WANCHO
Parallel Text Corpus
Linguistic Structures and Features

WANCHO 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:

Syeda Mustafiza Tamim
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

Wancho Parallel Text Corpus: Linguistic Features and Structures

Authors: Syeda Mustafiza Tamim, Dr. Rejitha K. S., Dr. Narayan Choudhary, Prof. Shailendra Mohan

Contributors: Lemnon Wangjen, Omwang Wangjen

e-ISBN: 978-81-69175-46-3
CIIL Publication No.: 1714

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: Syeda Mustafiza Tamim

Table of Contents

WANCHO PARALLEL TEXT CORPUS

Syeda Mustafiza Tamim

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 WANCHO: A BRIEF INTRODUCTION

Wancho is a Tibeto-Burman language predominantly spoken by the Wancho Naga community in the Longding district of Arunachal Pradesh and in some parts of the Mon district of Nagaland in Northeast India. According to the 2011 Census of India, there are 56,866 speakers who identified Wancho as their mother tongue, while the 2001 Census recorded 49,072 speakers, indicating a gradual increase in the speaker population. The Wancho people are ethnically related to the Nocte and Konyak Naga tribes and possess a rich cultural heritage. Traditionally, Wancho existed as an oral language and played an important role in preserving the community's folklore, songs, rituals, and oral traditions. The community is also known for its rich tradition of tattooing, and in the past the people practiced headhunting, a custom that has now been completely discontinued.

A significant development in the history of the Wancho language is the creation of its own writing system by Banwang Losu between 2001 and 2012. The Wancho script was later officially included in Unicode 12.0 in 2019, enabling its use in digital communication and documentation. The social structure of the Wancho community traditionally follows a hierarchical system led by a hereditary chief known as the Wangham. One of the most important cultural celebrations of the Wancho people is the Oriah festival, which is celebrated during March or April with animal sacrifices and ceremonial dances performed around a symbolic pole known as the Jangban. With the introduction of its script and ongoing documentation efforts, Wancho continues to gain importance in linguistic preservation and cultural identity.

1.4 WANCHO PARALLEL CORPUS

Wancho is a Tibeto-Burman language that exhibits several distinctive linguistic features common to many languages of the Northern Naga group. The language shows agglutinative tendencies where grammatical meanings are often expressed through suffixes attached to lexical roots. Wancho generally follows a Subject–Object–Verb (SOV) word order, although variations may occur depending on discourse context. Traditionally, Wancho existed as an oral language with rich oral traditions including folklore, songs, and ritual narratives. In recent years, the development of the Wancho script by Banwang Losu has enabled the language to be written and documented systematically. The script provides a structured representation of the phonological system of the language and has facilitated linguistic analysis and documentation.

Wancho provides an important resource for developing parallel text corpora with other Indian languages. A Wancho parallel corpus can support linguistic and computational research by enabling comparative studies of syntactic structures, morphological patterns, lexical correspondences, and translation alignment between languages. Such corpora are particularly valuable for natural language processing tasks including machine translation, language documentation, and corpus-based linguistic analysis. In multilingual contexts like India, the development of a Wancho parallel corpus contributes to the preservation of a lesser-documented language while also supporting broader research in computational linguistics and language technology.

1.5 SUMMARY OF WANCHO PARALLEL CORPUS

The Wancho 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, 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, Yerava, Yerukala/Yerukula, Yimchungre, Zeliang, Zemi, Zou.

There are 5,332 sentences/phrases in Wancho which are systematically structured based on 159 grammatical categories. Wancho has 34,611 word count and 167,197 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]https://language.census.gov.in/eLanguageDivision_VirtualPath/MTSI/ARUNACHAL%20PRADESH/Wancho.pdf
[3] Census of India. 2011. C-16 Population by Mother Tongue. Office of the Registrar General & Census Commissioner, Government of India.
[4] Census of India. 2001. Language Tables (Mother Tongue Data). Office of the Registrar General & Census Commissioner, Government of India.
[5] Unicode Consortium. 2019. The Unicode Standard Version 12.0 (includes the Wancho script block).
[6] Banwang Losu. Developer of the Wancho script (2001–2012).
[7] SIL International. Ethnologue: Languages of the World – Wancho Language Profile.
[8] Burling, Robbins. 2003. The Tibeto-Burman Languages of Northeast India. In The Sino-Tibetan Languages, Routledge.
[9] Elwin, Verrier. 1959. The Nagas in the Nineteenth Century. Oxford University Press.