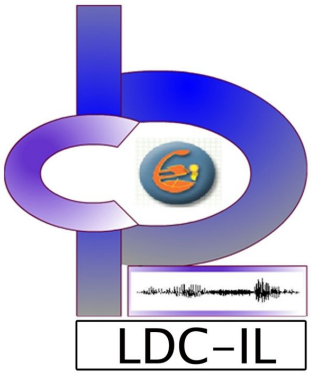
Urdu Parallel Text Corpus: Linguistic Features and Structures

URDU 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

Urdu Parallel Text Corpus: Linguistic Features and Structures

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

Contributors: Dr. Syed Majid Ali, Dr. Tasneem Zahera Haidry, Dr. Md. Farhan

e-ISBN: 978-81-69175-14-2
CIIL Publication No.:1717

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

URDU 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 URDU: A BRIEF INTRODUCTION

Urdu is one of the most prominent languages of South Asia, possessing deep cultural, literary, and historical significance. It originated in the Indian subcontinent during the medieval period through continuous interaction between local Indo-Aryan vernaculars and the Persian, Arabic, and Turkish languages introduced by traders, scholars, and rulers. The word “Urdu” is derived from the Turkish term Ordu, meaning “camp” or “army,” reflecting its early role as a contact language among linguistically diverse communities. Gradually, it developed from a practical lingua franca into a highly refined literary language with a rich intellectual and aesthetic tradition. Linguistically, Urdu belongs to the Indo-Aryan branch of the Indo-European language family and shares a common grammatical structure and basic vocabulary with Hindi, making the two mutually intelligible at the colloquial level. However, Urdu distinguishes itself through its Perso-Arabic script, written in the elegant Nastaliq style, and its extensive incorporation of Persian and Arabic vocabulary, which together shape its distinct cultural and literary identity.

Urdu has played a central role in shaping the literary heritage of South Asia, particularly in poetry and prose. Classical poetic genres such as the ghazal, nazm, marsiya, and qasida reached remarkable heights in Urdu, with eminent poets like Mir Taqi Mir, Mirza Ghalib, and Allama Iqbal elevating the language to exceptional artistic refinement. During the nineteenth and twentieth centuries, Urdu prose also flourished, producing significant works in fiction,

literary criticism, essays, and drama that engaged with themes of social reform, modernity, and political transformation. In contemporary India, Urdu is recognized as one of the Scheduled Languages under the Constitution and continues to be used in education, media, journalism, and cultural expression. It symbolizes the composite and syncretic heritage of the subcontinent and remains a dynamic, evolving language that preserves its classical elegance while adapting to modern intellectual and cultural contexts.

1.4 URDU PARALLEL CORPUS

Urdu is a morphologically rich language with complex grammatical and phonological features. Although it is less agglutinative than some Dravidian languages, it displays significant morphological variation through inflection, derivation, gender agreement, case marking, and the use of postpositions. Urdu also demonstrates notable morphophonemic alternations, particularly in plural formation, verb conjugation, and loanword adaptation from Persian and Arabic. The language generally follows a Subject–Object–Verb (SOV) word order, yet it allows a certain degree of flexibility for emphasis and stylistic variation. Urdu is written in the Perso-Arabic script, specifically in the Nastaliq style, which is known for its aesthetic elegance and calligraphic complexity. The script contains a rich inventory of consonants and vowel markers that represent subtle phonemic distinctions. These linguistic and orthographic features contribute to Urdu's structural depth and make it an important language for linguistic analysis in the Indian context.

Urdu also serves as a valuable resource for developing parallel text corpora with other Indian mother tongues and major languages such as Hindi and English. An Urdu parallel corpus can significantly support research in areas such as syntactic divergence, morphological variation, lexical borrowing, code-mixing, and script differences. Given India's multilingual environment, such a corpus is particularly useful for examining cross-linguistic influence and translation patterns. Furthermore, it can aid in the development of natural language processing tools, machine translation systems, and bilingual lexicons. Thus, Urdu parallel corpora hold substantial importance for both theoretical linguistic research and practical language technology applications.

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

1.5 SUMMARY OF URDU PARALLEL CORPUS

The Urdu parallel text corpus including English and 146 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, 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, Vaiphei, Wagdi, Wancho, Yerava, Yerukala/Yerukula, Yimchungre, Zeliang, Zemi, Zou.

There are 5,332 sentences/phrases in Urdu which are systematically structured based on 159 grammatical categories. Urdu has 34329 word count and 147375 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 Regional Language Distribution.