

KANGRI PARALLEL TEXT CORPUS:
LINGUISTIC FEATURES AND
STRUCTURES

KANGRI 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

Kangri Parallel Text Corpus: Linguistic Features and Structures

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

Contributors: Ravinder Singh, Nitish kumar

e-ISBN: 978-81-69099-99-8

CIIL Publication No.: 1641

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

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

KANGRI 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 KANGRI: A BRIEF INTRODUCTION

Kangri is a major language of the Himalayan region, spoken primarily in the Kangra district of Himachal Pradesh and surrounding areas. The language is deeply connected to local culture, folk life, and traditions. The Kangra region has historically been a rich cultural centre. Its temples, folk dances, folk songs, and fairs have preserved the Kangri language from generation to generation. As per the Census Report of 2011, around 27 states in India with a total population of 11,17,342 reported their mother tongue to be Kangri. With a majority of speakers residing in Himachal Pradesh, sub-districts such as Palampur, Dera Gopipur, Jawali, Nagrota Bagwan and Dharmshala are populated with Kangri speakers. However, it should be noted that Kangri is spoken in the whole Kangra region except Chhota and Bara Bangahal (Gopalakrishnan, 2022).

Linguistically, Kangri belongs to the Western Pahari branch of the Indo-Aryan language family. However, the linguistic foundation of Kangri is a unique hybrid, blending the structural framework of Shauraseni Prakrit with the ancient Khasya substratum. It is believed that the Khasya tribe in the Himalayan region came under Rajput rule and lost their original language through assimilation. This explains why Kangri shares grammatical similarities with North

Indian languages such as Hindi, but retains distinct phonetic and tonal qualities unique to the Himalayan region (Joshi, 2010). Historically, Takri served as the indigenous script for local use. However, over time Devanagari script became more prevalent, possibly due to reasons such as administrative uniformity, Hindi-medium education or the lack of printed or written materials. Furthermore, during the medieval period, the Kangra region was very significant. Its folk art and musical traditions lent stability to the language. Additionally, Kangra paintings and folk tales strengthened Kangri's cultural foundation.

1.4 KANGRI PARALLEL CORPUS

Kangri 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, it maintains its distinct identity through its independent form, phonetics, and popular usage.

Some features of Kangri are:

- **Phonetic Properties:** Kangri demonstrates the transformation of the sibilant /ʃ/ into /s/. For example, [ʃəkt̪i] → [səkt̪i:], [ʃer] → [ser]. Furthermore, one can find a large number of nasalised sounds in the Kangri's linguistic repository.
- **Morphological Properties:** Gender and number in Kangri, like Hindi, are clearly defined through suffix changes. For example, [cʰoru:] (boy.masc.sg), [cʰori:] (girl.fem.sg) and [cʰore/munde] (boys.masc.pl). Verbs inflect for tense, number and person. For example [mẽ ʃandə] (I go), [o əjo] (He came) and [t̪uf̪ā kərənə] (You will do it).
- **Syntactic Properties:** Sentences follow a Subject+Object+Verb (SOV) order.
- **Lexical Diversity:** The language's vocabulary shows visible influence from Hindi and Punjabi and serves as a cultural bridge. Some common words are [kuṭʰi] (where), [kinna] (how much), [nĩ:] (negation) and [paṇi:] (water), etc.

Kangri 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 KANGRI PARALLEL CORPUS

The development of the Kangri Parallel Corpus, Hindi was used as the source language, which was translated into Kangri. The Kangri 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, 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, 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 Kangri which are systematically structured based on 159 grammatical categories. Kangri has 31,135 word count and 154,850 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] Joshi, M.C. (2010). "On the origin of the neo Indo-Aryan Pahāḍī language." *Lingua Posnaniensis*.
- [2] LangLex. (n.d.). Kangri: Mother tongue profile (2011 Census report). Retrieved March 4, 2026, from <https://langlex.com/cens/MTPProfile.php?mtname=Kangri>
- [3] Office of the Registrar General, India. (2022). *Linguistic Survey of India: Himachal Pradesh*. (R. Pattanaik, Ed.). Language Division, Office of the Registrar General, 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.