Maring Parallel Text Corpus: Linguistic Features and Structures

Maring Parallel Text Corpus

MARING 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:

Amom Nandaraj Meetei

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

Maring Parallel Text Corpus: Linguistic Features and Structures

Authors: Amom Nandaraj Meetei, Dr. Rejitha K. S., Dr. Narayan Choudhary, Prof. Shailendra Mohan

Contributors: Kanshouwa Susie, Dangsawa Moran Maring, K Florida Maring, Ch Abose Moyon, Shimlung Makunga Maring

e-ISBN: 978-81-69099-45-5
CIIL Publication No.: 1603

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: Amom Nandaraj Meetei

Table of Contents

MARING PARALLEL TEXT CORPUS

Amom Nandaraj Meetei

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 in spoken form. They don't have script of their own. Following the requirements of the NEP-2020, LDC-IL initiated to collect sentences from different mother tongues that could give an idea about the structure of these 270 mother tongues.

A parallel corpus is a collection of texts in two or more languages that are carefully aligned so each segment matches its translation. 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 categories like Adjective, Adverb, Compounding, Interrogative, Imperative, Speech type, Statement, Structural Question, Coordination, and Subordination etc. Moreover, it includes linguistic features such as Adposition, Anaphora, Derivational Morphology, Inflectional Morphology and language specific features like Emphasis, Equative, Possession, Reciprocal, Reflexive etc. For more details, refer to LDC-IL Corpus Insights.

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 foundation 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 orange economy. Local languages often contain indigenous knowledge about agriculture, medicine,

environment, and culture. Bringing them into digital systems safeguards and shares this knowledge.

Supporting Low-resource languages makes AI more inclusive, intelligent, and representative of linguistic diversity. In India, where hundreds of mother tongues coexist, 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 use in their region. Some mother tongues may use multiple scripts or recently developed 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 MARING: A BRIEF INTRODUCTION

Maring [ISO 639-3 code nng and Glottolog code mari1416] is a Tibeto-Burman language spoken by the Maring, one of the ancient Naga tribes inhabiting the south-eastern region of the present state of Manipur in India along the border with Myanmar. The Maring people are primarily concentrated in Chandel district of Manipur, though smaller communities are also found in other districts such as Senapati, Ukhrul, and Tamenglong.

According to the 2011 Census of India, the population of the Maring Naga tribe in Manipur was approximately 25,800. While the majority reside in Chandel district, smaller groups are also found in neighbouring areas of Myanmar. Communities with similar physical and cultural features are also reported in other parts of India as well as in Bangladesh and Myanmar. Maring is written in the Latin (Roman) script, and in 2011, literacy rate of Sandangsemba Maring village was 59.57 % compared to 76.94 % of Manipur.

Traditional Maring society is closely associated with ritual practices that accompany major social and cultural events. Ritual rites are performed during festivals and on various customary occasions such as seed sowing, harvesting, house construction and inauguration, childbirth, the post-childbirth purification ceremony known as tuytrumkngei, marriage, and funeral or

condolence ceremonies. These practices form an integral part of the community's traditional and cultural life.

1.4 MARING PARALLEL CORPUS

Maring is an agglutinative language characterized by a high degree of morphological transparency and rich morphophonemic processes. Grammatical relations and functional categories are typically realized through the linear addition of suffixes, often resulting in productive compounding. The language exhibits systematic morphophonemic alternations at morpheme boundaries, reflecting phonological adjustments triggered by affixation. It also exhibits an elaborate case-marking system, productive nominal and verbal compounding, and a relatively flexible word order.

Although the language generally follows a canonical Subject–Object–Verb (SOV) pattern, variations occur for purposes of focus, emphasis, and topicalization. Taken together, these structural properties underscore the linguistic richness and complexity of Maring, contribute to Maring's uniqueness among the languages of India and highlight its importance in the study of linguistics

Maring 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 studying syntactic divergence, morphological richness, code-mixing phenomena, and script variation, which are particularly relevant in multilingual contexts such as India.

1.5 SUMMARY OF MARING PARALLEL CORPUS

The Maring 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, 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, Lotha, Lyngngam, Magadhi/Magahi, Maithili, Malayalam, Malvi, Manipuri, Mao, Mara, Maram, Marathi, 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 Maring which are systematically structured based on 159 grammatical categories. Maring has 25,073 word count and 160,556 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.