

Gojri/Gujjari/Gujar Parallel Text Corpus: Linguistic Features and Structures



GOJRI/GUJJARI/GUJAR 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. Mansoor Khan

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

Gojri/Gujjari/Gujar Parallel Text Corpus: Linguistic Features and Structures

Authors: Dr. Mansoor Khan, Dr. Rejitha K. S., Dr. Narayan Choudhary, Prof. Shailendra Mohan

Contributors: Naseem Akhtar, Mushtaq Khalid Chaudary, Tariq Hussain Abrar

e-ISBN: 978-81-69099-72-1

CIIL Publication No.: 1653

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. Mansoor Khan

Table of Contents

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

GOJRI/GUJJARI/GUJAR PARALLEL TEXT CORPUS

Dr. Mansoor Khan

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 GOJRI/GUJJARI/GUJAR: A BRIEF INTRODUCTION

Gojri (also spelled Gujari, Gojari, Gujri, Gujar, Guzari, etc.) is an Indo-Aryan language spoken primarily by the Gurjar/Gujjar ethnic community across South Asia. It belongs to the Indo-European → Indo-Iranian → Indo-Aryan → Western Rajasthani–Marwari group of languages and is noted for its rich oral traditions and cultural heritage.

The language is often treated as a distinct linguistic identity or as part of the broader Rajasthani dialect continuum, depending on the classification framework used. It shares linguistic similarities with Rajasthani, Punjabi, Haryanvi, Dogri, and Pahari languages, particularly in phonology and morphology, reflecting centuries of contact and shared history in the north western Indian subcontinent.

Geographic Distribution and Speakers

Gojri/Gujari is spoken across a broad geographic region including:

- Northern India — especially Jammu & Kashmir, Himachal Pradesh, Punjab, Uttarakhand, Uttar Pradesh, Haryana, Rajasthan, and Madhya Pradesh.
- Pakistan — most notably in Azad Kashmir, Gilgit-Baltistan, and Khyber Pakhtunkhwa.
- Eastern Afghanistan, where smaller Gujari-speaking communities exist.

According to Ethnologue and related language profiles, Gujari has an estimated speaker population ranging from around 1.5 million to nearly 2 million speakers globally, though exact numbers varies due to census categorizations and language reporting practices.

In the Indian census of 1961 and 1971 (often referenced via historical linguistic surveys), Gojri was recorded as being spoken by hundreds of thousands of people in Jammu & Kashmir and adjoining regions; at one point it appeared as the third-most spoken language there after Kashmiri and Dogri.

The Census lists Gojri/Gujjari/Gujar as one of the mother tongues with 12,27,901 persons reporting it as their mother tongue in India.

These figures come from the detailed language classification tables where many smaller languages or dialects are aggregated under larger language categories (in this case Hindi), which is a common practice in Indian Census language reporting.

1.4 GOJRI/GUJJARI/GUJAR PARALLEL CORPUS

Gojri/Gujjari/Gujar is a highly agglutinative language with rich morphophonemic features. It exhibits complex morphophonemic alternations, extensive case marking, productive compounding, and relatively flexible word order. In regions like Rajasthan and Himachal Pradesh, Gojri/Gujjari/Gujar is also written in Devanagari, the script used for Hindi.

In Jammu and Kashmir and Pakistan, Gojri is commonly written in the Perso-Arabic script (Urdu script). This script is widely used due to historical administrative and educational influence. 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 Gojri/Gujjari/Gujar's uniqueness among the languages of India and highlight its importance in the study of linguistics.

Gojri/Gujjari/Gujar 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. It is to be noted that for the development of Gojri/Gujjari/Gujar Parallel Corpus, Hindi was used as the source language which was translated into Gojri/Gujjari/Gujar.

1.5 SUMMARY OF GOJRI/GUJJARI/GUJAR PARALLEL CORPUS

The Gojri/Gujjari/Gujar 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, Garo, 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, 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 Gojri/Gujjari/Gujar which are systematically structured based on 159 grammatical categories. Gojri/Gujjari/Gujar has 29,139 word count and 142,950 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] [Gojri/Gujjari/Gujar Language Profile](#)
- [3] [Language Census Data](#)