



Maithili Parallel Text Corpus: Linguistic Features and Structures

MAITHILI 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:

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

Maithili Parallel Text Corpus: Linguistic Features and Structures

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

Contributors: Sharda Jha and Dharmbir

e-ISBN: 978-81-69099-09-7

CIIL Publication No.:1611

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: Ankita Tiwari

TABLE OF CONTENTS

MAITHILI PARALLEL TEXT CORPUS	5
1.1 LDC-IL PARALLEL TEXT CORPUS.....	5
1.2 CHALLENGES IN CREATING A PARALLEL CORPUS.....	5
1.3 MAITHILI: A BRIEF INTRODUCTION	6
1.4 MAITHILI: A BRIEF LINGUISTIC OVERVIEW	8
1.5 SUMMARY OF MAITHILI PARALLEL CORPUS	9
1.6 REFERENCE.....	10

MAITHILI 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 MAITHILI: A BRIEF INTRODUCTION

Maithili (ISO 639-3: mai)¹ traces its origins to the early medieval period, having evolved from Magadhan Prakrit and later Apabhramsha stages. By the fourteenth century, the devotional and lyrical compositions of the poet Vidyapati contributed significantly to the consolidation of Maithili as an independent literary language. Continuous contact with Sanskrit, Hindi, and other regional languages has further shaped its vocabulary, stylistic conventions, and syntactic structures over time. Recognized as one of the 22 Scheduled Languages of India, Maithili was reported to have 13,583,464 speakers in the 2011 Census². It is primarily spoken in the states of Bihar and Jharkhand and possesses a rich literary tradition supported by a well-developed grammatical framework. Maithili was written in multiple scripts, including Tirhuta (also known as Mithilakshara), Neware, Kaithi, and Devanagari. Among these, Tirhuta is regarded as the traditional and indigenous script of the language. However, in contemporary usage, Devanagari has become the dominant and widely accepted writing system, largely replacing the earlier scripts in formal and informal contexts.³

Maithili also exhibits substantial dialectal diversity shaped by regional and social factors. Notable varieties include Sotipura Maithili of northern Bihar, Khortha-influenced Maithili in areas bordering Jharkhand, Jogiara Maithili spoken by specific communities in Mithila, the relatively standardized Darbhanga variety associated with classical literature, Bajjika—

¹ <https://iso639-3.sil.org/code/mai>

² <https://censusindia.gov.in/nada/index.php/catalog/42561>

³ http://lisindia.ciiil.org/Maithili/Maith_hist.html

sometimes treated as a separate language—and Tirhutia Maithili of the Tirhut region. These dialects display distinct phonetic, lexical, and syntactic features, reflecting the broad and dynamic linguistic landscape of Maithili.⁴

1.4 MAITHILI: A BRIEF LINGUISTIC OVERVIEW

Maithili is geographically situated between several major linguistic groups: it is bordered by Bangla to the east, Magahi to the south, Bhojpuri to the west, and Sub-Himalayan languages to the north. This positioning has contributed to both contact-induced influences and the preservation of distinctive structural features.

Maithili is traditionally classified as an inflectional language due to its historical descent from Old Indo-Aryan (Sanskrit), which was characterized by a highly developed system of inflection. However, through the course of its linguistic evolution, Maithili has incorporated numerous grammatical particles, including postpositions, markers, and conjunctions. Consequently, within a twofold typological classification, it may be described as partly synthetic and partly analytic. In a fourfold typological framework, it exhibits both isolating and agglutinative tendencies. Its verbal system, in particular, demonstrates clear agglutinative features, as verb forms are constructed through the systematic addition of suffixes.

Maithili retains a richly inflected structure, especially in verb conjugation, where tense, aspect, and mood are explicitly marked. Syntactically, the language follows a canonical Subject–Object–Verb (SOV) word order, a characteristic common to Indo-Aryan languages. Case markers and postpositions play a central role in expressing grammatical relations, allowing for relatively flexible word order without ambiguity. Although Maithili shares many typological similarities with Hindi and related Indo-Aryan languages, it also preserves conservative and unique features that distinguish it within the subgroup.⁵

⁴ <https://www.ijital.org/images/issues/current-issues/current-issue-2024/current-issue-v10-issue-3/1.%20Prof%20Jha%20Issue%203.pdf>

⁵ <https://www.ijital.org/images/issues/current-issues/current-issue-2024/current-issue-v10-issue-3/1.%20Prof%20Jha%20Issue%203.pdf>

1.5 SUMMARY OF MAITHILI PARALLEL CORPUS

The Maithili 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 Anal, Angami, Apatani, Are, Assamese, Awadhi, Bagri Rajasthani, Bagri, Bagheli/Baghel Khandi, Balti, Bengali, Bhadrawahi, Bharmauri/Gaddi, Bhojpuri, Bilaspuri Kahluri, Bodo, 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, 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 Maithili dataset comprises 5,332 sentences/phrases, systematically structured according to 159 grammatical categories. The Maithili portion of the corpus contains 30,959 words and a total of 1,41,773 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] S. K. Jha. (2025). A Reflective Overview of Maithili Linguistics. International Journal of Innovations in TESOL and Applied Linguistics (IJITAL). Vol. 10, Issue 3. 2454-6887: <https://www.ijital.org/images/issues/current-issues/current-issue-2024/current-issue-v10-issue-3/1.%20Prof%20Jha%20Issue%203.pdf>
- [2] <https://censusindia.gov.in/nada/index.php/catalog/42561>
- [3] <https://iso639-3.sil.org/code/mai>
- [4] http://lisindia.ciiil.org/Maithili/Maith_hist.html
- [5] Ramawatar Yadav. (1996). A Reference Grammar of Maithili. Mouton de Gruyter
- [6] Rejitha K. S. and Narayan Kumar Choudhary. (ed.). (2025). LDC-IL Corpus Insights. CIIL, Mysore. 978-93- 48633-33-0.