



Kokbarak Parallel Text Corpus:

Linguistic Features and Structures



KOKBARAK 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:

*Umesh Chamling Rai,
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

Kokbarak Parallel Text Corpus: Linguistic Features and Structures

Authors: Umesh Chamling Rai, Dr. Rejitha K. S., Dr. Narayan Choudhary, Prof. Shailendra Mohan

Contributors: Jabanika Tripura, Prasanta Tripura, Salu Jamatia

e-ISBN: 978-81-69099-48-6

CIIL Publication No.: 1629

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: Umesh Chamling Rai

Table of Contents

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

KOKBARAK PARALLEL TEXT CORPUS

Umesh Chamling Rai

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 KOKBARAK: A BRIEF INTRODUCTION

Kokbarak/Kokborok is the mother tongue of the Tripuri (Tiprasa) people and one of the officially recognized languages of the Indian state of Tripura. The term 'Kokborok' literally means 'language of the people' (kok = language, borok = people). It plays a central role in the cultural identity, oral traditions, folklore, and contemporary literature of the indigenous communities of Tripura. In recent decades, Kokbarak has also gained visibility in education, broadcasting, and digital communication, which makes it significant for linguistic documentation.

Kokbarak belongs to the Tibeto-Burman branch of the Sino-Tibetan language family. Linguistically, it is closely related to Bodo and Dimasa languages of Northeast India. It is predominantly spoken in Tripura, but also in adjoining regions of Assam and Mizoram, and in some communities in Bangladesh. According to 2011 census estimates, Kokbarak has approximately 900,000 to 1 million speakers, making it one of the major indigenous languages of Northeast India.

1.4 KOKBARAK PARALLEL CORPUS

Kokbarak has several dialectal variations. Those are Debbarma (considered the standard variety), Jamatia, Reang (Bru), Noatia, and Murasing. While these dialects share a common grammatical structure and core vocabulary, they differ in pronunciation, intonation, and certain lexical items. For example, some dialects exhibit vowel length distinctions or tonal variations more prominently than others. Differences are also observed in kinship terminology and agricultural vocabulary.

Kokbarak follows a Subject-Object-Verb (SOV) word order. It is an agglutinative language, meaning that suffixes are added to root words to indicate tense, number, case, and other grammatical functions. There is no grammatical gender. Case markers are suffixed to nouns, and verbs are marked for tense and aspect rather than strict agreement with subjects. Postpositions are used instead of prepositions. Reduplication is a common linguistic feature used for emphasis or plurality.

In the process of interpretation, translation or grammatical analysis, languages such as Hindi, English, Assamese, and many others often contain abstract or technical terms that do not have direct equivalents in Kokbarak. In such cases, strategies such as semantic approximation, descriptive translation, or borrowing of terms were adopted. For example, modern technological or administrative terminology sometimes required explanatory phrases rather than a single-word translation in order to convey the intended meaning accurately.

Another linguistic challenge involves structural differences between the languages. While English follows a Subject-Verb-Object (SVO) word order, Kokbarak typically follows a Subject-Object-Verb (SOV) structure. Therefore, sentence restructuring is necessary during interpretation, translation, or grammatical analysis. In addition, English articles such as “a,” “an,” and “the” do not have direct equivalents in Kokbarak. As a result, researchers must rely on contextual interpretation to convey the intended meaning accurately.

Furthermore, tense and aspect markers require careful attention, as Kokbarak expresses aspect differently from the continuous or perfect forms commonly used in English. The above mentioned features contribute to Kokbarak’s uniqueness among the languages of India and highlight its importance in the study of linguistics.

Kokbarak 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 KOKBARAK PARALLEL CORPUS

The Kokbarak parallel text corpus connected with English and 146 mother tongues of India. Those mother tongues are Anal, Angami, 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, Kashmiri, Khandeshi, Khari Boli, Khasi, Khezha, Khiemnungan, Khortha/Khotta, Kisan, Kodava, 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, Maring, Mech/Mechhia, Mewari, Mewati, Miri/Mising, Mishmi, Mizo, Mongsen, Monpa, Mundari, Muwasi, Nawait, Nimadi, Nyishi, Nocte, Odia, Pahari, Paite, Palmuha, Pania, Pawari/Powari, Phom, Pnar/Synteng, Pochury, Poula, Punjabi, Purkhi, Rai, Rajasthani, Reang, Rengma, Rongmei, Sadan/Sadri, Sambalpur, 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 Kokbarak which are systematically structured based on 159 grammatical categories. Kokbarak has 27187 words count and 175194 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.
- [2] Pushpa Pai (Karapurkar), (1976), Kokborok Grammar, CIIL, Mysore, India.