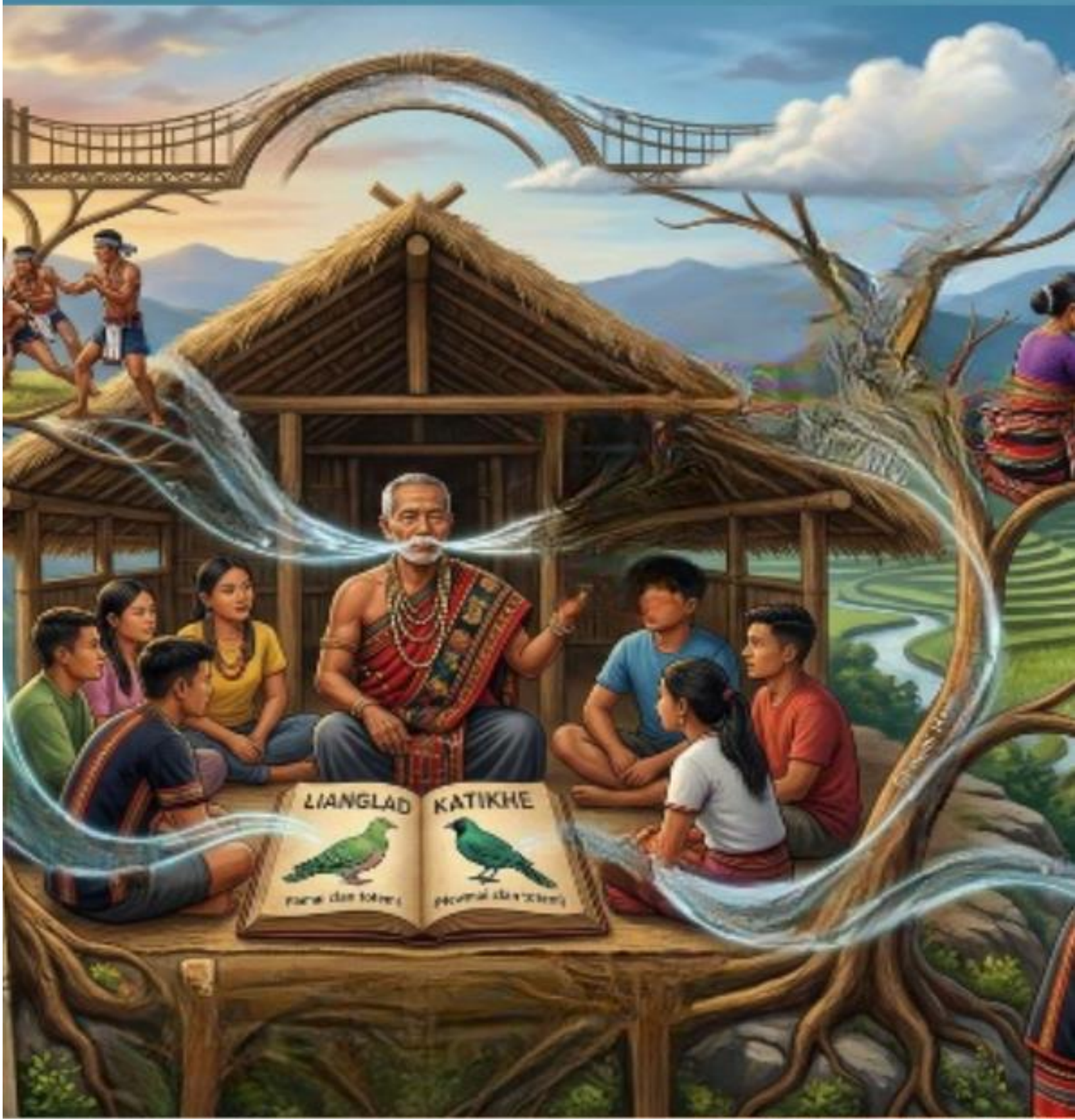
Liangmei Parallel Text Corpus: Linguistic Features and Structures



LIANGMEI 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

Liangmei Parallel Text Corpus: Linguistic Features and Structures

Authors: Dr. Kamaraj S, Dr. Rejitha K. S, Dr. Narayan Choudhary, Prof. Shailendra Mohan
Contributors: Lungphubou Abonmai, Wichamdinbo

e-ISBN:
CIIL Publication No.:1616

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

LIANGMEI **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 LIANGMEI: A BRIEF INTRODUCTION

Liangmei is a Tibeto-Burman language belonging to the Zeliangrong (Zeme–Liangmai–Rongmei) group of Naga languages. It is predominantly spoken in the northeastern Indian states of Manipur, Nagaland, and parts of Assam. Liangmei is recognized as one of the indigenous Naga languages and holds cultural and social importance among the Liangmei community. According to the 2011 Census of India, Liangmei has several tens of thousands of speakers who identify it as their mother tongue. Liangmei developed as part of the Tibeto-Burman linguistic tradition and shares historical and structural affinities with related Zeliangrong languages. Over time, due to close contact with neighboring communities and dominant regional languages, Liangmei has absorbed lexical influences from languages such as Meitei (Manipuri), Assamese, Hindi, and English. These contact-induced borrowings are particularly visible in domains such as education, administration, religion, and modern technology. Traditionally an oral language, Liangmei began to be written using the Roman script with the advent of missionary activities and the spread of Christianity in Northeast India. The Roman script is now widely used for education, religious texts, and literary production. Efforts toward orthographic standardization have contributed to the growth of written materials and literacy in the language.

Liangmei literature includes folk songs, folktales, myths, proverbs, and oral narratives that reflect the rich cultural heritage of the community. In recent decades, written literature such as poetry, short stories, translations, and religious writings has also developed. The preservation and documentation of oral traditions remain central to Liangmei literary identity. Liangmei is a language with deep cultural roots and strong contemporary relevance within its speech community. Its oral heritage, evolving written tradition, and cultural significance make it a unique and valuable language of Northeast India. With increasing educational initiatives, media presence, and community-driven language development efforts, Liangmei continues to flourish among its speakers while contributing to the linguistic diversity of India.

1.4 LIANGMEI PARALLEL CORPUS

Liangmei is a Tibeto-Burman language belonging to the Zeliangrong group of the Naga languages spoken mainly in the northeastern Indian states of Manipur, Nagaland, and parts of Assam. It is linguistically rich and structurally significant within the Tibeto-Burman language family. Liangmei is characterized by its agglutinative morphological structure, where grammatical functions are expressed through the systematic addition of affixes. It displays notable morphophonemic alternations, especially in affixation and verb agreement patterns. The language exhibits extensive case marking and a well-developed system of pronominal and verbal morphology. Compounding is productive, and word formation processes contribute to its lexical expansion. Liangmei generally follows a Subject–Object–Verb (SOV) word order, though pragmatic and discourse factors may allow certain flexibility. Liangmei is written using the Roman script in contemporary usage, though historically it was primarily an oral language. The script adequately represents its phonological system, including its consonant clusters, vowel distinctions, and tonal or suprasegmental features where applicable. Ongoing efforts in standardization and orthographic development have strengthened its written tradition.

Liangmei is a valuable resource for developing parallel text corpora with other Indian mother tongues. A Liangmei parallel text corpus can significantly contribute to linguistic research, particularly in areas such as syntactic divergence, morphological typology, morphophonemic processes, code-mixing phenomena, and script variation. Such a corpus is especially relevant in multilingual contexts like Northeast India, where language contact and cultural interaction are prominent. The development of parallel corpora involving Liangmei would not only support computational linguistics and natural language processing initiatives but also promote the documentation, preservation, and academic study of this important indigenous language.
This Liangmei 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 LIANGMEI PARALLEL CORPUS

The Liangmei 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, 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.

There are 5,332 sentences/phrases in Liangmei which are systematically structured based on 159 grammatical categories. Liangmei has 30096 word count and 168701 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.