



IRULA/IRULAR MOZHI PARALLEL TEXT CORPUS: Linguistic Features and Structures

IRULA/IRULAR MOZHI 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:

Dr. Amudha R

Dr. Rejitha K. S

Dr. Narayan Choudhary

Prof. Shailendra Mohan

**Linguistic Data Consortium for Indian Languages
Central Institute of Indian Language
Mysuru, India-570006**

भारतीयभाषासंस्थान

उच्चतरशिक्षाविभाग, शिक्षामंत्रालय, भारतसरकार, मानसगंगोत्री, हुन्सूररोड, 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

Irula/Irular Mozhi Parallel Text Corpus: Linguistic Features and Structures

Authors: Dr. Amudha R., Dr. Rejitha K. S., Dr. Narayan Choudhary, Prof. Shailendra Mohan

Contributors: Dr.Suresh P., Srinivasan K

e-ISBN: 978-81-69099-11-0

CIIL Publication No.: 1644

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 ManjuBhargavi, Unit-in-Charge, Printing Press & Publication Unit

M.N. Chandrashekar, Compositor, Printing Press & Publication Unit

Cover design: Amudha. R

Table of Contents

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

,

IRULA/IRULAR MOZHI PARALLEL TEXT CORPUS

Amudha. R

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 in spoken form. They don't have script of their own. Following the requirements of the NEP-2020, LDC-IL initiated to collect sentences from different mother tongues that could give an idea about the structure of these 270 mother tongues.

A parallel corpus is a collection of texts in two or more languages that are carefully aligned so each segment matches its translation. 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 categories like Adjective, Adverb, Compounding, Interrogative, Imperative, Speech type, Statement, Structural Question, Coordination, and Subordination etc. Moreover, it includes linguistic features such as Adposition, Anaphora, Derivational Morphology, Inflectional Morphology and language specific features like Emphasis, Equative, Possession, Reciprocal, Reflexive etc. For more details, refer to [LDC-IL Corpus Insights](#).

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 foundation 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 orange economy. Local languages often contain indigenous knowledge about agriculture, medicine,

environment, and culture. Bringing them into digital systems safeguards and shares this knowledge.

Supporting Low-resource languages makes AI more inclusive, intelligent, and representative of linguistic diversity. In India, where hundreds of mother tongues coexist, 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 use in their region. Some mother tongues may use multiple scripts or recently developed 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 IRULA/IRULAR MOZHI: A BRIEF INTRODUCTION

Irula (also referred to as Irular or Irulaa) is a member of the Dravidian language family, spoken by the Irula tribal community of southern India. It belongs to the Southern Dravidian subgroup and shares close affinities with Tamil, Kannada, and related languages. Traditionally, Irula is represented using the Tamil script.

The language is primarily spoken in the Nilgiri Hills and adjoining regions of Tamil Nadu, with smaller speaker populations located in Kerala and Karnataka. According to the 2011 Indian Census, approximately 11,870 individuals identified Irula as their mother tongue. However, this figure likely underrepresents actual usage, as many Irula speakers report dominant regional languages such as Tamil or Malayalam. While the Irula ethnic population exceeds 200,000, a substantial proportion of the community uses other regional languages in everyday communication.

The Irula people are a forest-dwelling tribal community traditionally engaged in hunting-gathering, with deep ecological knowledge. The name “Irula” means “people of the forest” or “people of the dark.” Research, particularly by Kamil Zvelebil, shows that Irula is an independent Dravidian language, preserving archaic phonological and lexical features from Proto-Dravidian.

Their culture is strongly oral, including songs, folklore, storytelling, and ecological knowledge, with skills in snake catching, traditional medicine, music, and dance. Beliefs combine animistic traditions with regional religious influences. Irula belongs to the Southern Dravidian branch, closely related to Tamil, but influenced by Kannada and Malayalam. Its phonology retains retroflex consonants and older vowel patterns. The language is agglutinative, marking tense, case, number, gender, and person through suffixes, and follows a Subject–Object–Verb order with postpositions. Verbs encode tense, aspect, mood, person, number, and sometimes gender, with complex predicates often using light or serial verbs.

The language is vulnerable, as younger speakers shift to dominant regional languages like Tamil, Malayalam, and Kannada. Despite this, Irula provides important insights into Dravidian grammar, language contact, morphosyntactic variation, and low-resource computational challenges.

1.4 IRULA/IRULAR MOZHI PARALLEL CORPUS

Irula/Irular Mozhi is a highly agglutinative language with rich morphophonemic features. It exhibits complex morphophonemic alternations, extensive case marking, productive compounding, and relatively flexible word order. Irula/Irular Mozhi is written in Irula/Irular Mozhi script derived from the Brahmi tradition. It has an extensive set of consonant and vowel symbols that precisely capture its phonemic distinctions. It generally follows the Subject-Object-Verb pattern but it is a linguistically rich and structurally complex language. These features contribute to Irula/Irular Mozhi's uniqueness among the languages of India and highlight its importance in the study of linguistics.

Irula/Irular Mozhi 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 IRULA/IRULAR MOZHI PARALLEL CORPUS

The Irula/Irular Mozhi 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, Chirri, Chungli, Churahi, Coorgi/Kodagu, Deori, Dhundhari, Dimasa, Dogri, Gangte, Garhwali, Garo, Gojri/Gujjari/Gujar, Gujarati, Gujar, Halabi, Handuri, Hara/Harauti, Haryanvi, Hindi Multani, Hindi, 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, Mongsan, 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, Sambalpur, Sangtam, Sanskrit, Santali, Saurashtra/Saurashtri, Sema, Shina, Sindhi, Sirmauri, Sugali, Surjapuri, Talgalo, Tangkhul, Tamil, Telugu, Thado, Tibetan, Tikhir, Tripuri, Tulu, Urdu, Vaiphei, Wagdi, Wancho, Yerava, Yerukala/Yerukula, Yimchungre, Zeliang, Zemi, Zou.

There are 5,332 sentences/phrases in Irula/Irular Mozhi which are systematically structured based on 159 grammatical categories. Irula/Irular Mozhi has 22,909 word count and 164,665 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.