

Chakru/Chokri

Parallel Text Corpus:

Linguistic Features and Structures



CHAKRU/CHOKRI 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:

***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

Chakru/Chokri Parallel Text Corpus: Linguistic Features and Structures

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

Contributors: Thukuvelu Sakhamo, Huvo Thuluo, Kukhrunelu Theluo

e-ISBN:

CIIL Publication No.:1668

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: Dr.Kamaraj s

Table of Contents

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

CHAKRU/CHOKRI PARALLEL TEXT CORPUS

Dr. Kamaraj S.

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 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 CHAKRU/CHOKRI: A BRIEF INTRODUCTION

Chakri/Chokri language is a Tibeto-Burman language predominantly spoken by the Chakhesang (Chokri) Naga community in the Phek district of Nagaland. It belongs to the Angami–Pochuri group of the Tibeto-Burman language family and represents an important component of the linguistic mosaic of Northeast India. Chokri is recognized as one of the indigenous Naga languages and plays a significant role in the cultural and social life of its speech community. According to the Census of India (2011), Chokri has several thousand native speakers, primarily concentrated in rural and semi-urban areas of Nagaland. The language developed within the broader Tibeto-Burman linguistic tradition and evolved through centuries of oral transmission. Unlike languages that experienced heavy classical influence, Chokri developed largely within its indigenous cultural framework, though in modern times it has borrowed lexical items from English, Nagamese, and other neighboring languages due to contact and multilingual interaction. Traditionally, Chokri was an oral language with a rich heritage of folk songs, folktales, myths, ritual chants, and customary narratives. With the advent of Christianity and formal education in the region during the colonial period, the Roman script was adopted to write Chokri. The present orthography is based on the Roman alphabet, adapted to represent the specific phonological

features of the language. Chokri literature today includes translated religious texts, hymns, folk narratives, educational materials, poetry, and emerging modern writings. While it does not have an ancient written literary text comparable to classical traditions, its oral literature preserves historical memory, customary law, moral values, and collective identity of the Chakhesang people.

Chokri is a language with deep cultural roots and strong contemporary relevance within its community. Its oral traditions, evolving written form, and cultural significance make it a unique and valuable language of India. It continues to be used in homes, churches, community events, and local media, and efforts are ongoing to promote its documentation, preservation, and development for future generations.

1.4 CHAKRU/CHOKRI PARALLEL CORPUS

Chokri language is a Tibeto-Burman language spoken by the Chokri Naga community of Nagaland. It is characterized by rich agglutinative morphology and distinctive phonological features. The language exhibits complex morphophonemic alternations, especially in the formation of tense, aspect, mood, plurality, and case relations. Like many Naga languages, Chokri makes extensive use of suffixation, and grammatical relationships are often marked morphologically rather than through rigid word order. Chokri demonstrates productive compounding and derivational processes that contribute to its lexical expansion. It generally follows a Subject–Object–Verb (SOV) word order, though pragmatic and discourse factors may allow flexibility. The language also shows significant dialectal variation, reflecting the socio-cultural diversity of the Chokri-speaking communities. Chokri is primarily written in the Roman script, introduced through missionary and educational influences, and adapted to represent its specific phonemic distinctions, including tone and vowel length where necessary. Although traditionally oral in nature, recent efforts in documentation and literacy development have contributed to the growth of written materials in Chokri.

Chokri is a valuable resource for developing parallel text corpora with other Indian mother tongues. A parallel corpus involving Chokri can serve as an important research tool for studying syntactic divergence, agglutinative structures, morphophonemic processes, tone–meaning relationships, code-mixing phenomena, and script variation. Such research is particularly significant in multilingual contexts like India, where lesser-documented tribal languages play a crucial role in understanding linguistic diversity and structural complexity.

This Chakru/Chokri 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 CHAKRU/CHOKRI PARALLEL CORPUS

The Chakru/Chokri 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, 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, 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 chakru/chokri which are systematically structured based on 159 grammatical categories. chakru/chokri has 28640 word count and 130608 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).

1.3 REFERENCE

- [1] Rejitha K. S. and Narayan Kumar Choudhary. (ed.). (2025). LDC-IL Corpus Insights. CIIL, Mysore. 978-93- 48633-33-0.