

Muwasi

Parallel Text Corpus :

Linguistic Features and Structures



# MUWASI 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:**

***Bhageshree K. Khandale***

***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

### Muwasi Parallel Text Corpus: Linguistic Features and Structures

*Authors:* Bhageshree K. Khandale, Dr. Rejitha K. S., Dr. Narayan Choudhary, Prof. Shailendra Mohan

*Contributors:* Arjun Banke, Sachin Banke

*e-ISBN:* 978-81-69099-92-9

*CIIL Publication No.:* 1593

*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:* Tejaswini A.G & Bhageshree K. Khandale

## Table of Contents

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

## MUWASI PARALLEL TEXT CORPUS

*Bhageshree K. Khandale*

### 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 MUWASI: A BRIEF INTRODUCTION

The term Muwasi (also spelled Mowasi/mouasi/Mawasi) refers specifically to one of the eastern dialects of Korku language, which is found in areas including the Chhindwara, Khandwa, Betul, and Narmadapuram districts of Madhya Pradesh, as well as in Himachal Pradesh, Rajasthan, Chhattisgarh, and other states. Related Korku dialects are also spoken in northern parts of Maharashtra where many tribal Korku people use it in everyday communication. As per the 2011 Census of India, 35,827 speakers identified Muwasi as their mother tongue within the broader Korku language group.

The Muwasi mother tongue belongs to the Austro-Asiatic language family, making it linguistically related to languages of eastern India rather than the surrounding Indo-Aryan languages. It is considered the westernmost Austro-Asiatic language in India and is written using a form of the Devanagari script (Balbodh style).

## 1.4 MUWASI PARALLEL CORPUS

The Muwasi variety represents one of the regional speech forms within this broader linguistic tradition rather than a completely separate language system. In the context of Madhya Pradesh (India), what's often referred to as the Muwasi language is not a separate language in its own right, but rather a variant or dialect of the Korku language spoken by the indigenous Korku tribal community of central India. However, the Muwasi dialect differs from Korku, particularly in the formation of negative verbs. Muwasi language share large number of features at all grammatical levels. The linguistic similarities like three number system, animate and inanimate distinction, word order Subject-Object-Verb (SOV) etc., may be concluded that Muwasi resemblance with the other Austro-Asiatic languages.

Muwasi is traditionally an oral language, though the use of the Devanagari script is becoming more common in modern contexts. Literacy in Muwasi remains relatively low. Linguistic surveys indicate that the Muwasi language is spoken by several hundred thousand people, mainly in Madhya Pradesh and Maharashtra. Some sources classify Muwasi as a vulnerable language in terms of its vitality. Due to the influence of Hindi and Marathi, most of the Muwasi speakers are bilingual and also a large number of vocabularies of Hindi and Marathi are found in this language.

Muwasi 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 MUWASI PARALLEL CORPUS

The Muwasi 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, Chirr, 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, 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 Muwasi which are systematically structured based on 159 grammatical categories. Muwasi has 30,762 word count and 148,990 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] The Munda Languages, Gregory D. S. Anderson, 2008.
- [3] Linguistic Survey of India, Madhya Pradesh, 2024.
- [4] [Language Census Data](#)
- [5] <https://languex.com/cens/MTPProfile.php?mtname=Muwasi>