Lotha Parallel Text Corpus: Linguistic Features and Structures

LOTHA 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. Kamaraj S,
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

Lotha Parallel Text Corpus: Linguistic Features and Structures

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

Contributors: Thungchopeni N Murry, Yanpomo R Humtsoe, Nzanmongi Z Ezung,
Dr. Yantsubeni Ngullie

e-ISBN:
CIIL Publication No.: 1614

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: Dr.Kamaraj S

Table of Contents

LOTHA PARALLEL TEXT CORPUS

Dr.Kamaraj S

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 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 LOTHA: A BRIEF INTRODUCTION

Lotha language is a Tibeto-Burman language predominantly spoken by the Lotha Naga community in the Wokha district of Nagaland. It is one of the recognized Naga languages of the state and forms an important part of the linguistic and cultural identity of the Lotha people. According to the Census of India 2011, Lotha has approximately 1.5–1.7 lakh speakers who identify it as their mother tongue. Lotha belongs to the Sino-Tibetan language family and developed as a distinct speech form within the Naga linguistic group over centuries. Unlike classical literary languages of India, Lotha evolved primarily as an oral language, preserving its history, customs, and traditions through folk songs, folktales, proverbs, and oral narratives. With the advent of Christian missionaries in the 19th and early 20th centuries, the language began to be written systematically using the Roman script. The present writing system of Lotha is based on the Roman alphabet, adapted to represent its specific phonological features.

Lotha literature initially grew through translated religious texts, especially the Bible, hymn books, and catechism materials. Over time, original writings such as poetry, short stories, essays, and educational textbooks emerged. Though the written literary tradition is comparatively recent,

the oral literary heritage of Lotha is rich and culturally significant. Folk narratives and traditional songs continue to play a vital role in preserving community values and indigenous knowledge. The language holds deep cultural importance among the Lotha people. It is actively used in daily communication, local church activities, community gatherings, and cultural festivals. Lotha also finds place in school education within the community and appears in local print and broadcast media in Nagaland. Despite being a regional language with limited global spread, Lotha maintains vibrant contemporary relevance among its speakers. Its unique linguistic structure, oral traditions, and cultural depth make it an important component of the Naga linguistic landscape. Continuous efforts in documentation, education, and digital preservation are contributing to the strengthening and promotion of the Lotha language for future generations.

1.4 LOTHA PARALLEL CORPUS

Lotha language is an agglutinative Tibeto-Burman language with distinct morphological and phonological features. It exhibits suffixation for grammatical functions such as tense, aspect, mood, number, and case. Like many Naga languages, Lotha demonstrates morphophonemic alternations that occur during affixation, particularly in verb forms and plural marking. Case marking is present, though not as extensive as in some Dravidian languages, and word formation processes such as compounding and derivation are productive. Lotha is written in the Roman script, which was introduced through missionary activity and later standardized for educational and literary purposes. The orthography represents its consonant and vowel distinctions through adapted Roman letters, sometimes using digraphs to capture specific phonetic sounds. Tone and vowel length play a functional role in meaning differentiation, though tone marking is not always consistently represented in everyday writing. The basic word order of Lotha generally follows the Subject–Object–Verb (SOV) pattern. Postpositions are commonly used, and modifiers typically precede nouns. Despite its relatively smaller speaker base, Lotha is linguistically rich and structurally significant within the Tibeto-Burman language family. Its verbal morphology, agreement patterns, and syntactic structures make it an important language for comparative and descriptive linguistic studies.

Lotha serves as a valuable resource for developing parallel text corpora with other Indian mother tongues, particularly those from different language families such as Indo-Aryan and Dravidian. A parallel corpus involving Lotha can provide insights into syntactic divergence, morphological typology, translation strategies, and cross-linguistic structural variation. It is especially useful for studying agglutinative morphology, tone-related meaning distinctions, and challenges of representing oral languages in written and digital formats. In a multilingual country like India, the development of a Lotha parallel text corpus would contribute significantly to research in language documentation, computational linguistics, machine translation, and preservation of indigenous linguistic heritage.

This Lotha parallel corpus constitutes a systematically translated dataset derived from an original English corpus, ensuring structured cross-linguistic alignment and semantic correspondence.

1.5 SUMMARY OF LOTHA PARALLEL CORPUS

The Lotha parallel text corpus is aligned with English and 146 mother tongues of India. These 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, 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, 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 Lotha which are systematically structured based on 159 grammatical categories. Lotha has 31952 word count and 171180 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.