Monpa Parallel Text Corpus:

Linguistic Features and Structures

MONPA 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:
Hemlata Daimary
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

Monpa Parallel Text Corpus: Linguistic Features and Structures

Authors: Hemlata Daimary, Dr. Rejitha K. S., Dr. Narayan Choudhary, Prof. Shailendra Mohan
Contributors: Nima Dorjee, Pema chotton

e-ISBN: 978-81-69099-13-4

CIIL Publication No.: 1595

First published: AD 2026 March :: *Phalguna 1947 Śaka*

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: Hemlata Daimary

Table of Contents

MONPA PARALLEL TEXT CORPUS

Hemlata Daimary

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 expressionand 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 thereforebringing 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 MONPA : A BRIEF INTRODUCTION

Monpa Language belongs to the Tibeto-Burman branch of the Sino-Tibetan language family. It is an indigenous language spoken by the Monpa people of northeastern India, primarily in the districts of Tawang and West Kameng in Arunachal Pradesh, with smaller communities residing in parts of Assam.As a vital linguistic thread in the cultural fabric of the Eastern Himalayas, it serves as the primary medium of expression for the Monpa people, a community renowned for its deep spiritual heritage and historic ties to Tibetan Buddhism. According to the Census of India (2011), the Monpa-speaking population is total 13,703 numbers, though precise figures vary due to classification practices and multilingualism within the region.

The language is part of the East Bodish branch, representing a unique evolutionary path distinct from its neighbors. Historically, Monpa has flourished through a symbiotic relationship with Classical Tibetan, which has profoundly influenced its philosophical and religious lexicon. Traditionally recorded using the Uchen script—a formal, block-style script of Tibetan origin—the language preserves ancient wisdom and monastic traditions. In contemporary settings, there

is a growing movement to document and standardize its various dialects to ensure their resilience in a rapidly changing world.

It encompasses several dialectal varieties, reflecting the geographical spread and historical development of Monpa communities. The language serves not only as a medium of daily communication but also as a vessel of oral literature, ritual practice, indigenous knowledge systems, and collective memory. Its vocabulary and expressive forms encode ecological knowledge, social organization, and spiritual worldviews that have been transmitted across generations.

Like many indigenous languages of the Himalayan belt, Monpa faces challenges arising from modernization, migration, and the increasing dominance of regional and national languages. The preservation and documentation of such languages are vital, not only for safeguarding cultural identity but also for enriching the broader understanding of human linguistic diversity. Each language embodies unique ways of interpreting the world and contributes to the intellectual and cultural mosaic of humanity.

1.4 MONPA PARALLEL CORPUS

Monpa language exhibits complex morphological and syntactic structures that reflect its historical development within the eastern Himalayan region. The language demonstrates agglutinative tendencies, with the use of suffixation, case markers, and verbal inflections to encode grammatical categories such as tense, aspect, mood, and other functional distinctions.

Monpa generally follows a Subject-Object-Verb (SOV) word order, although variation may occur depending on discourse context and emphasis. Its phonological system reflects features characteristic of Himalayan Tibeto-Burman languages, and dialectal diversity within Monpa highlights the influence of geography and long-standing cultural interactions. Traditionally maintained through oral transmission, Monpa carries a rich heritage of folklore, ritual discourse, and indigenous knowledge systems. In written contexts, it has been represented using adapted scripts, including Tibetan and Roman scripts, particularly in educational and documentation initiatives. This evolving orthographic practice provides important opportunities for systematic linguistic recording and digital archiving.

The structural richness and cultural depth of Monpa make it highly significant for linguistic documentation and comparative research. The development of a Monpa parallel corpus, aligned with other Indian languages, can facilitate the study of syntactic alignment, morphological patterns, lexical correspondences, language contact phenomena, and script adaptation processes. Such a corpus is especially valuable in multilingual settings, where issues of translation equivalence and code-mixing frequently arise.

By supporting corpus development and scholarly engagement, this initiative contributes not only to academic research but also to the preservation and revitalization of Monpa as a living linguistic heritage within India’s diverse cultural landscape..

1.5 SUMMARY OF MONPA PARALLEL CORPUS

The Monpa 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, 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, 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, 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 Monpa which are systematically structured based on 159 grammatical categories. Monpa has 30,510 word count and 164,566 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] Rejitha K. S. and Narayan Kumar Choudhary. (ed.). (2025). LDC-IL Corpus Insights. CIIL, Mysore. 978-93- 48633-33-0.