

CHAMBEALI/CHAMRALI
PARALLEL TEXT CORPUS:
LINGUISTIC FEATURES AND
STRUCTURES



CHAMBEALI/CHAMRALI 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

Chambeali/Chamrali Parallel Text Corpus: Linguistic Features and Structures

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

Contributors: Dr. Abdul Kreem, Dr. Hardeep Singh, Shahnaz Akhtar

e-ISBN: 978-81-69099-36-3

CIIL Publication No.: 1667

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

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

CHAMBEALI/CHAMRALI PARALLEL TEXT CORPUS

Dr. Satyaendra Kumar Awasthi, Teena Verma

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 CHAMBEALI/CHAMRALI: A BRIEF INTRODUCTION

Chambeali is an ancient language of the Himalayan region, spoken in the Chamba district of Himachal Pradesh, which is situated to the north of the Kangra district. As per the census report of 2011, a total population of 1,25,746 reported their mother tongue to be Chambeali.

Like other Himalayan languages, Chambeali is also written using the Devanagari script. Though in the past, Takri script was used in the hilly areas, its use is now very limited. The Chamba region is also known for the Chamba Rumal, which is a form of hand embroidery on a kerchief. This form of art flourished in the early 18th and 20th century (Sharma, 2021). As is common with other highland languages in Himachal, written records are hard to find, which means the bulk of its literature has survived mainly through a long-standing oral tradition. Today, Chambeali is still woven into the fabric of daily life, heard in family homes and fields, sung in folk music, spoken during wedding rituals, and used to honor the ancient traditions of local gods and goddesses.

1.4 CHAMBEALI/CHAMRALI PARALLEL CORPUS

Chambeali is a well-defined Indo-Aryan language characterised by a clear grammatical structure and distinct phonetic features. It is a culturally rich language with clear gender and number markings, verb aspect markings and a systematic use of word-order. Despite its close relationship to Hindi, and its shared borders with various Tibeto-Burman influences, it maintains its distinct identity through its independent form, phonetics, and popular usage.

Some features of Chambeali/Chamrali are-

- **Phonetic Properties:** One very distinct feature of Chambeali/Chamrali is the high number of nasalised vowels in the vocabulary. For example, words such as [piʃ^hō] (behind), [rūnā] (to cry, weep), [tā] (them), [uʃ^hānā] (to raise) etc. This is possibly due to the language's affinity to the Tibeto-Burman group.
- **Morphological Properties:** Nouns undergo declension as gender and number in Chambeali are clearly defined through suffix changes. For example, [ch^hokəɾə] (boy.masc.sg) and [kuɾi:] (girl.fem.sg). Plurals are often formed with /ã/ or /e/, such as [kuɾijã] (girls.fem.pl). The verb system is built on a series of participles. The Present Tense is often formed using the present participle, while the Past Tense is formed from the past participle.
- **Syntactic Properties:** Sentences strictly follow a Subject-Object-Verb (SOV) order, a common feature across the Indo-Aryan Group.
- **Lexical Diversity:** The language's vocabulary shows visible influence from Hindi, Dogri and Punjabi, serving as a cultural bridge between the hills and the plains. Common lexical items include [puʈ^həɾ] (son), [ki:] (why), [kuɾi:] (a girl) and [səɾ] (the head).

Chambeali 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 CHAMBEALI/CHAMRALI PARALLEL CORPUS

For the development of the Chambeali/Chamrali Parallel Corpus, Hindi was used as the source language, which was translated into Chambeali/Chamrali. The Chambeali/Chamrali parallel text corpus is 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, 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, Sambalpur, 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 Chambeali/Chamrali which are systematically structured based on 159 grammatical categories. Chambeali has 31,853 word count and 155,017 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] Grierson, G. A. (1916). Linguistic Survey of India: Vol. IX, Indo-Aryan Family, Central Group, Part IV, Specimens of Pahārī Languages and Gujurī. Office of the Superintendent of Government Printing, India.
- [2] LangLex. (n.d.). Chambeali/Chamrali: Mother tongue and its regions (2011 Census report). Retrieved March 5, 2026, from <https://langlex.com/cens/mtregion.php?mtname=Chambeali%2FChamrali>
- [3] Office of the Registrar General & Census Commissioner. (2011). Census of India 2011: Language and Mother Tongue. Government of India.
- [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.
- [5] Sharma, H. (2021, February 7). Chamba rumal: The embroidery that resembles paintings! Satyam Fashion Institute. <https://satyamfashion.ac.in/blog/chamba-rumal-the-embroidery-that-resembles-paintings/>