

Kachchhi Parallel Text Corpus

**Linguistic Features
and Structures**

KACHCHHI 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:

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

Kachchhi Parallel Text Corpus: Linguistic Features and Structures

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

Contributors: Fatima Imtiaz Dhafrani, Imtiaz Dhafrani.

e-ISBN: 978-81-69099-89-9

CIIL Publication No.: 1642

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: Saurabh Varik

Table of Contents

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

KACHCHHI 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 KACHCHHI: A BRIEF INTRODUCTION

Kachchhi (also spelled Kutchi or Kachchi) is an Indo-Aryan language belonging to the Indo-Iranian branch of the Indo-European language family, primarily spoken in the Kachchh (Kutch) region of Gujarat, as well as in parts of Rajasthan, Sindh (Pakistan), and among diaspora communities. It is derived from one of the prakrit, or ancient vernacular languages of India that evolved into the modern Indo-Aryan languages of the Northwestern subdivision. Linguistically, it occupies a transitional position between Sindhi and Gujarati and shows influences from Sindhi, Gujarati, Rajasthani, and Hindi-Urdu, along with loanwords from Persian and Arabic due to historical trade and cultural contact. Kachchhi is mainly an oral language, though it is written using the Gujarati script in India and the Perso-Arabic script in Pakistan, with older scripts like Khojki and Khudabadi now largely obsolete. According to data presented on the LangLex website, Kachchhi is listed as a mother tongue in the Indian Census (2011) with a little over one million speakers, often grouped under the broader Sindhi language category for census classification. LangLex primarily provides statistical and census-based information, highlighting the demographic presence of Kachchhi rather than detailed grammatical description, but its data

underlines the language’s significant speaker base and its continued relevance in the multilingual landscape of India. (Source: <https://kachchh.nic.in/people-of-kachchh/>.)

Kutchi is sometimes considered a dialect of Sindhi by some scholars and a dialect of Gujarati by others. The language shares many words with both Sindhi and Gujarati and speakers of these languages can understand Kutchi to some extent. Kutchi is spoken mainly in the Kutch region and is influenced by nearby languages such as Sindhi, Kathiawari (a dialect of Gujarati), and Marwari (a dialect of Rajasthani), which together create a mixed linguistic character (Government of India Ministry of Home Affairs, 2001).

According to the 2001 Census of India, Kutchi is listed as a sub-language of Sindhi. However, many people believe Kutchi should be recognized as an independent language. Efforts to preserve and promote Kutchi include institutions like the Kutchi Sahitya Academy and publications such as the Kutchi-Mitra Daily newspaper. Out of about 2.5 million Sindhi speakers in India, around 823,058 people reported Kutchi as their mother tongue. Sindhi itself is one of the 22 official languages listed in Schedule VIII of the Indian Constitution and is also widely spoken in Pakistan (Government of India Ministry of Home Affairs, 2001; Maheshwari, 2010; University of California, n.d.).

1.4 KACHCHHI PARALLEL CORPUS

Collecting a Kachchhi parallel corpus is difficult mainly because of the socio-linguistic and resource constraints surrounding the language. Kachchhi is largely a spoken language, used in homes, markets, and community interactions, while written usage is limited. There is no single, strongly standardized writing tradition—people write Kachchhi in Gujarati, Devanagari, or sometimes Perso-Arabic scripts, which makes alignment with a single source language inconsistent. In addition, most available texts are informal, short, or mixed with Gujarati, Sindhi, or Hindi, reducing the availability of clean, sentence-level parallel data required for corpus creation.

Another major challenge is language mixing and variation. Many speakers unconsciously insert words from dominant regional languages due to limited written vocabulary in Kachchhi or habit formed through education in other languages. Dialectal variation across regions (Kutch, urban vs. rural areas, migrant communities) further complicates one-to-one translation and alignment. From a project perspective—especially in institutional efforts like those under LDC-IL—there are also issues of finding trained translators, ensuring linguistic consistency, handling non-standard spellings, and obtaining copyright-free, reusable source texts. All these factors together make building a sizable and high-quality Kachchhi parallel corpus a slow and demanding process.

Due to very limited reference books and dictionaries in Kachchhi, it is often difficult to find exact native Kachchhi words; therefore, Gujarati or other language words are sometimes used to make the text workable. The scarcity of resources, along with poor internet connectivity and other technical issues, has also created additional challenges during the work.

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

1.5 SUMMARY OF KACHCHHI PARALLEL CORPUS

The Kachchhi 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, Chirri, 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, 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, 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 Kachchhi which are systematically structured based on 159 grammatical categories. Kachchhi has 30,410 word count and 143,091 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://kachchh.nic.in/people-of-kachchh/>
- [3] Government of India Ministry of Home Affairs, 2001; Maheshwari, 2010; University of California, n.d.).
- [4] <https://languex.com/cens/MTPProfile.php?mtname=Kachchhi>
- [5] <https://language.census.gov.in/showMTSIData#>