

Magadhi/Magahi Parallel Text Corpus: Linguistic Features and Structures



MAGADHI/MAGAHI 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:

***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

Magadhi/Magahi Parallel Text Corpus: Linguistic Features and Structures

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

Contributors: Dr. Anamika Kumari, Madhulika

e-ISBN: 978-81-69099-38-7

CIIL Publication No.: 1612

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: Sonali Sutradhar

Table of Contents

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

MAGADHI/MAGAHİ 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 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 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 MAGADHI/MAGAHİ: A BRIEF INTRODUCTION

Magahi (also known as Magadhi) is a vibrant Indo-Aryan language primarily spoken in the eastern part of India, specifically in the Magadh region of Bihar and parts of Jharkhand. It carries a deep historical weight, being the direct descendant of Magadhi Prakrit, the language believed to be used in the ancient Mauryan Empire. It originated from the eastern dialect of Apabhramsha called Magadhi. The language has several dialects and can be written in Devanagari and Kaithi. In Bihar, presently Magahi is spoken in districts of Patna, Nalanda, Gaya, Jehanabad, Aurangabad, Sheikhpura and Nawada. It expands its boundaries to Lakhisaray and Bhagalpur. The language is also spoken in districts of Hazaribag, Giridih, Palamau and Ranchi belonging to the newly formed state of Jharkhand, which was once a part of state of Bihar. The language is also spoken in Mayurbhanj and Bamra in Orissa. Magahi also has some speakers in Malda and Purulia districts of West Bengal. According to the 2011 census 1,27,06,825 people has recognized Magadhi/Magahi as their mother tongue.

Magadhi/Magahi is part of the Bihari group, making it a sibling to Bhojpuri and Maithili. While often labeled as a "dialect" of Hindi, linguistically it is a distinct language with its own unique

grammar and syntax. Like other Eastern Indo-Aryan languages, Magahi often prefers the "unrounded" vowel sounds. It lacks the complex clusters found in Sanskrit, opting for simplified phonetic structures. Magahi has a rich oral tradition, including folk songs, folktales, and regional literature. Folk songs like Phaag and Chaita are sung all over Bihar. While Phaag are sung in the month of Phalgun, Chaita are sung in the month of Chait.

1.4 MAGADHI/MAGAHİ PARALLEL CORPUS

Magadhi/Magahi belongs to the Indo-Aryan language family. It is a highly inflectional language. Magahi phonology shows several characteristics typical of eastern Indo-Aryan languages, such as short and long vowel distinctions and Nasalized vowels. Magahi is famous among linguists for its complex verb morphology. Unlike Hindi, the verb in Magahi doesn't just change based on the subject; it also changes based on the social status of the person being spoken to. Magahi is famous among linguists for its complex verb morphology. Unlike Hindi, the verb in Magahi doesn't just change based on the subject; it also changes based on the social status of the person being spoken to. Unlike Hindi, Magahi does not assign grammatical gender to inanimate objects.

Magadhi/Magahi 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 the development of the language which is particularly relevant in multilingual contexts such as India.

1.5 SUMMARY OF MAGADHI/MAGAHİ PARALLEL CORPUS

The Magadhi/Magahi 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, 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, Kodava, Kokbarak, Kolami, Koli, Kom, Konda, Konkani, Konyak, Koya, Kudubi/Kudumbi, Kuki, Kurmali Thar, Ladakhi, Lepcha, Liangmei, Limbu, Lotha, Lyngngam, 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 Magadhi/Magahi which are systematically structured based on 159 grammatical categories. Magadhi/Magahi has 32616 word count and 149260 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.