STRUCTURES**



BUNDELI/BUNDEL KHANDI 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. Satyaendra Kumar Awasthi

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

Bundeli/Bundel khandi Parallel Text Corpus: Linguistic Features and Structures

Authors: Dr. Satyaendra Kumar Awasthi, Dr. Rejitha K. S., Dr. Narayan Choudhary, Prof. Shailendra Mohan

Contributors: Anjana Kishanpuri, Shivanee

e-ISBN: 978-81-69099-59-2

CIIL Publication No.: 1669

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: Teena Verma

Table of Contents

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

BUNDELI/BUNDEL KHANDI PARALLEL TEXT CORPUS

Dr. Satyaendra Kumar Awasthi

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 BUNDELI/BUNDEL KHANDI: A BRIEF INTRODUCTION

Bundeli is an Indo-Aryan language and one of the most important and widely spoken regional languages of Uttar Pradesh and Madhya Pradesh. As per the 2011 Census of India, 56,26,356 individuals declared Bundeli as their mother tongue, which is grouped under the larger language category of Hindi. The present-day Bundelkhand region is divided between Uttar Pradesh and Madhya Pradesh. Uttar Pradesh includes the districts of Jhansi, Jalaun, Lalitpur, Mahoba, Banda, Hamirpur, Chitrakoot, and Banda, while Madhya Pradesh includes, Sagar, Chhatarpur, Tikamgarh, and Orchha.

This language is known as the cultural soul of the Bundelkhand region and is deeply connected to its folk life, history, and traditions. Historically, this region has been renowned for its heroism, folk tales, and cultural richness, the influence of which is clearly visible in Bundeli language and literature. Bundeli literature has primarily developed in the form of folk poetry and heroic ballads. "Alha-Khand" is Bundeli's most famous folk ballad, describing the valor of Alha and Udal. This ballad is still sung during the rainy season and on special occasions. Its powerful style and heroic language reflect the power and influence of Bundeli.

In addition, folk tales, hymns, ethnic songs, and proverbs are an important part of Bundeli literature. Rural life, nature, social relations, and folk beliefs have been prominent themes in Bundeli works. Even in modern times, some writers have attempted to keep it alive by writing poetry, plays, and stories in Bundeli.

Bundeli is the language of folk life. Wedding songs, Kajri, Phaag, Devi songs, and Varsha songs are still popular in villages. In particular, "Alha" singing has become a hallmark of Bundelkhand. Folk songs vividly depict heroism, love, devotion, and nature.

The historical context of Bundelkhand, such as the valor of Maharani Lakshmibai, is also expressed through Bundeli folk songs and stories. Regional pride and cultural identity have been preserved through the language.

1.4 BUNDELI/BUNDEL KHANDI PARALLEL CORPUS

Bundeli is a major language of the Indo-Aryan language family. Linguistically, it is considered one of the major dialects of the Western Hindi group. Bundeli is believed to have evolved from Shauraseni Apabhramsha, and its distinctive phonetic, morphological, and syntactic structure distinguishes it from standard Hindi.

A key feature of Bundeli is the vocalic shift, where the standard Hindi /a:/ ending often becomes an /o:/ ending; for example, the word jaisā (/dʒəisa:/) is pronounced as jaiso (/dʒəiso:/). There is also a tendency toward final vowel shortening in words like bhuk (/bʰʊk/) for hunger, haat (/ɦa:t/) for hand, and dood (/d̪u:d/) for milk. Other phonetic traits include the transformation of the dental /s/ into the aspirated /tʰ/, such as sidhi (/sidhi:/) becoming chhidhi (/tʰidhi:/). The use of 'ko' (ko), khaan (khaan), and khaan (khaan) instead of 'ke liye' (to bring) are some of the characteristics of this dialect.

The following are the main linguistic features of Bundeli:

- O-vowel predominance: Like Braj Bhasha, Bundeli is also 'o'-vowel predominant; 'o' is used extensively at the end of nouns and adjectives. This is clearly reflected in this data: kəo: 'kəo' is used in place of 'कहा', rəo: 'rəo' is used in place of 'रहा'.
- Case markers: Accusative- kʰõ, kʰɑ̃ (Hindi को), Locative- पे (Hindi पर) etc.
- Specific use of pronouns: In Bundeli, the standard Hindi pronouns वह (/vəɦ/) is replaced by बो (/bo:/). उसने (/usne:/) is replaced by ओने (/o:ne:/).
- Phonetic Changes: In Bundeli, common phonetic shifts from standard Hindi include the replacement of the postalveolar fricative /ʃ/ with the dental /s/. For example, rəme:ʃ changes to rəme:s and muruge:ʃ becomes muruge:s. Furthermore, the labiodental /v/ is replaced with the bilabial /b/, such as vəɦ becomes bo.

Bundeli 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 BUNDELI/BUNDEL KHANDI PARALLEL CORPUS

The development of Bundeli Parallel Corpus, Hindi was used as the source language which was translated into Bundeli. The Bundeli 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, Chakru/Chokri, Chambeali/Chamrali, Chang, Chhattisgarhi, Chirri, Chungli, Churahi, Coorgi/Kodagu, Deori, Dhundhari, 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, 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.

There are 5,332 sentences/phrases in Bundeli which are systematically structured based on 159 grammatical categories. Bundeli has 32,595 word count and 140,228 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] LangLex. (n.d.). Bundeli/Bundel khandi: Mother tongue profile (2011 Census report). Retrieved March 4, 2026, from <https://langlex.com/cens/MTPProfile.php?mtname=Bundeli%2FBundel+khandi>
- [2] Office of the Registrar General & Census Commissioner, India. (2018). C-16: Population by mother tongue (Table). Ministry of Home Affairs, Government of India. <https://language.census.gov.in/showMTSIdata>
- [3] Verma, D. (1954). Hindi bhasha ka itihās [History of the Hindi language]. Hindustani Academy.
- [4] Rejitha K. S. and Narayan Kumar Choudhary. (ed.). 2025. LDC-IL Corpus Insights. Central Institute of Indian Languages, Mysore. 978-93-48633-33-0.