Nawait Parallel Text Corpus:

Linguistic Features and Structures

NAWAIT 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:

Mr. Saurabh Varik
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

Nawait Parallel Text Corpus: Linguistic Features and Structures

Authors: Mr. Saurabh Varik, Dr. Rejitha K. S., Dr. Narayan Choudhary, Prof. Shailendra Mohan
Contributors: Mohammad Zubair Jukaku, Sayed Mohiddin Affan, Khateebi Mohammed Imran

e-ISBN: 978-81-69099-66-0
CIIL Publication No.: 1592

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: Saurabh Varik

Table of Contents

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

NAWAIT PARALLEL TEXT CORPUS

Saurabh Varik

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 NAWAIT: A BRIEF INTRODUCTION

The Nawayathi language (also called Nawayati or Navayati) is spoken by the Nawayath Muslim community mainly in the coastal regions of Karnataka, especially in Bhatkal, Honnavar, Kumta, and nearby areas. It belongs to the Indo-Aryan branch of the Indo-European language family and is closely related to Konkani and Marathi, but it has developed as a distinct community language. Historically, the Nawayath people had strong trade relations with Arab merchants, and this contact influenced their language and culture. Because of this interaction, Nawayathi contains many Arabic and Persian loanwords, especially in religious, social, and cultural vocabulary. (Ethnologue; Glottolog; Linguistic Survey of India)

Linguistically, Nawayathi shows grammatical similarities with Konkani, including sentence structure and some vocabulary, but its pronunciation and lexical items reflect strong influence from Urdu and Arabic. Traditionally the language was written using the Perso-Arabic script, particularly for religious texts and community records, although today it is mostly used as a

spoken language within families and the community. Most Nawayath speakers are multilingual and also speak Kannada, Urdu, Konkani, or English depending on education and region. Despite its limited written use, Nawayathi remains an important marker of identity and heritage for the Nawayath community in coastal Karnataka. (Ethnologue; Glottolog; Masica 1991; Linguistic Survey of India) (Sources: Ethnologue: Languages of the World; Glottolog; G.A. Grierson – Linguistic Survey of India; Colin P. Masica – The Indo-Aryan Languages (1991)).

According to the Census of India 2011, a total of 13,123 people in India reported Nawait as their mother tongue, although it is officially grouped under the broader language name Konkani. These speakers are spread across 13 states, with the vast majority residing in Karnataka (12,775), followed by smaller populations in Kerala (98), Odisha (72), Goa (71), Maharashtra (55), Tripura (23), Tamil Nadu (21), Uttar Pradesh (3), NCT of Delhi (1), Rajasthan (1), Nagaland (1), Assam (1), and Chhattisgarh (1). Given its relatively small numbers, examining population distribution at the sub-district level would provide a more accurate picture of the language's presence and vitality (source: https://langlex.com/cens/MTProfile.php?mtname=Nawait+).

1.4 NAWAIT PARALLEL CORPUS

While creating a Nawait Language Parallel Corpus, several challenges are encountered at both linguistic and technical levels. Nawait is a low-resource language, and the availability of written resources such as grammars, dictionaries, textbooks, or standardized orthography is extremely limited, which makes accurate translation and alignment difficult. In many cases, native speakers rely heavily on Urdu, Konkani, or Arabic loanwords, leading to inconsistency in vocabulary and sentence structure across the corpus. Grammatical variation is another major issue, as Nawait is mostly used in spoken form and shows dialectal differences, lack of fixed spelling conventions, and flexible word order, making normalization and annotation complex. On the technical side, lack of digitized data, Unicode compatibility issues, font limitations, and poor internet connectivity slow down data collection and validation processes. Additionally, limited availability of trained native speakers with linguistic expertise affects quality control. Despite these challenges, parallel corpus creation for Nawait is crucial for language documentation, preservation, and future NLP applications, and systematic efforts can gradually improve resource availability and standardization.

Despite being a low-resource language, some valuable Nawait resources and contributors are already available and actively ready to provide data and support corpus creation, which is a strong and positive step toward systematic documentation and future linguistic work on Nawait.

The Nawait parallel corpus has been created using Urdu as the source language, and the Nawait text is written in Perso-Arabic script, reflecting the commonly used writing practice among native speakers.

1.5 SUMMARY OF NAWAIT PARALLEL CORPUS

The Kudubi/Kudumbi parallel text corpus includes English and 146 other mother tongues of India. 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, 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 Nawait which are systematically structured based on 159 grammatical categories. Nawait has 29,110 word count and 145,844 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] Ethnologue: Languages of the World; Glottolog; G.A. Grierson – Linguistic Survey of India; Colin P. Masica – The Indo-Aryan Languages (1991).
[3] https://langlex.com/cens/MTProfile.php?mtname=Nawait+
[4] https://language.census.gov.in/showMTSIdata#
[5] (From The Nawayaths of Kanara: Victor D'Souza, 1955)
[6] https://twocircles.net/2017jun24/411956.html
[7] https://joshuaproject.net/people_groups/17767/IN
[8] https://bhatkallys.com/news/read/nawayaths-now-have-their-missing-letters-on-phonetic-keyboard/