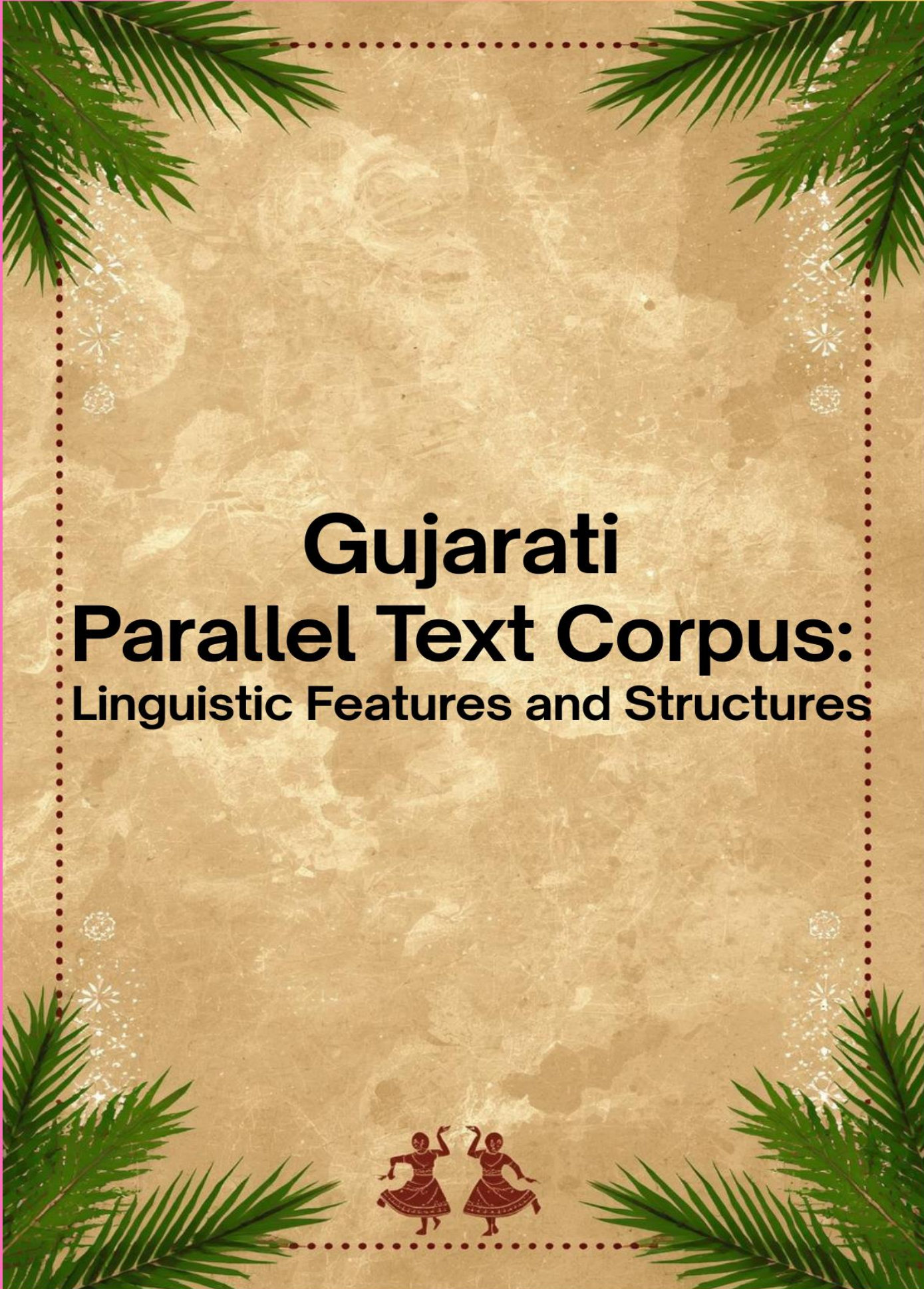




Gujarati

Parallel Text Corpus:

Linguistic Features and Structures



GUJARATI 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:

***Dr. Vijayalaxmi F Patil,
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

Gujarati Parallel Text Corpus: Linguistic Features and Structures

Authors: Dr. Vijayalaxmi F Patil, Dr. Rejitha K. S., Dr. Narayan Choudhary, Prof. Shailendra Mohan

Contributors: Himani Kanojia, Natwarlal Modha, Dr. Rozy Patel

e-ISBN: 978-81-69099-85-1

CIIL Publication No.: 1652

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: Dr. Vijayalaxmi F Patil

Table of Contents

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

GUJARATI PARALLEL TEXT CORPUS

Dr. Vijayalaxmi F Patil

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 GUJARATI: A BRIEF INTRODUCTION

Gujarati is one of the important modern Indo-Aryan languages of India, spoken by 5,503,6204 people according to the census of India report 2011. Its primary home is the western Indian state of Gujarat, where it functions as the dominant language of daily life, education, administration and cultural expression. Beyond Gujarat, Gujarati-speaking communities are also found in neighbouring states such as Rajasthan and Maharashtra particularly in border districts and major cities. Due to centuries of trade, migration, and entrepreneurship, Gujarati has developed a strong global presence. Significant diaspora communities exist in countries including the United Kingdom, United States, Canada and South Africa. In many of these countries Gujarati is maintained as a heritage language used in homes, religious institutions, community organizations and ethnic media.

Gujarati belongs to the Indo-Aryan branch of the Indo-European language family which also includes languages such as Hindi, Marathi, Bengali and Punjabi. Historically, Gujarati developed from Old Gujarati between the 12th and 14th centuries. Old Gujarati itself evolved from earlier Middle Indo-Aryan forms known as Prakrit which ultimately trace back to Sanskrit, the ancient classical language of India.

Over time, Gujarati absorbed vocabulary and stylistic influences from several foreign languages. Persian and Arabic influences entered during medieval Islamic rule in western India. Portuguese loanwords became part of the language through early maritime trade and colonial contact along the western coast. Later, during British rule, English contributed administrative, educational and technical vocabulary. As a result, modern Gujarati reflects layers of historical interaction while retaining its Indo-Aryan grammatical foundation.

Gujarati is written in the Gujarati script which evolved from the Devanagari script. While both scripts share a common origin and similar letter forms, the Gujarati script is visually distinct because it lacks the horizontal headline (Shiro Rekha) that typically runs across the top of Devanagari letters. This gives Gujarati text a more rounded and open appearance.

The script is fully standardized and used widely in schools, universities, newspapers, books, government documents, digital media and business communication. It supports a rich print culture and a growing presence in online platforms and technology.

Dialects and Regional Varieties:

Gujarati is not uniform across all regions; it includes several major dialects:

- Kathiawari – Spoken in the Saurashtra (Kathiawar) region; known for distinct pronunciation and vocabulary.
- Surti – Used in South Gujarat, especially around Surat; influenced historically by trade and maritime contact.
- Charotari – Spoken in Central Gujarat; often considered close to standard Gujarati.
- Kutchi – Spoken in the Kutch region; sometimes regarded as a separate but closely related language due to its unique phonological and lexical features.

Despite these variations, speakers of different dialects generally understand one another.

Gujarati possesses a rich and continuous literary tradition. Early literature includes devotional (bhakti) poetry dedicated to Hindu deities, as well as Jain religious writings. Over time, the language expanded into folk literature, narrative poetry, drama, essays, novels and modern short stories. The 19th and 20th centuries saw the growth of reformist, nationalist, and modernist writing.

Gujarati was the mother tongue of prominent national leaders such as Mahatma Gandhi and Sardar Vallabhbhai Patel, both of whom used the language in speeches and writings to connect with the masses.

Gujarati is one of the 22 scheduled languages recognized by the Constitution of India. It serves as the official language of the state of Gujarat and is used in government administration, legislative proceedings, education, print and electronic media and everyday communication.

Today, Gujarati remains a vibrant and dynamic language. It continues to evolve while preserving its historical depth and cultural identity. With a strong literary heritage, institutional support and an active global diaspora, Gujarati maintains both regional strength and international presence.

1.4 GUJARATI PARALLEL CORPUS

Gujarati is a highly agglutinative language with rich morphophonemic features. It exhibits complex morphophonemic alternations, extensive case marking, productive compounding, and relatively flexible word order. Gujarati language is written in Devanagari script derived from the

Brahmi tradition. It has an extensive set of consonant and vowel symbols that precisely capture its phonemic distinctions. It generally follows Subject-Object-Verb pattern but it is linguistically rich and structurally complex language. These features contribute to Gujarati's uniqueness among the languages of India and highlight its importance in the study of linguistics.

Gujarati 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 GUJARATI PARALLEL CORPUS

The Gujarati 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, Chirri, Chungli, Churahi, Coorgi/Kodagu, Deori, Dhundhari, Dimasa, Dogri, Gangte, Garhwali, Garo, Gojri/Gujjari/Gujar, Gujarati, 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, Sambalpur, 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 Gujarati which are systematically structured based on 159 grammatical categories. Gujarati has 29,060 word count and 144,914 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.