Coorgi/Kodagu Parallel Text Corpus: Linguistic Features and Structures

COORGI/KODAGU 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. Saritha S.L

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

Coorgi/Kodagu Parallel Text Corpus: Linguistic Features and Structures

Authors: Dr. Saritha S. L., Dr. Rejitha K. S., Dr. Narayan Choudhary, Prof. Shailendra Mohan
Contributors: Thashma K. P., Lavina G. S., Dr. Bhamini Raghavaiah K.

e-ISBN: 978-81-69099-88-2
CIIL Publication No. :1661

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: Saritha.S.L

Table of Contents

COORGI/KODAGU PARALLEL TEXT CORPUS

Saritha. S. L.

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 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 COORGI/KODAGU: A BRIEF INTRODUCTION

Coorgi/Kodagu is a Dravidian language spoken by about 114,000 people in Karnataka state in southern India, in particular in the district of Kodagu, and also in the Bengaluru, Dakshina Kannada and Mysuru districts. In 2011 there were about 114,000 speakers of Coorgi Kodava, which is also known as Kodava, Coorgi Kodava, Kadagi, Khurgi, Kodagu, Kotagu, Kurja or Kurug. The people of Coorg are known as Kodavas or Coorgies. According to the 2011 Census

of India, the status of the Coorgi/Kodagu (Kodava) language is characterized by its classification as a "Mother Tongue" and its specific demographic concentration in Karnataka.

The 2011 Census recorded the following figures for speakers who identified their mother tongue as Coorgi Kodagu: Total Speakers in India: 113,857. Over 95% of these speakers reside in the state of Karnataka, specifically in the Kodagu district. When comparing the 2011 data to the 2001 Census (which recorded 166,187 speakers), there appeared to be a significant statistical decline. However, experts suggest this might be due to migration Speakers moving to cities and identifying with dominant languages.

1.4 COORGI/KODAGU PARALLEL CORPUS

In the field of Natural Language Processing (NLP), a Parallel Corpus is a collection of texts in two or more languages where each sentence in one language is paired with its translation in another. For the Coorgi/Kodagu language, a parallel corpus is more than a technical database; it is an act of cultural resistance against the homogenizing force of the internet. It provides the "raw material" needed for Coorgi to live in smartphones, classrooms, and global archives. By building these digital bridges today, we ensure that the unique voice of the Coorgi people continues to resonate in the interconnected world of tomorrow.

The 2011 Census status highlights why a parallel corpus is so important. Since Coorgi/Kodagu is grouped under Kannada in the census, it often loses out on independent government funding for digital development. Creating a corpus helps prove its unique linguistic identity and provides the data needed for its preservation.

Creating a parallel corpus for Coorgi/Kodagu is challenging due to its status as a primarily oral, low-resource language with limited written text, lack of standardized script, and significant lexical variation influenced by surrounding Kannada and English. Some of the terminologies in the regional language are anglicised and it becomes tough to judge whether the original word should be used or alternate words should be used. The scarcity of bilingual, high-quality parallel data complicates natural language processing (NLP) applications.

For the Coorgi language, a parallel corpus is more than a technical database; it is an act of cultural resistance against the homogenizing force of the internet. It provides the "raw material" needed for Kodava to live in smartphones, classrooms, and global archives. By building these digital bridges today, we ensure that the unique voice of the Coorgi people continues to resonate in the interconnected world of tomorrow.

1.5 SUMMARY OF PARALLEL CORPUS

The Coorgi/Kodagu 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, 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, Maring, Mech/Mechhia, Mewari, Mewati, Miri/Mising, Mishmi, Mizo, Mongsen, Monpa, Mundari, Muwasi, Nawait, Nepali, Nimadi, Nissi, Nocte, Odia, Pahari, Paite, Palmuha, 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 Coorgi/Kodagu which are systematically structured based on 159 grammatical categories. Coorgi/Kodagu has 23,604 word count and 156040 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). LDC-IL Coorgi/Kodagu parallel corpus is not just a collection of words; it is a foundational infrastructure that enables Coorgi/Kodagu to succeed in the digital age, ensuring its linguistic shades are preserved, understood, and accessible for future generations of researchers and speakers alike.

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. Government of Karnataka (2014), Human Development Report Kodagu-2014.

3. Mallikarjun B. 2019. Linguistic Ecology of Karnataka. Language in India. Vol, 19:7.

4. Census of India 1872 to 2011-Coorg and Karnataka.

5. Conner.1870. Memoir of Codugu Survey: Commonly Written Koorg 2 Parts. Bangalore Central Jail Press.