Kudubi/Kudumbi Parallel Text Corpus: Linguistic Features and Structures

KUDUBI/KUDUMBI 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:

Mr. Saurabh Varik

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

Kudubi/Kudumbi Parallel Text Corpus: Linguistic Features and Structures

Authors: Mr. Saurabh Varik, Dr. Rejitha K. S., Dr. Narayan Choudhary, Prof. Shailendra Mohan
Contributors: Santhosh Kumar, Shwetha K.

e-ISBN: 978-81-69099-10-3
CIIL Publication No.: 1621

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: Saurabh Varik

Table of Contents

KUDUBI/KUDUMBI PARALLEL TEXT CORPUS

Saurabh Varik

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 KUDUBI/KUDUMBI: A BRIEF INTRODUCTION

The Kudumbi (also called Kudubi, Kurumbi, or Kunbi) are a traditional farming community believed to have originally migrated from North Goa. They are historically associated with the Konkani-speaking Kunbi–Kurmi agricultural community, but the Kudubi language belongs to the Dravidian language family. Over time, several groups migrated from Goa and settled along the western coast of India. Some communities settled in the coastal districts of Uttara Kannada, Dakshina Kannada, and Udupi in Karnataka, while others moved further south to Kerala. One of the earliest groups reached Cherai Island near Paravur in Ernakulam district, and later spread to areas such as Kochi, Vypeen, Kodungallur, Thrissur, Kozhikode, Kannur, Alappuzha, Kottayam, Kollam, and Thiruvananthapuram. The largest settlement of Kudumbis is in Vypeen near Kochi. Traditionally, they are farmers and agricultural labourers, known for their expertise in paddy cultivation in the low-lying fields of the Kerala backwaters, and for introducing the "Chettiverippu" variety of paddy rice from the Konkan region (Thurston, 1909; Census of India, 1961).

Today, Kudubi/Kudumbi speakers are mainly found in Karnataka, Kerala, and Odisha, with a total population of about 66,918 in India, most of whom live in Karnataka. A smaller number have migrated to cities such as Bangalore, Mangalore, Mumbai, and Delhi in search of better livelihood opportunities. In Karnataka, the Kudubi community is not recognized as a Scheduled Tribe, and community members have demanded such recognition due to social and economic challenges (Ravindranath Rao). The Kudubi language is recognized in Karnataka, but it does not have its own script, and speakers usually write it using the Kannada script. Rituals are an important part of Kudumbi social life, as they believe that performing traditional rituals helps maintain physical, mental, and spiritual well-being by protecting individuals and the community from harmful spirits. Linguistically, the phonological system of Kudubi includes a set of distinctive sound units (phonemes), and the description of its phonemic inventory is based on field data collected from speakers in the Udupi district of Karnataka.

(**Source**: E. Thurston (1909) – *Castes and Tribes of Southern India*, Volume 4., Census of India **(**1961**)** – Volume VII: Kerala, p. 210., Dr. Y. Ravindranath Rao – Caste Practices and Influences Affecting Tribals: A Case Study of Kudubis of Karnataka., Field data collected from Kudubi speakers in Udupi district, Karnataka.)

Creating a parallel corpus for the Kudubi/Kudumbi language in India is relatively little difficult because the language is highly under-resourced and largely oral in nature. Kudubi is mainly spoken by the Kudumbi community in parts of Goa, coastal Karnataka, and Maharashtra, and it does not have a well-standardized written form. Speakers often use spoken varieties that vary from region to region, and when writing is attempted, it is usually influenced by Konkani, Marathi, or Kannada. For example, a native Kudubi sentence may include core Kudubi vocabulary but borrow function words or verbs from Konkani or Marathi due to lack of confidence in written Kudubi. This creates inconsistency when aligning source and target texts in a parallel corpus.

Another difficulty is the lack of reference materials such as grammars, dictionaries, or standardized orthography. While creating parallel sentences, it is often hard to verify whether a word is genuinely Kudubi or borrowed. For instance, a concept like “government office” or “development scheme” may not have an established Kudubi equivalent, forcing the annotator to use Marathi or Konkani terms, which reduces linguistic purity. Additionally, limited digital literacy among native speakers, scarcity of trained annotators, and technical issues such as typing support and internet connectivity further slow down corpus creation. As a result, although parallel corpus development for Kudubi is possible, it requires careful validation, community involvement, and flexible linguistic guidelines to accommodate natural code-mixing while preserving the core structure of the language.

The Kudubi/Kudumbi parallel corpus has been created using Kannada as the source language, and the Kudubi/Kudumbi text is written in Kannada script, reflecting the commonly used writing practice among native speakers.

1.4 SUMMARY OF KUDUBI/KUDUMBI PARALLEL CORPUS

The Kudubi/Kudumbi parallel text corpus includes English and 146 other 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, 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 Kudubi/Kudumbi which are systematically structured based on 159 grammatical categories. Kudubi/Kudumbi has 24,371 word count and 148,209 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).

1.5 REFERENCE

[1] Rejitha K. S. and Narayan Kumar Choudhary. (ed.). (2025). LDC-IL Corpus Insights. CIIL, Mysore. 978-93- 48633-33-0.
[2] https://language.census.gov.in/showMTSIdata#
[4] https://langlex.com/cens/MTProfile.php?mtname=Kachchhi
[5] https://language.census.gov.in/showMTSIdata#
[6] E. Thurston (1909) – Castes and Tribes of Southern India, Volume 4.
[7] Census of India (1961) – Volume VII: Kerala, p. 210., Dr. Y. Ravindranath Rao – Caste Practices and Influences Affecting Tribals: A Case Study of Kudubis of Karnataka., Field data collected from Kudubi speakers in Udupi district, Karnataka.)