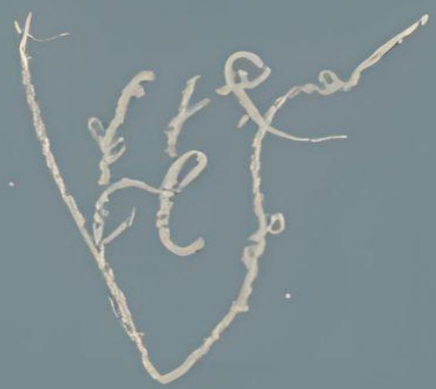
Khandeshi Parallel Text Corpus:

Linguistic Features and Structures



KHANDESHI 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

Khandeshi Parallel Text Corpus: Linguistic Features and Structures

Authors: Mr. Saurabh Varik, Dr. Rejitha K. S., Dr. Narayan Choudhary, Prof. Shailendra Mohan
Contributors: Dr. Gajanan Suresh Wankhede, Dr. Urmila Patil.

e-ISBN: 978-81-69099-20-2
CIIL Publication No.: 1637

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

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

KHANDESHI 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 KHANDESHI: A BRIEF INTRODUCTION

Khandeshi (also known as Khandeshi Bhili or Ahirani) is an Indo-Aryan language spoken mainly in the Khandesh region of north-western Maharashtra, particularly in the districts of Dhule, Jalgaon, and Nandurbar, with some speakers also found in parts of Madhya Pradesh and Gujarat (Census of India 2011; Ethnologue). Linguistically, Khandeshi belongs to the Western Indo-Aryan subgroup and shows strong influence from neighboring languages such as Marathi, Gujarati, and Bhili, as well as some influence from Hindi due to regional contact (Masica 1991; Ethnologue).

Structurally, Khandeshi generally follows the Subject–Object–Verb (SOV) word order and uses postpositions rather than prepositions, which is typical of most Indo-Aryan languages (Masica 1991). Nouns in Khandeshi inflect for grammatical categories such as gender, number, and case, while verbs agree with the subject in gender, number, and person and also mark tense, aspect, and mood (Masica 1991; Ethnologue). Phonologically, the language preserves several regional and archaic features that distinguish it from Standard Marathi, including specific vowel

realizations and vocabulary items that reflect contact with Bhili and Gujarati varieties (Masica 1991; Ethnologue).

According to the LangLex Census Mother Tongue Profile, the total number of Khandeshi speakers reported in the Census of India 2011 is 1,860,236. The data also shows that Ahirani (1,636,465 speakers) is the largest component, followed by Dangi (150,674), Gujari (57,171), and Khandeshi proper (10,670) (LangLex Census MT Profile; Census of India 2011).

This census-based profile indicates that Khandeshi is recorded as a language group consisting of several related mother tongues, with Ahirani forming the majority of speakers within the Khandeshi category (LangLex Census MT Profile; Census of India 2011).

Khandeshi is widely used in everyday oral communication, folk songs, and local storytelling traditions, while written usage remains limited and is usually represented in the Devanagari script (Ethnologue; Census of India 2011). Due to limited standardization and the scarcity of formal educational and digital resources, Khandeshi is considered an under-resourced language in linguistic and technological domains despite having a relatively large number of speakers (Census of India 2011; Ethnologue).

1.4 KHANDESHI PARALLEL CORPUS

Creating a Parallel Corpus for the Khandeshi language in India is relatively difficult due to its under-resourced nature and strong dependence on oral usage. Khandeshi has very limited standardized written material such as books, newspapers, or official documents, which makes it hard to find source texts for parallel alignment with languages like English, Hindi, or Marathi. For example, while Marathi has well-established translations of government schemes or literary texts, the same content rarely exists in written Khandeshi, so translators must rely on oral renditions and native speaker intuition. Another challenge is high dialectal variation across regions like Jalgaon, Dhule, and Nandurbar, where the same sentence may be expressed differently, making consistency difficult. For instance, a simple sentence like “He went to the market” may have multiple valid Khandeshi forms depending on local speech. Additionally, Khandeshi speakers often mix Marathi, Hindi, or Gujarati vocabulary due to language contact, so finding pure Khandeshi equivalents for technical or abstract terms becomes challenging. Lack of trained linguists, limited digital tools, and low awareness about standardized spelling further complicate corpus creation. As a result, building a reliable Khandeshi parallel corpus requires extensive fieldwork, native speaker validation, and careful normalization to balance linguistic authenticity with consistency.

During PCP work, I faced notable confusion between Khandeshi, Ahirani, and other Khandeshi dialects, which created challenges in data validation. Some resources rejected data by stating that it was Ahirani and not Khandeshi, while other resources rejected the same data by claiming it was Khandeshi and not Ahirani. This lack of clear linguistic boundaries among closely related dialects led to inconsistency and delays in acceptance. In addition, many contributors had limited technical knowledge, such as basic computer handling, keyboard typing, and familiarity with digital tools, which affected data quality and productivity. Other recurring issues included the frequent use of spoken forms in written texts, variation between spoken and written conventions,

and confusion in representing aspiration and non-aspiration sounds correctly. These linguistic and technical challenges together made the creation of a clean and consistent Khandeshi Parallel Corpus more complex and time-consuming.

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

1.5 SUMMARY OF KHANDESHI PARALLEL CORPUS

The Khandeshi 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, 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 Khandeshi which are systematically structured based on 159 grammatical categories. Konkani has 27,088 word count and 153,178 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] Census of India. 2011. C-17 Population by Mother Tongue. Government of India.
[3] Ethnologue: Languages of the World. Entry for Khandeshi / Ahirani.
[4] Masica, Colin P. 1991. The Indo-Aryan Languages. Cambridge University Press.
[5] https://language.census.gov.in/showMTSIdata#
[6] https://langlex.com/cens/MTProfile.php?mtname=Khandeshi
[7] LangLex. Census Mother Tongue Profile: Khandeshi. . https://langlex.com/cens/MTProfile.php?mtname=Khandeshi