Bhojpuri Parallel Text Corpus:
Linguistic Features and Structures

BHOJPURI 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:
Ankita Tiwari
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

Bhojpuri Parallel Text Corpus: Linguistic Features and Structures

Authors: Ankita Tiwari, Dr. Rejitha K. S., Dr. Narayan Choudhary, Prof. Shailendra Mohan

Contributors: Ravi Prakash Suraj, Rajesh Kumar, Anupama Tiwari

e-ISBN: 978-81-69099-95-0
CIIL Publication No.: 1673

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: Ankita Tiwari

TABLE OF CONTENTS

BHOJPURI PARALLEL TEXT CORPUS

Ankita Tiwari

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 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 BHOJPURI: A BRIEF INTRODUCTION

The name Bhojpuri for the language seems to be established by the 17th century and first documented in writing in 1789. The language is variously known under different alternative names such as Bajpuri, Bhojapuri, Bihari, Deswali, Khotla, and Piscimas. Bhojpuri is primarily spoken in western Bihar, the eastern (Poorvanchal) region of Uttar Pradesh, and the north-western areas of Jharkhand. It also has a significant presence outside India, particularly in Nepal and in Diaspora communities in Guyana, Suriname, Fiji, Trinidad and Tobago, and Mauritius. The language derives its name from the town of Bhojpur in Bihar, which later became the name of the district in that region. Bhojpuri is closely related to Maithili and Magahi, and these languages are collectively categorized as the Bihari group. It belongs to the Eastern branch of the Indo-Aryan language family and is further placed within the Bihari subgroup (Grierson, 1903). As an Eastern Indo-Aryan language, it also shares historical and structural affinities with Bengali and Odia. The Surinamese variety of Bhojpuri is commonly known as Sarnami Hindi or Sarnami.

Bhojpuri was written in the Kaithi script from the 16th century until the mid-twentieth century. Since 1894, however, Devanagari has gradually become the dominant script, particularly in administrative and official domains. According to the 2011 Census of India, Bhojpuri has 50,579,447 speakers and is officially regarded as a dialect of Hindi.[1]

[1] https://censusindia.gov.in/nada/index.php/catalog/42561

1.4 BHOJPURI: A BRIEF LINGUISTIC OVERVIEW

Bhojpuri (ISO 639-3 code: bho)[2] follows a Subject–Object–Verb (SOV) word order, with indirect objects generally preceding direct objects and determiners or modifiers placed before nouns. Its lexicon is primarily derived from Sanskrit, while also incorporating loanwords from Hindi, Bengali, and English. Bhojpuri utilizes word-formation processes such as affixation, reduplication, and compounding.

Phonologically, Bhojpuri distinguishes a drawled or lengthened vowel a, which is conventionally represented in writing using the Avagraha symbol (ऽ). For example, the expression *sun'la* (you listen) is written as *सुनऽलऽ*, with the Avagraha marking the elongated vowel in contrast to the short a.

According to Grierson Bhojpuri is relatively simpler than many other Indo-Aryan languages. Nouns occur in three forms: short, long, and redundant, while adjectives remain invariant and do not reflect gender distinctions. Plurals are formed either by adding the suffixes *-na* or *-ni* to nouns or by employing collective terms such as *sabh (all)* or *"lōg (people)*. Numerals often appear with classifiers such as *gō* and *ṭhō*, which emphasize both countability and totality.

Bhojpuri exhibits a three-tiered system of second-person pronouns, similar to many other Indian languages. The highest honorific forms, *rauraa/rauā* and *raaur*, convey the greatest degree of respect, while *tohaar* functions as an medium honorific, and *tu* (with its possessive form *tor*) serves as the non-honorific or intimate pronoun. The highly honorific forms *rauraa/rauā* are typically employed when addressing elders, socially respected individuals, strangers, or in formal contexts. Grammatically, these pronouns trigger the addition of the suffix *-ii* to the verb, as in *rauraa/rauā calii* (you go).

1.5 SUMMARY OF BHOJPURI PARALLEL CORPUS

The Bhojpuri Parallel Text Corpus has been developed using Hindi as the source language. This Corpus connected with English and 146 mother tongues of India. Those mother tongues are

[2] https://iso639-3.sil.org/code/bho

Anal, Angami, Apatani, Are, Assamese, Awadhi, Bagri Rajasthani, Bagri, Balti, Bengali, Bhadrawahi, Bharmauri/Gaddi, Bagheli/Baghel Khandi, 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.

The Bhojpuri dataset comprises 5,332 sentences/phrases, systematically structured according to 159 grammatical categories. The Bhojpuri portion of the corpus contains 31,893 words and a total of 1,41,872 characters. Moreover, the corpus is aligned with 146 Indian mother tongues in addition to English. Altogether, the multilingual parallel corpus encompasses 4,404,845 words (over 4.4 million tokens) and 23,374,289 characters (approximately 23.3 million).

1.6 REFERENCE

[1] Ranjan Kumar. (2024). Linguistic Variation and Change in Bhojpuri. International Journal for Multidisciplinary Research (IJFMR). 2582-2160

[2] https://censusindia.gov.in/nada/index.php/catalog/42561

[3] Tariq Khan and Manish Kumar Singh, Variations in Bhojpuri: A Sociolinguistic Study

[4] Ram Baksh Mishra. (2003). Sociology of Bhojpuri Language.

[5] J. Rajathi & P. Persumal Samy. (2021). A Linguistic Description of Bhojpuri Mother Tongue Spoken in Bihar: https://www.academia.edu/45186266/A_Linguistic_Description_of_Bhojpuri_Mother_Tongue_Spoken_in_Bihar

[6] Shaivya Singh, Rajesh Kumar and Lata Atreya. (March, 2014). Politeness in Language of Bihar: A Case Study of Bhojpuri, Magahi, and Maithili. International Journal of Linguistics and Communication. 2372-479X

[7] Rejitha K. S. and Narayan Kumar Choudhary. (ed.). (2025). LDC-IL Corpus Insights. CIIL, Mysore. 978-93- 48633-33-0.