KHORTHA/KHOTTA PARALLEL TEXT CORPUS: LINGUISTIC FEATURES AND STRUCTURES

KHORTHA/KHOTTA 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. Satyaendra Kumar Awasthi
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

Khortha/Khotta Text Corpus: Linguistic Features and Structures

Authors: Dr. Satyaendra Kumar Awasthi, Dr. Rejitha K. S., Dr. Narayan Choudhary, Prof. Shailendra Mohan

Contributors: Siddharth Sanket, Dr. Shilpa, Dr. Ritu Ghansi, Basant Kumar

e-ISBN: 978-81-69099-22-6
CIIL Publication No.: 1632

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: Teena Verma

Table of Contents

KHORTHA/KHOTTA PARALLEL TEXT CORPUS

Dr. Satyaendra Kumar Awasthi

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 KHORTHA/KHOTTA: A BRIEF INTRODUCTION

Khortha is an Indo-Aryan language and one of the most important and widely spoken regional languages of Jharkhand. It is not merely a language; rather, it serves as a significant medium of cultural identity and expression for millions of people. Khortha is primarily used as a link language in the districts of North Chotanagpur, Santhal Pargana, and Palamu divisions of Jharkhand. According to Khortha scholars, it covers an area of approximately 34,000 square miles. Its main districts include Bokaro, Chatra, Deoghar, Dhanbad, Dumka, Giridih, Godda, Hazaribagh, Koderma, Pakur, and Ramgarh.. As per the 2011 Census of India, 8,038,735 individuals declared Khortha/Khotta as their mother tongue, which is grouped under the larger language category of Hindi.

Khortha folk culture is vibrant and rooted in the soil. Its most prominent characteristic is that it revolves around nature and labor. Khortha folk songs are not merely a means of entertainment, but are also connected to every ritual and agricultural cycle of life. They can be broadly divided into the following categories: Sanskar Geet, Seasonal and Festival Geet, Karam Geet, Sohrai (Chanchar) Geet and Tusu Geet. Sanskar Geet are songs from birth to death. Khortha wedding

songs are dominated by humor and satire. Karam is a major festival of Jharkhand. Sisters sing for the well-being and prosperity of their brothers. Its songs reflect a deep sense of separation and reunion. Sohrai is a festival of animal worship. The songs sung during this festival are called "Gewar" or "Chanchar," which are sung to the beat of the dhol and dhak. Tusu Geet, sung during Makar Sankranti, symbolizes collective unity. Labor songs are songs sung to relieve fatigue while working in the fields, harvesting, or sowing crops. On the occasion of a wedding, when the men leave for the wedding procession, the women of the house perform a play called "Dumkach."

Khortha folk tales feature more ordinary people, animals, and flora than kings and queens. These tales often contain a moral lesson. Some of the prominent folk tales include "Budhi Aur Okar Naati," and others. Khortha folk literature is steeped in humanism and love for nature. Even in their poverty, the common people here know how to celebrate through songs.

1.4 KHORTHA/KHOTTA PARALLEL CORPUS

Khortha is an Indo-Aryan language with a clear grammatical structure and strong linguistic features. Its literature is written in the Devanagari script. Folk literature is abundant in Khortha, featuring clear gender and number forms, a case system, a pronoun system, and a simple SOV sentence structure. It's written in the Devanagari script.

Khortha is an Aryan language, but it is deeply influenced by tribal languages like Mundari, Santhali, and Kudukh. It has some unique features:

- The palatal //ʃ/ and the cerebral /ʂ/ are not pronounced in Khortha; instead, the dental /s/ is used. Sometimes, the /kʰ/ is also used in place of the cerebral /ʂ/. For example- bʱɑʂɑ (languge) - bʱɑsɑ / bʱɑkʰɑ, viʂ (poison)- bis / bikʰ.
- In Khortha, the retroflex /ɳ/ is not pronounced; the dental /n/ is used instead: gənɳeʃ - gənes, guɳɑ - gunɑ.
- Conversion of /v/ to /b/: If /v/ is at the beginning of words, it becomes /b/: e.g., van - ban, var – bar.
- Another interesting fact is the tendency to add /i/ or /u/ sounds before, after, or in the middle of words. For example, skuːl (school)-iskuːl, kʰɑ (eat)-kʰɑi, ɖɑl (dal)-ɖɑil.
- To form the plural, gulɑ, guliː, gulәin, gilәin, gilin, ɑin etc. are added.

Khortha 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 KHORTHA/KHOTTA PARALLEL CORPUS

The development of Khortha Parallel Corpus, Hindi was used as the source language which was translated into Khortha. The Khortha 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, 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, 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 Khortha which are systematically structured based on 159 grammatical categories. Khortha has 31,131 word count and 141,319 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] DANGI, A.K. (2012) Khortha Bhasha: Ek Parichay (Khortha Language: An Identity), Jamshedpur, Jharkhand; Spardha Publications.
[2] LangLex. (n.d.). Khortha/Khotta: Mother tongue profile (2011 Census report). Retrieved March 5, 2026, from https://langlex.com/cens/MTProfile.php?mtname=Khortha%2FKhotta
[3] Office of the Registrar General & Census Commissioner, India. (2018). C-16: Population by mother tongue (Table). Ministry of Home Affairs, Government of India. https://language.census.gov.in/showMTSIdata
[4] OHDAR, B. (2012) Khortha Bhasha Evam Sahitya: Udbhav Evam Vikash (Khortha Language and Literature: Origin and Growth), Ramgarh: Khortha Bhasha Sahitya Academy.
[5] Rejitha K. S. and Narayan Kumar Choudhary. (ed.). 2025. LDC-IL Corpus Insights. Central Institute of Indian Languages, Mysore. 978-93-48633-33-0.