BRAJBHASHA PARALLEL TEXT CORPUS: LINGUISTIC FEATURES A STRUCTURES

BRAJBHASHA 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:

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 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

Brajbhasha Parallel Text Corpus: Linguistic Features and Structures

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

Contributors: Shyam Ratan, Pramod Rathor, Omprakash

e-ISBN: 978-81-69099-60-8
CIIL Publication No.: 1670

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: Teena Verma

Table of Contents

BRAJBHASHA 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 BRAJBHASHA: A BRIEF INTRODUCTION

Brajbhasha is an Indo-Aryan language generally classified as a dialect of the Western Hindi group. It occupies an important position among the Western Hindi varieties. Derived from ‘Shauraseni Prakrit’, the form spoken in Mathura, Aligarh, and Agra is regarded as the standard or pure variety of Brajbhasha. Its geographical influence extends southward to regions such as Agra, Aligarh, Bharatpur, Dholpur, Karauli and Mathura. However, the speech in many of these areas represents regional varieties shaped by Brajbhasha rather than its pure form. According to the Census of India 2011, about 15,56,314 people reported Brajbhasha as their mother tongue.

Brajbhasha is renowned for its rich and extensive literary heritage. It appears in classical works associated with Brajbhasha and Pingal traditions, including texts like Prithviraj Raso, Sandesh Rasak, and Kirtilata. It also served as a principal literary medium for the Nath saints and major Bhakti poets such as Surdas, Tulsidas, Nanddas, Abdul Rahim Khan-i-Khana (Rahim), and Raskhan, as well as Ritikal poets like Bihari Lal, Dev, and Ratnakar. Owing to this distinguished literary contribution, it gained recognition under the name “Braj Bhasha.”

1.4 BRAJBHASHA PARALLEL CORPUS

Brajbhasha is a well-formed Indo-Aryan language with a clear grammatical structure, clear gender and number forms, case system and a simple SOV sentence structure. Despite its close relationship to Hindi, it maintains its distinct identity through its independent form, phonetics, and popular usage. Brajbhasha is written in the Devanagari script

From the sixteenth to the eighteenth century, Hindi literature was largely based on Brajbhasha. Dr. Dhirendra Varma considers the actual beginning of Brajbhasha literature to be around 1519 CE. The disciples of Vallabhacharya and Vitthalnath composed devotional works immersed in Krishna-bhakti primarily in Brajbhasha. Surdas and the other Ashtachap poets established Braj as a major literary language.

Brajbhasha is predominantly characterized by “o-ending” (okārānt) forms such as bhlo, baḍo, chhoro, hoygo, and haigo. In addition, some distinctive grammatical features of this language include the use of 'nɛ' in place of 'ne', 'kuː □' instead of "ko' , 'so', instead of 'se' , and 'pɛ' instead of 'pəɾ'.

Some notable grammatical features of Brajbhasha are as follows:

- Frequent use of o-ending or au-ending forms in verbs, nouns, and adjectives.
- Use of case markers such as ko, ko̰, and kɔ in accusative and dative constructions.
- Possessive forms like meɾo (my), t̪eɾo (your) and ɦəməɾo (our) in the genitive case.
- Use of emphatic particles such as ke ɦoɦut̪o and ɦut̪e to express certainty.
- Addition of the suffix -n for plural forms, as in cərnən (feet) and nɛnən (eyes).
- Use of u as a variant or altered form in words such as kuᶜʰ (some) → kəcʰu.

Brajbhasha 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 BRAJBHASHA PARALLEL CORPUS

The development of Brajbhasha Parallel Corpus, Hindi was used as the source language which was translated into Brajbhasha. The Brajbhasha parallel text corpus connected with English and 146 mother tongues of India. Those mother tongues are Anal, Angami, Apatani, Are, Assamese, Bagheli/Baghel Khandi, Bagri Rajasthani, Bagri, Balti, Bengali, Bhadrawahi, Bharmauri/Gaddi, Bhojpuri, Bilaspuri Kahluri, Bodo, 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 Brajbhasha which are systematically structured based on 159 grammatical categories. Brajbhasha has 32,625 word count and 143,975 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, George Abraham (Ed.) (1919) Linguistic Survey of India. Volume-11 19/9 Indo-Aryan Family (Central Group), Reprint Year-2005, 2012. Delhi, India: Low Price Publications

[2] LangLex. (n.d.). Brajbhasha: Mother tongue profile (2011 Census report). Retrieved March 4, 2026, from https://langlex.com/cens/MTProfile.php?mtname=Brajbhasha

[3] Office of the Registrar General & Census Commissioner, India. (2018). C-16: Population by mother tongue (Table). Ministry of Home Affairs, Government of India. https://language.census.gov.in/showMTSIdata

[4] Verma, D. (1954). Hindi bhasha ka itihas [History of the Hindi language]. Hindustani Academy.

[5] Rejitha K. S. and Narayan Kumar Choudhary. (ed.). 2025. LDC-IL Corpus Insights. Central Institute of Indian Languages, Mysore. 978-93-48633-33-0.