Konkani Parallel Text Corpus

Linguistic Features and Structures

KONKANI 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

Konkani Parallel Text Corpus: Linguistic Features and Structures

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

Contributors: Ashalata Navelkar, Yashwant Gawas, Siddhi Gawas, Pooja Tople, Shweta Parab, Sunita Gurudas Kanekar

e-ISBN: 978-81-69099-83-7
CIIL Publication No.: 1624

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

KONKANI 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 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 KONKANI: A BRIEF INTRODUCTION

Konkani is an Indo-Aryan language spoken mainly along the western coastal belt of India. It is the official language of Goa and is also spoken in parts of Karnataka, Maharashtra, and Kerala. Due to migration, Konkani-speaking communities are also found in other parts of India and abroad.

Konkani has a strong oral tradition. Folk songs, proverbs, stories, and everyday conversations play an important role in preserving the language. Written literature developed later but has grown steadily in modern times.

Konkani belongs to the Indo-Aryan branch of the Indo-European language family. It shares several grammatical and lexical features with neighboring languages such as Marathi and Kannada, while maintaining its own distinct identity. Early descriptions of the Konkani language show that it is spoken along the western coastal strip of India between Malvan and Karwar, bounded by the Western Ghats on the east and the Arabian Sea on the west, with speakers also

settled in places such as Mumbai, Mangaluru, and Kochi. Historical works indicate that the language had several names in earlier centuries; Portuguese missionaries in the 16th–17th centuries often referred to it as lingua canarim or lingua bramana, while some texts used forms such as lingua concana. Evidence of the language's early literary presence appears in the 1678 Amsterdam edition of Hortus Malabaricus by Hendrik van Rheede, where plant names and a testimonial written in Devanagari by Konkani Brahmins demonstrate the continued use of the language outside its core region. These works collectively show that Konkani had a wide geographical distribution and a complex naming history in early linguistic and missionary literature.

1.4 KONKANI PARALLEL CORPUS

Creating a Parallel Corpus for the Konkani language in India is relatively easy in the electronic or digital domain due to the strong multilingual ecosystem and widespread digitization of language resources. Konkani is used alongside languages such as English, Hindi, and Marathi in government communication, education, and media, resulting in a large volume of already translated and aligned content. Government notifications, textbooks, public information materials, and news content are often available in multiple languages in digital form, which provides ready-made parallel data that can be efficiently extracted and aligned using existing tools.

Additionally, the availability of Unicode-compliant scripts, growing digital publishing practices, and active academic and community participation further simplify corpus creation. Digital resources reduce manual effort, ensure consistency, and allow easy scalability and reuse. With existing language technology tools and institutional support in India, building a high-quality Konkani Parallel Corpus electronically becomes cost-effective, accurate, and sustainable for long-term language technology and research needs.

While creating the Konkani Parallel Text Corpus for Konkani at LDC-IL, it is sometimes observed that contributors use spoken or colloquial forms of Konkani in written data, which is completely acceptable and reflects natural language use. In some cases, words from other languages are also used due to limited vocabulary awareness or common contact influence. Such variations are a normal part of language usage and provide valuable insights into real-world Konkani language practices.

The Konkani parallel corpus has been created using English as the source language, and the Konkani text is written in Devanagri script, reflecting the commonly used writing practice among native speakers.

1.5 SUMMARY OF KONKANI PARALLEL CORPUS

The Konkani 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, 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 Konkani which are systematically structured based on 159 grammatical categories. Konkani has 25,384 word count and 144,901 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] https://language.census.gov.in/showMTSIdata#
[3] https://langlex.com/cens/MTProfile.php?mtname=Konkani
[4] George Cardona and Dhanesh Jain (eds.). 2003. The Indo-Aryan Languages. London: Routledge.
[5] Census of India. 2011. Language Census of India – Primary Census Abstract for Languages. Office of the Registrar General & Census Commissioner, Government of India.
[6] Madhav M. Deshpande. 1993. Sociolinguistic Attitudes in India: An Historical Reconstruction. Ann Arbor: Karoma Publishers.
[7] Government of Goa, Official Language Cell. 1987. The Goa, Daman and Diu Official Language Act (Konkani as Official Language). Panaji: Government of Goa.
[8] Sahitya Akademi. 2009. Encyclopaedia of Indian Literature: Konkani Language and Literature. New Delhi: Sahitya Akademi.
[9] Lewis, M. Paul, Gary F. Simons & Charles D. Fennig (eds.). 2024. Ethnologue: Languages of the World – Konkani. SIL International.
[10] Grierson, George A. 1903–1928. Linguistic Survey of India, Vol. VII: Indo-Aryan Languages (Southern Group – Konkani). Calcutta: Office of the Superintendent of Government Printing.
[11] Konkani language and literature, J. Gerson-Da-Cunha