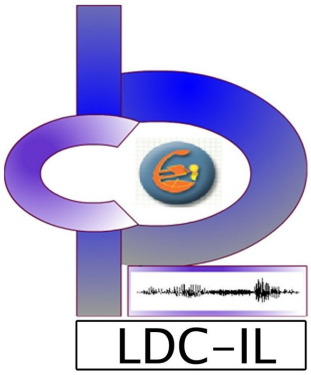
Nimadi Parallel Text Corpus: Linguistic Features and Structures

NIMADI 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:
Dr. Shahnawaz Alam
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

Nimadi Parallel Text Corpus: Linguistic Features and Structures

Authors: Dr. Shahnawaz Alam, Dr. Rejitha K. S., Dr. Narayan Choudhary, Prof. Shailendra Mohan

Contributors: Nikita, Rohit Patel, Ajay Patel, Shivpal Kalam,

e-ISBN: 978-81-69099-80-6
CIIL Publication No.: 1590

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: Shahnawaz Alam

Table of Contents

NIMADI PARALLEL TEXT CORPUS

Shahnawaz Alam

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 NIMADI: A BRIEF INTRODUCTION

Nimadi (also spelled Nimaadi or Nemadi) is an Indo-Aryan language spoken predominantly in the Nimar region of western Madhya Pradesh. The Nimar region, located along the fertile plains of the Narmada valley, includes districts such as Khandwa, Khargone, Barwani, and Dhar. Nimadi serves as a primary medium of daily communication among rural and semi-urban communities in this region. Linguistically, Nimadi belongs to the Indo-Aryan language family and is often considered a variety closely related to Rajasthani and Malvi, while also showing contact influences from Hindi and Marathi. Its phonological and lexical structure reflects a blend of Rajasthani features with regional innovations shaped by historical migration, trade, and cultural interaction.

The language is primarily transmitted through oral tradition, with rich expressions in folk songs, proverbs, storytelling, and devotional poetry. Although Devanagri is commonly used for writing Nimadi, there is no fully standardized orthography, and written materials remain limited. As a result, much of its linguistic heritage survives in oral narratives rather than formal documentation.

Socio-linguistically, Nimadi plays a crucial role in preserving regional identity and cultural continuity. It is widely used in informal domains such as home, marketplace, and local gatherings, while Hindi dominates formal education, administration, and media. This functional distribution has influenced patterns of bilingualism and code-switching among speakers.

In recent years, growing academic interest in regional languages and digital documentation efforts has highlighted the need to preserve and study Nimadi systematically. Developing linguistic resources such as dictionaries, grammars, and corpora is essential for safeguarding its vitality and ensuring its representation within India's multilingual framework.

1.4 NIMADI PARALLEL CORPUS

Nimadi is a morphologically rich Indo-Aryan language with distinct grammatical and phonological characteristics shaped by regional contact and historical development. Nimadi exhibits considerable morphological variation through inflectional processes such as gender agreement, number marking, case relations, and the use of postpositions. Verbal morphology reflects tense, aspect, mood, and agreement patterns typical of Indo-Aryan languages. The language also demonstrates morphophonemic alternations, particularly in plural formation, verb conjugation, and lexical borrowing from neighboring languages such as Hindi and Marathi. Nimadi generally follows a Subject–Object–Verb (SOV) word order, though pragmatic factors allow flexibility for emphasis and discourse functions. It is primarily written in the Devanagri script, but orthographic practices are not fully standardized, reflecting its strong oral tradition and limited formal documentation.

Nimadi also holds significant potential for the development of parallel text corpora with other Indian languages, particularly Hindi and English. A Nimadi parallel corpus would facilitate research on syntactic divergence, dialectal variation, lexical borrowing, and contact-induced change within the multilingual ecology of India. Given its status as a regionally dominant yet under-resourced language, such a corpus would be valuable for examining cross-linguistic influence, translation strategies, and structural correspondences. Moreover, it could contribute to the development of natural language processing tools, machine translation systems, speech technologies, and bilingual lexicons. Therefore, building a Nimadi parallel corpus is crucial not only for theoretical linguistic inquiry but also for advancing inclusive language technology and preserving regional linguistic heritage.

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

1.5 SUMMARY OF NIMADI PARALLEL CORPUS

The Nimadi parallel text corpus integrates English with 146 other Indian mother tongues. 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, 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 Nimadi which are systematically structured based on 159 grammatical categories. Nimadi has 32253 word count and 137392 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 REFERENCES

[1] Rejitha K. S. and Narayan Kumar Choudhary. (ed.). (2025). LDC-IL Corpus Insights. CIIL, Mysore. 978-93- 48633-33-0.
[2] Census of India (2011). Language Data and Linguistic Survey Reports.
[3] George Abraham Grierson. Linguistic Survey of India.