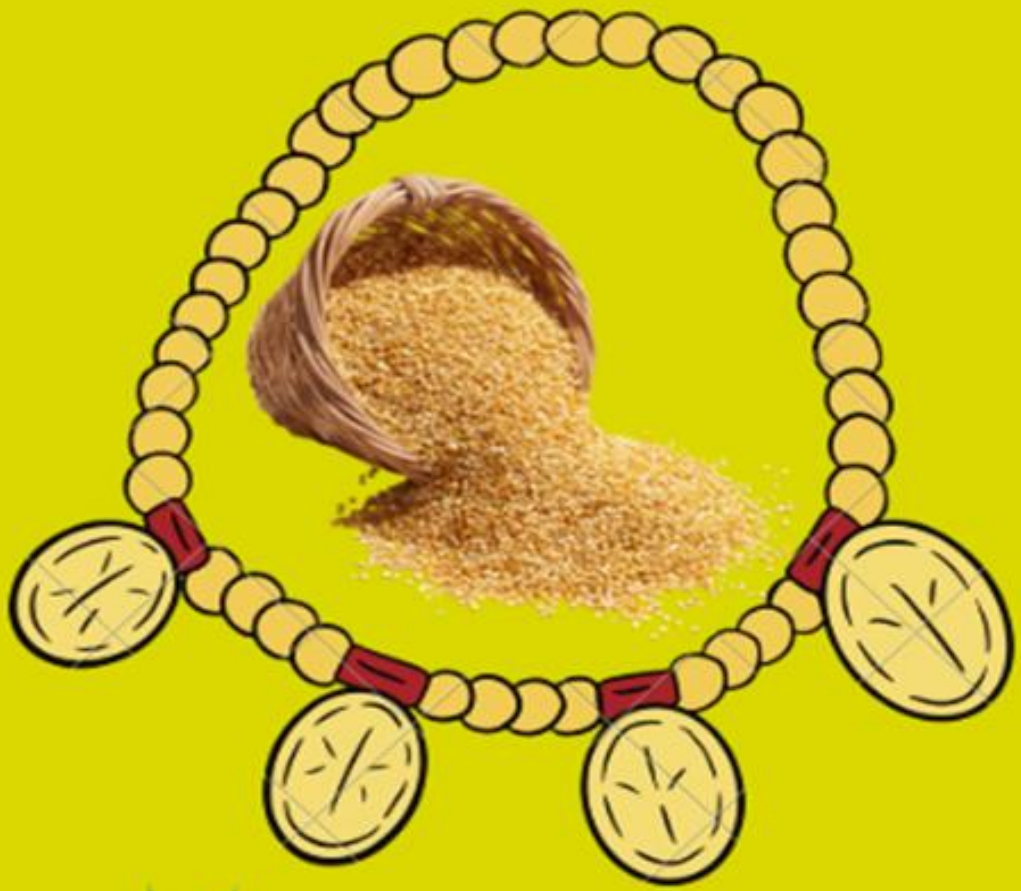
Yerava Parallel Text Corpus: Linguistic Features and Structures



YERAVA 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. Saritha S.L

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

Yerava Parallel Text Corpus: Linguistic Features and Structures

Authors: Dr. Saritha S. L., Dr. Rejitha K. S., Dr. Narayan Choudhary, Prof. Shailendra Mohan
Contributors: Poovaiah M. P., Accamma M. S.

e-ISBN: 978-81-69175-40-1
CIIL Publication No.: 1713

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: Saritha.S.L

Table of Contents

YERAVA PARALLEL TEXT CORPUS

Saritha. S. L.

1.1 LDC-IL PARALLEL TEXT CORPUS

LDC-IL have 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 YERAVA: A BRIEF INTRODUCTION

According to 2011 Census approximately 24,574 speakers in Karnataka (Total Ravula speakers, including those in Kerala, were recorded at roughly 26,563). According to the Current Estimates (2024–2026) based on the decadal growth of the Scheduled Tribe (ST) population in Karnataka, the number of people identifying as Yerava is estimated to be between 31,000 and 35,000. While

the community size is increasing, researchers estimate that only about 20,000 to 22,000 are "active" speakers who use Yerava as their primary daily language.

In the 2011 Census of India, the Yerava language is classified as a mother tongue under the Malayalam language group. It belongs to the South Dravidian sub-family. It shares a high degree of lexical similarity with Malayalam but also incorporates elements of Kannada and Kodava due to geographic proximity. In Karnataka, the speakers are largely found in the Virajpet and Somwarpet taluks of Kodagu. In Kerala, they are mostly situated in the Mananthavady region of Wayanad. Yerava is traditionally an unwritten language. The Kannada script is the standard medium for writing the language. Children from the Yerava community attend schools where Kannada is the medium of instruction. It is rich cultural heritage, including folk songs, oral histories, and community rituals, is passed down through spoken word rather than text. Historically, the Yerava people were agricultural laborers. Today, there are increasing efforts in computational linguistics and documentation to preserve the language and create parallel corpora to bridge it with more widely spoken languages.

While the overall literacy rate in Karnataka was approximately 75.37% in 2011, the literacy rate among the Yerava and other forest-based tribes in Kodagu was significantly lower. Estimates suggest literacy rates for the Yerava tribe hovered around 40-50%, trailing behind the state average. There is a notable disparity between genders. Literacy among Yerava men is higher than among women, though government initiatives in the last decade have begun to narrow this gap. Children grow up speaking Yerava at home but are taught in Kannada from Grade 1. This can lead to lower comprehension in early years. Because Yerava has no script, there are no textbooks or stories in their mother tongue to help bridge the gap to Kannada.

The Lack of Technical Equivalents is a major challenge when translating a sentence from English to Yerava. For example, when translating the sentence "I am going to the laboratory," the translator faces a major challenge. There is no word for "Laboratory" in Yerava. In this case the translator would likely use the Kannada word “Prayōgālaya” (ಪ್ರಯೋಗಾಲಯ) or just say the English word "Lab" with a Yerava accent.

1.4 YERAVA PARALLEL CORPUS

The most significant characteristic of Yerava is its lack of a native script. It is an oral language, preserved through the "Baane" (folk songs) and oral histories passed down by elders. In modern contexts, this poses a challenge for preservation. Today, the Kannada script is used as the primary medium for documentation, education, and administration within the community. This reliance on a "bridge script" means that some of the language's unique phonetic nuances specifically its distinct nasal and retroflex sounds risk being standardized into the phonology of the dominant regional language.

The Yerava language is more than just a means of communication; it is a repository of the community's history, forest-based wisdom, and tribal identity. While the 2011 Census provides a numerical snapshot of its existence, the true story of Yerava lies in its struggle to survive in a rapidly modernizing world. Preserving this language requires more than just counting speakers; it requires the creation of digital resources, like parallel corpora, and a renewed effort to document its oral traditions before they are silenced by the passage of time. Yerava is structurally and lexically influenced by Malayalam, Kannada, and Kodava.

Creating the Yerava Parallel Corpus, a digital database of sentences aligned between Yerava and a major language such as English and Kannada, is no ordinary task. Since it is entirely oral, the sentences must be translated into Kannada scripts. When done this way, it not only loses its linguistic beauty but also its naturalness. Most Indian mother tongues recorded in the Census are low-resource (meaning they have little digital data). Yerava is a representative model for these languages. 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 YERAVA PARALLEL CORPUS

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

There are 5,332 sentences/phrases in Yerava which are systematically structured based on 159 grammatical categories. Yerava has 24453 word count and 153083 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. Census of India.,(2011)., Kerala, data highlight; the scheduled tribes., http://www.censusindia.gov.in/tables-published/scst/dh-st-kerala.pdf.
3. Census of India.,(2011).,basic datasheet, district Wayanad , Kerala., http://www.censusindia.gov.in/dis-file/datasheet-3203.pdf.
4. Mallikarjun B. 1993. A Descriptive Analysis of Yerava. Central Institute of Indian Languages, Mysore.
5. Mallikarjun B. 2019. Linguistic Ecology of Karnataka. Language in India. Vol, 19:7.
6. “Kritads kaipusthakam”., (2017-2018).