

Mech/Mechhia

Parallel Text Corpus:

Linguistic Features and Structures



MECH/MECHHIA 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:

Hemlata Daimary

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

Mech/Mechhia Parallel Text Corpus: Linguistic Features and Structures

Authors: Hemlata Daimary, Dr. Rejitha K. S., Dr. Narayan Choudhary, Prof. Shailendra Mohan

Contributors: Ananta Champromary, Dr. Bikash Narjinary

e-ISBN: 978-81-69099-39-4

CIIL Publication No.: 1602

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: Hemlata Daimary

Table of Contents

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

MECH/MECHHIA PARALLEL TEXT CORPUS

Hemlata Daimary

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 MECH/MECHHIA: A BRIEF INTRODUCTION

The Mech language is an indigenous language spoken by the Mech community, an ethno-linguistic group inhabiting parts of eastern India. It is generally classified within the Tibeto-Burman branch of the Sino-Tibetan language family and closely related to Bodo language. The earliest scholarly identification of the Mech with the Bodo people appears in the writings of B.H Hodgson. In his work “Essay the First: On the Kocch, Bodo and Dhimal Tribes” (1847), Hodgson clearly stated: “The Mech is a name imposed by strangers. The people call themselves Bodo.”

This statement strongly affirms that “Bodo” is the self-designation of the community, while “Mech” is an external appellation.

The Mech people are traditionally associated with the plains and foothill regions of the eastern Himalayan belt and share close cultural and historical ties with other communities of the Bodo–Garo group. Over generations, the Mech community has preserved a distinct cultural identity expressed through their language, oral traditions, social practices, and relationship with the natural environment.

The language is spoken primarily in northern parts of West Bengal, particularly in the districts of Alipurduar, Jalpaiguri, Darjeeling district (Terai region) and Cooch Behar, as well as in adjoining areas of western Assam. As per the 2011 Census of India, there are 11,546 speakers identified Mech as their mother tongues in the country. The language functions as an important marker of community identity and serves as a medium through which traditional knowledge, folklore, songs, and collective memory are transmitted across generations.

Like many languages of Sino-Tibetan language family, Mech faces increasing challenges due to the influence of dominant regional and national languages, patterns of migration, and ongoing socio-economic change. These factors often contribute to a gradual decline in everyday language use, especially among younger generations. The language has been preserved primarily through oral tradition, making its documentation particularly significant. In such circumstances, efforts toward documentation and preservation become essential. Safeguarding the Mech language is important not only for protecting the cultural identity of the community but also for maintaining the unique systems of knowledge and worldview embedded within the language.

1.4 MECH/MECHHIA PARALLEL CORPUS

The Mech language demonstrates distinctive grammatical features, including the use of suffixes and markers to express grammatical relationships. It is an agglutinative language with rich morphophonemic features. It exhibits complex morphophonemic alternations, extensive case marking, productive compounding, and relatively flexible word order. Mech generally follows a Subject–Object–Verb (SOV) sentence pattern, though variation may occur depending on context and emphasis in natural speech. These structural characteristics reflect the linguistic traditions of the Tibeto-Burman languages of northeastern India.

The language has been preserved primarily through oral transmission within the community. Oral narratives, folk songs, traditional expressions, and everyday communication have played a significant role in maintaining the continuity of the language across generations. In recent years, efforts have been made to document and represent the language in written form, often through adapted scripts such as Devanagari, Roman and Brahmi, particularly for educational and research purposes. Despite limited written resources, the language continues to serve as an important medium for expressing the cultural identity, values, and heritage of the Mech people.

Mech also offers valuable potential for the development of parallel text corpora with other Indian languages. A Mech parallel corpus can contribute to linguistic research by enabling the study of syntactic variation, vocabulary correspondences, translation patterns, and language contact phenomena. Such a corpus is especially relevant in multilingual contexts like India, where interaction between languages often results in code-mixing and script diversity. The development of a Mech parallel corpus therefore represents an important step toward strengthening linguistic documentation, promoting scholarly research, and supporting the preservation of this significant indigenous language.

1.5 SUMMARY OF MECH/MECHHIA PARALLEL CORPUS

The Mech/Mechhia 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, 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, Maithili, Malayalam, Malvi, Manipuri, Mao, Mara, Maram, Marathi, Maring, 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 Mech/Mechhia which are systematically structured based on 159 grammatical categories. Mech/Mechhia has 24,707 word count and 163,754 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.