Pania Parallel Text Corpus: Linguistic Features and Structures

PANIA 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

Pania Parallel Text Corpus: Linguistic Features and Structures

Authors: Dr. Saritha S. L., Dr. Rejitha K. S., Dr. Narayan Choudhary, Prof. Shailendra Mohan
Contributors: Nandakumar S., Anjali Bhaskaran

e-ISBN: 978-81-69175-19-7
CIIL Publication No.: 1751

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

PANIA 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 PANIA: A BRIEF INTRODUCTION

Pania Community, the largest community among the Scheduled Tribes of Kerala is mainly distributed in Wayanad district. They also live in Kannur, Kozhikode and Malappuram Districts, with significant populations in the Nilgiris (Tamil Nadu) and Coorg (Karnataka). According to the 2011 Census, the population of the community is 88,450. (74.49%) among them 42,775 are

males and 45,675 are females. The sex ratio is 1068 females per 1000 males. Their literacy rate is 63.2 percent and most of the Panias are settled in Wayanad district.

The term ‘Pania’ may have been derived from the word ‘Pani’ meaning ‘Work’ and the term means ‘Workers’ as opposed to the landlords. According to their legends, the Pania came from ‘Ippimala’, a mountain near Banasura peak. They speak a language of Malayalam mixed with Tamil words which are called ‘Bhasha’. Pania’s used to live in small thatched huts with or without walls popularly known as ‘Pire’, erected in the land of the landlord. Traditionally Pania’s are bonded laborers. The Bonded Labor System has existed among tribal communities of Wayanad, especially Pania’s and Adiyans. Pania’s who were treated as slaves of their respective Land Lords and also victims of bonded labor system. They have been freed by the enactment of the abolition of bonded labor system since 1976; they still subsist on agriculture labor or any other manual labor.

Pania is mainly an oral language. Pania does not have a formal script of its own. It is generally written using the script of the region in which the speaker lives: When written, it adopts the script of the surrounding dominant language, Kerala: Malayalam script, Karnataka: Kannada script, Tamil Nadu: Tamil script. Pania is considered a "Developing" or stable indigenous language, typically used for communication at home and within the community. While older generations are primary speakers, younger generations are increasingly using Malayalam for education and interaction with outsiders.

Building a Pania parallel corpus involves hurdles that go far beyond simple translation. Because the language is oral, tribal, and "low-resource," Creators face a unique set of technical, linguistic, and ethical obstacles. Both Malayalam and Pania are agglutinative, meaning they build long words by stacking multiple suffixes onto a root.

Pania community lives in the border triangle of Kerala, Karnataka, and Tamil Nadu, their version of Malayalam is "contaminated" (in a linguistic sense) by all three. Some words that exist in Malayalam are replaced by Kannada words in Pania (e.g., using ‘Neeru’ for water instead of ‘vellam’).

1.4 PANIA PARALLEL CORPUS

A Pania parallel corpus is a highly specialized resource used primarily for language preservation, educational bridging, and social inclusion. Language spoken by the Pania tribe in Kerala and parts of Tamil Nadu/Karnataka, these corpora act as a vital link between the tribal community and the mainstream. Creating a parallel corpus for Pania language is a specialized task because Pania is an oral language.

"The Pania Parallel Corpus is a bilingual linguistic resource designed to bridge the gap between a high-resource global language and an endangered, low-resource tribal dialect. Given that Pania is a primarily oral language with no indigenous script, the corpus utilizes a pivot-based translation model. Pania 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.

The LDC-IL (Linguistic Data Consortium for Indian Languages) has developed the Pania corpus as part of its broader mission to document and create digital resources for Indian mother tongues.

1.5 SUMMARY OF PANIA PARALLEL CORPUS

The Pania Parallel Corpus is a structured linguistic dataset designed to align sentences from a source language with their direct translations in Pania. As Pania is a "low-resource" and traditionally unwritten language, this corpus serves as a vital bridge for digital inclusion and language preservation. Without such a resource, the Pania community the largest scheduled tribe in Kerala remains linguistically isolated from the benefits of digital growth. Ultimately, the corpus ensures that when a Pania speaker interacts with technology, AI systems can understand not just their words, but their specific cultural and linguistic identity.

The LDC-IL Pania Parallel Corpus represents a significant milestone in Indian computational linguistics. Because Pania is an endangered, oral-only language, creating a structured text corpus requires a unique methodology that balances linguistic accuracy with digital preservation.

The Pania 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, 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, 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 Pania which are systematically structured based on 159 grammatical categories. Pania has 21731 word count and 163599 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. 1. E. Thurston and Ranachari A.K.,(1909)., caste and tribes of southern India: vol;4-KM., government press, Madras
3. Census of India., (2011)., Kerala, data highlight; the scheduled tribes., http://www.censusindia.gov.in/tables-published/scst/dh-st-kerala.pdf.
4. Census of India. (2011), basic datasheet, district Wayanad, Kerala., http://www.censusindia.gov.in/dis-file/datasheet-3203.pdf. “Kirtads kaypusthakam”, (2017-18).
5. K N K Sharma .,(1992)., A note on the paniyan of Kerala, man in India., (vol:72(.,(pp:3) ., Quarterly Journal of Anthropology, Ranchi.
6. Nambodiri. D,(2006)., Wayanad indicative a situational study and feasibility report for the comprehensive development of Adivasi communities of Wayanad., Kozhikode: IIM and KIRTADS.