Bengali Parallel Text Corpus: Linguistic Features and Structures

BENGALI 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:
Sonali Sutradhar,
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

Bengali Parallel Text Corpus: Linguistic Features and Structures

Authors: Sonali Sutradhar, Dr. Rejitha K. S., Dr. Narayan Choudhary, Prof. Shailendra Mohan

Contributors: Poulami Das, Arpita Poddar, Monalisa Paul, Agnisha Majumder, Kuntala Ghosh Dastidar

e-ISBN: 978-81-69099-26-4
CIIL Publication No.: 1676

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: Sonali Sutradhar

Table of Contents

BENGALI PARALLEL TEXT CORPUS

Sonali Sutradhar

1.1 LDC-IL PARALLEL TEXT CORPUS

LDC-IL have 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 in spoken form. They don't have script of their own. Following the requirements of the NEP-2020, LDC-IL initiated to collect sentences from different mother tongues that could give an idea about the structure of these 270 mother tongues.

A parallel corpus is a collection of texts in two or more languages that are carefully aligned so each segment matches its translation. 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 categories like Adjective, Adverb, Compounding, Interrogative, Imperative, Speech type, Statement, Structural Question, Coordination, and Subordination etc. Moreover, it includes linguistic features such as Adposition, Anaphora, Derivational Morphology, Inflectional Morphology and language specific features like Emphasis, Equative, Posession, Reciprocal, Reflexive etc. For more details, refer to LDC-IL Corpus Insights.

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 foundation 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 orange economy. Local languages often contain indigenous knowledge about agriculture, medicine,

environment, and culture. Bringing them into digital systems safeguards and shares this knowledge.

Supporting Low-resource languages makes AI more inclusive, intelligent, and representative of linguistic diversity. In India, where hundreds of mother tongues coexist, 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 use in their region. Some mother tongues may use multiple scripts or recently developed writing systems, 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 BENGALI: A BRIEF INTRODUCTION

Bengali (also called Bangla) is the language native to the region of Bengal, which includes the Indian states of West Bengal, Tripura, and southern Assam and Bangladesh. Bengali is one of the most spoken languages in the world, ranking currently as the fifth most spoken native language with approximately 230 million speakers. The language is written using the Bengali-Assamese script. It is one of the Scheduled Languages and is also recognized as a Classical Language of India by the Government of India on 3 October 2024. As per the 2011 Census of India, there are 9,61,77,835 speakers identified Bengali as their mother tongues in the country. It developed as a distinct language with significant influences from Sanskrit. Later, through contact with English, French, Arabic, Persian, Portuguese, it borrowed numerous lexical items from them.

Bengali has developed over more than 1,400 years. Bengali literature, with its millennium-old literary history, was extensively developed during the Bengali Renaissance and is one of the most prolific and diverse literary traditions in Asia. The Bengali language movement from 1948 to 1956 demanding that Bengali be an official language of Pakistan fostered Bengali nationalism in East Bengal leading to the emergence of Bangladesh in 1971. In 1999, UNESCO recognised 21 February as International Mother Language Day in recognition of the language movement.

1.4 BENGALI PARALLEL CORPUS

Bengali is a highly inflectional language with rich morphophonemic features. The standard literary form of Modern Bengali was developed during the 19th and early 20th centuries based on the west-central dialect spoken in Shantipur region of the Nadia district. Modern Bengali shows a high degree of diglossia, with the literary and standard form differing greatly from the colloquial speech of the regions that identify with the language. It has an extensive set of consonant and vowel symbols that precisely capture its phonemic distinctions. It generally follows Subject-Object-Verb pattern but it is linguistically rich language. These features contribute to Bengali's uniqueness among the languages of India and highlight its importance in the study of linguistics.

Bengali 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 and code-mixing phenomena, which are particularly relevant in multilingual contexts such as India.

1.5 SUMMARY OF BENGALI PARALLEL CORPUS

The Bengali 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, 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, 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 Bengali which are systematically structured based on 159 grammatical categories. Bengali has 26950 word count and 146668 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.