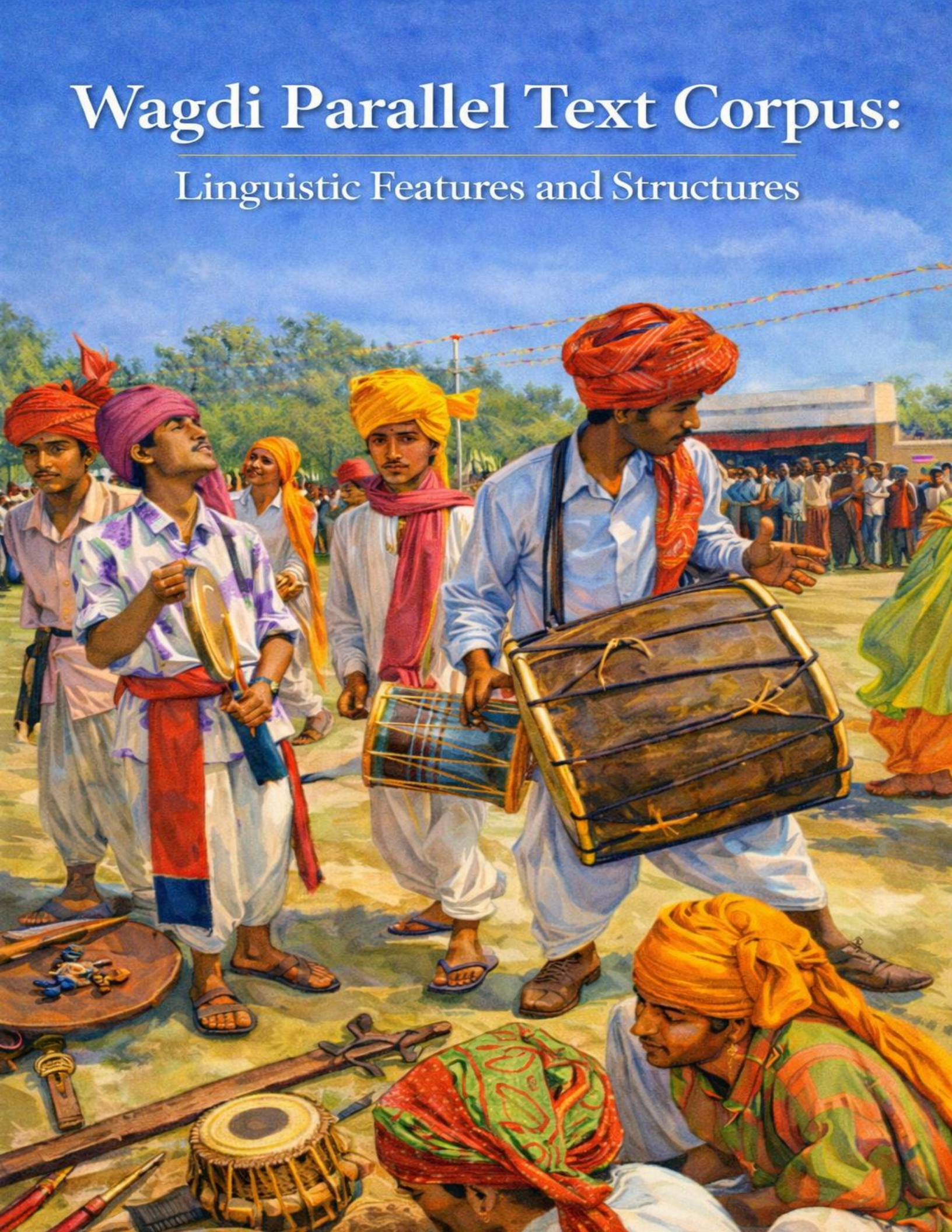
Wagdi Parallel Text Corpus:

Linguistic Features and Structures



WAGDI 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:

Chetan Baji

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

Wagdi Parallel Text Corpus: Linguistic Features and Structures

Authors: Chetan Baji, Dr. Rejitha K. S., Dr. Narayan Choudhary, Prof. Shailendra Mohan

Contributors: Hoshang Panchal, Kushagra Panchal

e-ISBN: 978-81-69175-13-5
CIIL Publication No.: 1715

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: Chetan Baji

Table of Contents

WAGDI PARALLEL TEXT CORPUS

Chetan Baji

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 systems, 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 WAGDI: A BRIEF INTRODUCTION

Wagdi is an Indo-Aryan language spoken primarily by the Bhil tribal communities of western India. It forms part of the Bhil subgroup within the Western Indo-Aryan branch of the Indo-European language family. The language derives its name from the historic Vagad (Wagad) region, which encompasses parts of present-day southern Rajasthan and adjoining areas of Gujarat and Madhya Pradesh. With approximately 3 to 3.5 million speakers according to the 2011 Census of India, Wagdi represents an important yet under-documented linguistic tradition of India.

Geographic Distribution and Speech Community; Wagdi is predominantly spoken in Dungarpur, and Banswara districts of Rajasthan, Dahod and Panchmahal districts of GujaratBordering tribal regions of western Madhya Pradesh. It shares structural affinities with Gujarati and Rajasthani, while retaining distinctive Bhil linguistic traits. Due to prolonged geographical contact, Wagdi demonstrates features of a transitional contact language within the Western Indo-Aryan zone.

Script and Writing Practices: Historically, Wagdi has been a predominantly oral language with no indigenous script of its own. In contemporary usage, it is written in Devanagari script (especially in Rajasthan), Gujarati script (in Gujarat region), Occasionally Roman script for

linguistic documentation. The absence of a standardized orthography has posed challenges for literary development and formal educational implementation.

In recent years, Wagdi has gained attention in linguistic documentation and language technology initiatives. Efforts toward corpus creation and computational resource development are being undertaken by organizations such as Linguistic Data Consortium for Indian Languages (LDC-IL, CIIL). The development of parallel text corpora, machine translation systems, and inclusive AI models for Indian languages has highlighted the importance of resource creation for low-resource languages like Wagdi.

Challenges and Preservation: Major challenges faced by the Wagdi language including lack of standardized grammar and orthography, limited written literature and underrepresentation in digital platforms.

Wagdi stands as an important representative of the Bhil subgroup within the Western Indo-Aryan linguistic tradition. Rooted in the Vagad region and closely tied to the cultural identity of the Bhil community, it embodies a dynamic blend of indigenous and Indo-Aryan linguistic elements. While historically oral and under-documented, Wagdi is gradually gaining recognition in academic research, corpus development, and inclusive language technology initiatives.

1.4 WAGDI PARALLEL CORPUS

Wagdi is morphologically rich Indo-Aryan language of the Bhil subgroup, characterized by notable morphophonemic variation and extensive use of postpositional case marking. It exhibits phonological alternations influenced by regional contact, productive compounding, and a relatively flexible word order within an overall Subject–Object–Verb (SOV) syntactic pattern. Traditionally, Wagdi has been an oral mother tongue with no independent script; however, in written contexts it is commonly represented using Devanagari and, in some regions, the Gujarati script. Its phonological system includes a range of vowels, retroflex and aspirated consonants, and nasalized sounds that reflect both Indo-Aryan structure and regional linguistic influence. Owing to its hybrid lexical composition, shaped by sustained contact with Gujarati, Rajasthani, and Hindi, Wagdi is linguistically diverse and structurally dynamic.

As a mother tongue of the Bhil community, Wagdi constitutes a valuable resource for the development of parallel text corpora with other Indian languages. Such parallel corpora are particularly useful for examining syntactic variation, morphological complexity, code-mixing patterns, dialectal diversity, and script representation in multilingual settings like India.

The inclusion of Wagdi in corpus-based linguistic research and language technology initiatives can significantly contribute to the documentation of low-resource tribal languages and support inclusive and representative natural language processing for Indian mother tongues.

This Wagdi parallel corpus constitutes a systematically translated dataset derived from an original English corpus, ensuring structured cross-linguistic alignment and semantic correspondence.

1.5 SUMMARY OF WAGDI PARALLEL CORPUS

The Wagdi parallel text corpus is aligned with English and 146 mother tongues of India. And 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, 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 Wagdi t which are systematically structured based on 159 grammatical categories. Wagdi has 31752 word count and 130757 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] **Grierson, G. A. (1908–1928).**
Linguistic Survey of India.
Government of India.
https://ruralindiaonline.org/library/resource/linguistic-survey-of-india---rajasthan-part-i/