



PAHARI PARALLEL TEXT CORPUS: LINGUISTIC FEATURES AND STRUCTURES



PAHARI 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 वभाग, 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

Pahari Parallel Text Corpus: Linguistic Features and Structures

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

Contributors: Syed Murtuza Hussain, Dr. Syed Mahir Hussain Jafri

e-ISBN: 978-81-69175-84-5

CIIL Publication No.: 1754

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

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

PAHARI 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 PAHARI: A BRIEF INTRODUCTION

The Pahari language of Jammu and Kashmir is one of the most beautiful languages that have been traditionally indigenous to the region and Himalayan foothills. It possesses its own distinct linguistic structure, as well as a rich cultural tradition, folklore, and historical heritage. This is a non-scheduled language spoken across various parts of Jammu and Kashmir, particularly in the districts of Poonch and Rajouri. It is also used in regions such as Hamirpur, Karsog, Kupwara, Mendhar, Nadaun, Nowshah, Nurpur, Rajauri and Theog. According to the Census of India 2011, the total number of Pahari speakers was recorded as 3,253,889. The Pahari is an Indo-Aryan language classified under the Western Pahari languages group.

According to Masoodi (1985), historical developments and related evidence indicate that Pahari is not a recently emerged language but possesses considerable antiquity, extending back nearly a millennium or more. The language is believed to have been in use since early human settlement in the Himalayan region. Pahari are deeply connected with their land and environment, their main food is maize, vegetables, rajma, curd and butter (lassi-makhan) their attire includes kurta, shalwar, waistcoat, and turban for men, and long frock-style dress with head covering, heavy

woolen shawl for women in winter. Pahari is rich in Folklore, and oral tradition, like, geet Mahiya (wedding song).

1.4 PAHARI PARALLEL CORPUS

Pahari is a well-formed Indo-Aryan language with a clear grammatical structure and good linguistic features. Pahari has SOV (subject-object-verb) word order. Pahari is a verb-final language; the verb is the final component of the sentence. However, the rich agreement system makes the word order free. Word order can be observed in the following data:

- mōhən-sən zahid-ā cɪtʰi lɪkʰ-i
Mohan-ERG Zahid-DAT letter write-PST
Mohan wrote a letter to zahid.
- mōhən-sən cɪtʰi zahid-ā lɪkʰi
Mohan-ERG letter Zahid-DAT write-PST

The pronominal system of the language exhibits distinctive structural features. The third-person pronoun ‘o’ (ओ) functions as a generalized form used for masculine, feminine, and, in certain contexts, plural reference. This indicates a reduced gender distinction in third-person pronouns compared to Standard Hindi. It can be observed in the following data:

- o ek məʃur pʰənəkər fɛ (She is a famous artist)
- o ek məʃur əd̪i:b fɛ (He is a renowned writer)
- o bəʃər ge fɛ (They have gone to the market)

Pahari 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 PAHARI PARALLEL CORPUS

The development of Pahari Parallel Corpus, English was used as the source language which was translated into Pahari. This Pahari 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, 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, 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 Pahari which are systematically structured based on 159 grammatical categories. Pahari has 32,270 word count and 139,064 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. (1919). *Linguistic survey of India* (Vol. 8, Part 2: Indo-Aryan family, North-Western group—Pahari). Calcutta: Office of the Superintendent of Government Printing, India.
- [2] Karnai, M. A. (2018). Pahari phonology and morphology: A linguistic study. Srinagar: Valley Publications.
- [3] Khan, A. Q. (n.d.). A preliminary study of Pahari and its sound system. Department of English, University of Azad Jammu & Kashmir.
- [4] LangLex. (n.d.). Pahari: Mother tongue and its regions (2011 Census report). Retrieved March 5, 2026, from <https://langlex.com/cens/mtregion.php?mtname=Pahari>
- [5] Masoodi, G. M. (2012). Sociolinguistic survey of Pahari in Jammu region. Srinagar: Valley Book House.
- [6] Rejitha K. S. and Narayan Kumar Choudhary. (ed.). 2025. LDC-IL Corpus Insights. Central Institute of Indian Languages, Mysore. 978-93-48633-33-0.