

Bodo Parallel Text Corpus: Linguistic Features and Structures



BODO 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:

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

Bodo Parallel Text Corpus: Linguistic Features and Structures

Authors: Hemlata Daimary, Dr. Rejitha K. S., Dr. Narayan Choudhary, Prof. Shailendra Mohan

Contributors: Dr. Bridul Basumatary, Mr. Farson Daimari, Dr. Bihung Brahma, Mr. Mahananda Brahma

e-ISBN: 978-81-69099-37-0

CIIL Publication No.: 1671

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: Hemlata Daimary

Table of Contents

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

BODO 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 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 BODO: A BRIEF INTRODUCTION

Bodo (also known as Boro) is a prominent Tibeto-Burman language primarily spoken in the state of Assam, particularly within the Bodoland Territorial Region (BTR), as well as in parts of West Bengal and neighboring states in Northeast India. It holds a distinguished position as one of the Scheduled Languages of India, serving as a vital medium of administration, education, and cultural expression. According to the 2011 Census of India, there are 14,82,929 speakers who identified Bodo as their mother tongue, reflecting its status as a significant linguistic identity in the region.

The language belongs to the Boro-Garo group and traces its roots to the ancient inhabitants of the Brahmaputra Valley. While Bodo has maintained its core structural integrity, it has historically interacted with Assamese and Bengali, leading to a nuanced exchange of lexical items. In earlier periods, various scripts were utilized by the community, including the Roman script and regional variants. However, in 1974, the Devanagari script was officially adopted for writing Bodo, modified with specific diacritics to precisely capture its unique tonal phonemes and aspirate sounds.

Bodo literature is a vibrant tapestry of oral traditions, folk songs, and modern written works. Many prominent writer are among the pioneering figures who laid the foundation for contemporary Bodo letters, transitioning the language from its rich folk roots into the realms of modern poetry, drama, and fiction. The *Bodo Sahitya Sabha*, established in 1952, has played a monumental role in the standardization and promotion of the language. Today, Bodo is a resilient and flourishing tongue with a robust presence in digital media and journalism, standing as a testament to the enduring heritage and intellectual vitality of the Bodo people.

1.4 BODO PARALLEL CORPUS

Bodo is a linguistically sophisticated and structurally distinctive language. It is renowned for tonal complexity and a highly efficient agglutinative system, where various grammatical markers are systematically layered onto a root word. The language exhibits a remarkable range of morphological depth, characterized by precise case endings, expressive verbal suffixes, and a tradition of productive compounding. While Bodo predominantly adheres to a Subject-Object-Verb (SOV) structure, its syntax is inherently dynamic, allowing for rhythmic flexibility that reflects its deep-seated oral and literary heritage.

The modern Bodo language is primarily written in the Devanagari script, which has been thoughtfully adapted to represent its specific phonemic inventory, including its unique aspirated sounds and tonal variations. This transition to a standardized written form has further solidified Bodo's role as a major medium of intellectual and cultural discourse in Northeast India. The language's intricate phonology and structured grammar offer profound insights into the cognitive frameworks of its speakers, highlighting its critical position within the broader tapestry of the Sino-Tibetan linguistic phylum.

Bodo language represents an important resource for the development of parallel text corpora with other Indian languages. A Bodo parallel corpus can serve as a valuable tool for linguistic research, enabling the study of syntactic variation, vocabulary correspondences, translation practices, and patterns of language contact. Such resources are particularly relevant in multilingual settings like India, where interactions among languages often lead to code-mixing and script diversity. The creation of a Bodo parallel corpus therefore contributes not only to academic research but also to the broader goals of language documentation, preservation, and the promotion of linguistic diversity.

1.5 SUMMARY OF BODO PARALLEL CORPUS

The Bodo 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, 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, 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 Bodo which are systematically structured based on 159 grammatical categories. Bodo has 24,560 word count and 166,299 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.