

Pawari/Powari

Parallel Text Corpus:

Linguistic Features and Structures



PAWARI/POWARI 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

Pawari/Powari Parallel Text Corpus: Linguistic Features and Structures

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

Contributors: Dr. Tufan singh pardhi, Dr. Shobha

e-ISBN: 978-81-69175-95-1

CIIL Publication No.: 1749

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

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

PAWARI/POWARI 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 PAWARI/POWARI: A BRIEF INTRODUCTION

Pawari, also known as Powari, Panwari is an important regional language that belongs to the Indo-Aryan branch of the Indo-European language family. The total population of Pawari is 325,772 as per the census of India report 2011. It is considered to be one of approximately 120 dialects of Hindi. This dialect is primarily spoken in the southeastern region of Madhya Pradesh and in the adjoining southern border areas of Maharashtra. Pawari is widely spoken in the districts of Balaghat and Seoni in Madhya Pradesh, as well as in Gondia, Bhandara, Nagpur, Wardha, and Yavatmal districts of Maharashtra. In addition, Pawari-speaking communities reside in major metropolitan cities such as Mumbai, Pune, Bengaluru, Raipur, Bhopal, and Indore, and even abroad.

Geographically, the principal settlement area of the Pawari-speaking population has been around the banks of the Wainganga River in Madhya Pradesh.

The Pawari/Powari-speaking community identifies itself as belonging to the Paramara Rajput (Kshatriya) lineage. The community considers King Vikramaditya and the illustrious King Bhoja of the Paramara dynasty as their ancestors. According to historical traditions, their original

homeland is believed to have been the Malwa region, particularly around Dhar. Over time, due to political, social, and economic reasons, the community migrated from Malwa toward Rajasthan, Gujarat, and Maharashtra. In the Vidarbha region, their significant settlement was at Nagardhan (in the Ramtek area). From there, they gradually expanded into Balaghat, Seoni, and other nearby districts. With time, they abandoned their earlier martial occupation and adopted agriculture and animal husbandry as their primary means of livelihood.

Due to migration from Malwa, the Pawari language shows strong similarities with the Malvi dialect. In fact, Pawari is considered very close to Malvi. Along with this, the dialect incorporates vocabulary from Bundeli, Bagheli, Marathi, and other regional languages.

In the Parsawada, Baihar, and Birsa areas of Balaghat district, Pawari shows noticeable influence from the tribal language Gondi and the “Khaltahi” variety of Chhattisgarhi. In Seoni district, elements of Bundeli are evident in Pawari. In the coastal areas of Gondia, Bhandara, Nagpur, and Wardha, Pawari reflects a stronger influence of Marathi. Thus, Pawari serves as a powerful example of regional linguistic contact and cultural exchange.

Traditionally, Pawari has been predominantly an oral language. Its cultural richness has been preserved through folk songs, seasonal songs, proverbs, myths, legends, and narratives passed down from generation to generation. Oral performance plays a vital role during festivals, marriages, harvest celebrations, and other community rituals. When written, Pawari is generally recorded in the Devanagari script, which is also used for Hindi and Marathi. However, a fully standardized orthography has not yet been developed, and written usage remains limited compared to oral tradition. Scholars, local activists, and community members have shown growing interest in documenting its vocabulary, grammar, and oral literature. Such efforts aim to preserve Pawari as an important component of the linguistic diversity of Madhya Pradesh and central India.

Renowned linguists George Abraham Grierson and R. V. Russell also located the Pawari/Powari-speaking community around the Vidarbha region in their studies. Pawari society is known for its distinct social traditions, customs, and cultural heritage. In this community, language is not merely a means of communication but also a foundation of collective identity and historical memory.

1.4 PAWARI/POWARI PARALLEL CORPUS

Pawari is a highly agglutinative language with rich morphophonemic features. It exhibits complex morphophonemic alternations, extensive case marking, productive compounding, and relatively flexible word order. Pawari 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 Pawari’s uniqueness among the languages of India and highlight its importance in the study of linguistics.

Pawari 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 PAWARI/POWARI PARALLEL CORPUS

The Pawari 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, Gujar, 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, Kurmal 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, 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, Tikhri, Tripuri, Tulu, Urdu, Vaiphei, Wagdi, Wancho, Yerava, Yerukala/Yerukula, Yimchungre, Zeliang, Zemi, Zou.

There are 5,332 sentences/phrases in Pawari which are systematically structured based on 159 grammatical categories. Pawari has 31,903 word count and 140,675 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] Grierson, George Abraham. Linguistic Survey of India. Vol. 9: Indo-Aryan Languages: Central Group. Calcutta: Office of the Superintendent of Government Printing, India, 1908.
- [2] Russell, R. V., and Hiralal. The Tribes and Castes of the Central Provinces of India. London: Macmillan & Co., 1916.
- [3] Chatterji, Suniti Kumar. Indo-Aryan and Hindi: Historical and Linguistic Studies. Calcutta: Firma K. L. Mukhopadhyay, 1970.
- [4] टेंभरे, ज्ञानेश्वर. पवारी ज्ञानदीप. नागपुर: हिमालया पब्लिकेशन, 2011.
- [5] Balaghat District Gazetteer. Bhopal: Government of Madhya Pradesh, 1971.
- [6] Rejitha K. S. and Narayan Kumar Choudhary. (ed.). (2025). LDC-IL Corpus Insights. CIIL, Mysore. 978-93- 48633-33-0.