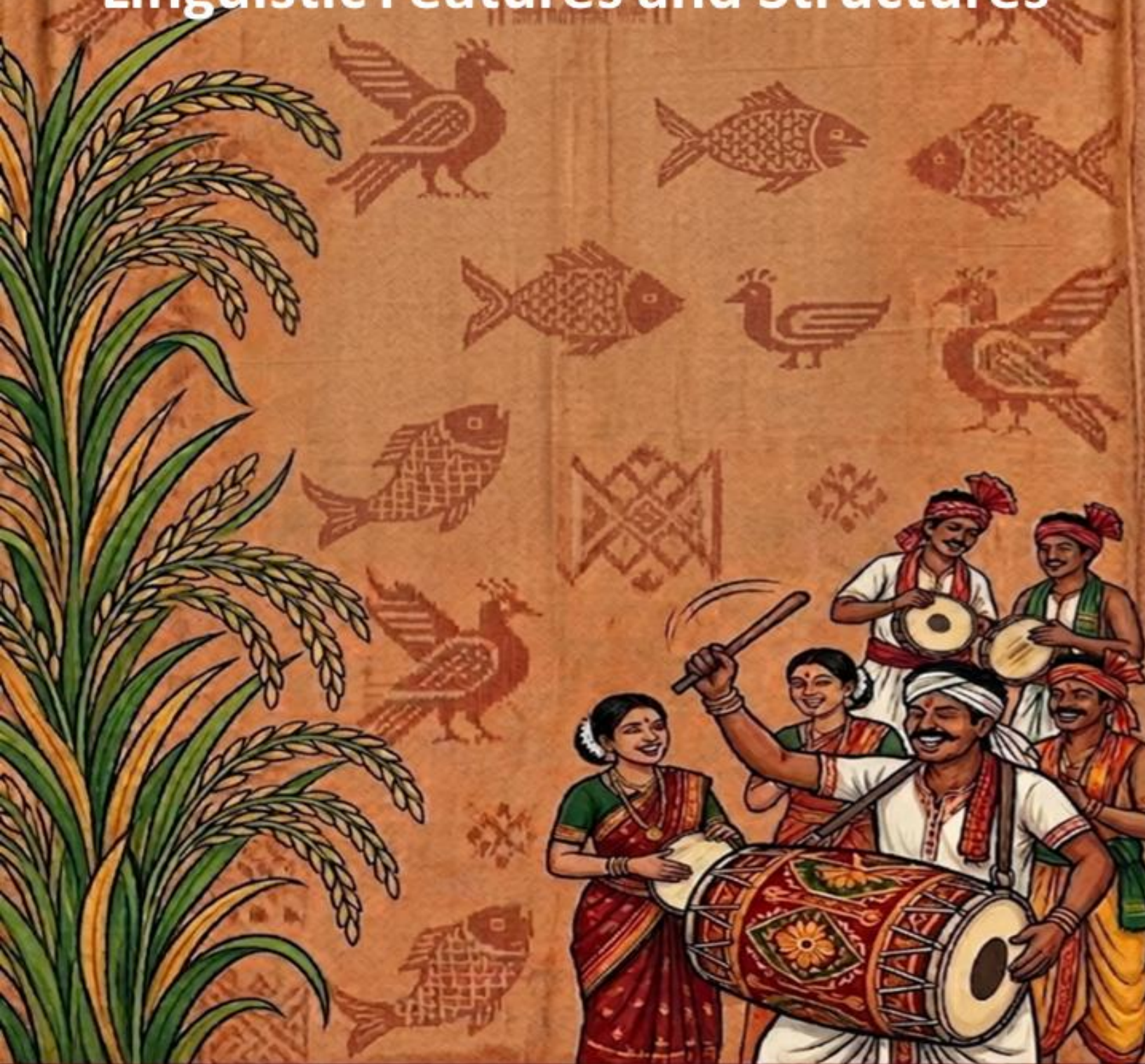


Sambalpuri Parallel Text Corpus: Linguistic Features and Structures



SAMBALPURI 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:

***Poulami Das,
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

Malayalam Parallel Text Corpus: Linguistic Features and Structures

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

Contributors: Suman Rani Panda, Astha Ayusha Mall

e-ISBN: 978-81-69175-58-6

CIIL Publication No.: 1737

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: Sonali Sutradhar

Table of Contents

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

SAMBALPURI PARALLEL TEXT CORPUS

Sonali Sutradhar

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 in spoken form. They don't have script of their own. Following the requirements of the NEP-2020, LDC-IL initiated to collect sentences from different mother tongues that could give an idea about the structure of these 270 mother tongues.

A parallel corpus is a collection of texts in two or more languages that are carefully aligned so each segment matches its translation. 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 categories like Adjective, Adverb, Compounding, Interrogative, Imperative, Speech type, Statement, Structural Question, Coordination, and Subordination etc. Moreover, it includes linguistic features such as Adposition, Anaphora, Derivational Morphology, Inflectional Morphology and language specific features like Emphasis, Equative, Possession, Reciprocal, Reflexive etc. For more details, refer to [LDC-IL Corpus Insights](#).

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 foundation 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 orange economy. Local languages often contain indigenous knowledge about agriculture, medicine,

environment, and culture. Bringing them into digital systems safeguards and shares this knowledge.

Supporting Low-resource languages makes AI more inclusive, intelligent, and representative of linguistic diversity. In India, where hundreds of mother tongues coexist, 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 use in their region. Some mother tongues may use multiple scripts or recently developed 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 SAMBALPURI: A BRIEF INTRODUCTION

Sambalpuri: Sambalpuri (also known as Kosli) is a vibrant Indo-Aryan language variety spoken primarily in Western Odisha. While it is often classified as a dialect of Odia by linguists and the government, most of its 2.6 million native speakers consider it a distinct language with its own unique literary and cultural identity. As per the 2011 Census of India, there are 26,30,381 speakers identified Sambalpuri as their mother tongues in the country. Sambalpuri is a vibrant, soulful language that serves as the cultural heartbeat of Western Odisha, known for its rhythmic cadence and deep connection to folk traditions. Sambalpuri is the primary medium for some of India's most iconic folk expressions. Satya Narayan Bohidar is often hailed as the pioneer of Sambalpuri literature, having standardized its grammar and sparked a literary renaissance in the early 20th century. While it is linguistically categorized as an Indo-Aryan language, its soul is deeply influenced by the tribal demographics of the region, creating a beautiful "bridge" between different cultural identities in Eastern India. Traditionally, Sambalpuri is written using the Odia script. The vast majority of Sambalpuri literature from the 19th century poems in Sambalpur Hiteisini to the modern works of Padma Shri Haldhar Nag is written using the Odia script.

Sambalpuri literature has a rich tapestry that blends ancient oral traditions with a powerful modern movement. For decades, it was preserved through folk songs and stories, but in the last century, it has evolved into a sophisticated literary field that commands national respect. Satya

Narayan Bohidar, known as the "pioneer" of Sambalpuri literature. He was the first to systematically work on Sambalpuri grammar and a dictionary, providing the linguistic foundation for future writers. His birthday, August 1st, is celebrated as Sambalpuri Din (Sambalpuri Day). Hema Chandra Acharya often called the "Kosli Balmiki", who translated the Ramayana into Sambalpuri, titled "Ram Raha." This work was monumental in proving that the language could handle epic, classical themes. Dr. Nilmadhab Panigrahi, a giant of the language who wrote the "Kosli Mahabharat," ensuring that both major Indian epics were accessible in the mother tongue of Western Odisha.

1.4 SAMBALPURI PARALLEL CORPUS

Sambalpuri is a member of the Indo-Aryan language family. While it shares a common ancestor with Standard Odia, it has evolved distinct phonological, morphological, and syntactical features that make it a unique linguistic entity. Sambalpuri tends to drop the "h" sound in aspirated consonants. It features a very specific vowel harmony where the vowel in one syllable influences the vowel in the next. One of the most fascinating features of Sambalpuri is its Austroasiatic (Munda) and Dravidian influence. Because Western Odisha is a tribal heartland, the language has absorbed unique echo words, specific classifiers for counting objects and a rich vocabulary for forest produce, agriculture, and local geography that doesn't exist in Sanskritic Odia.

Sambalpuri 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.

1.5 SUMMARY OF SAMBALPURI PARALLEL CORPUS

The Sambalpuri 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, 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, 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 Sambalpuri which are systematically structured based on 159 grammatical categories. Sambalpuri has 27783 word count and 138676 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.