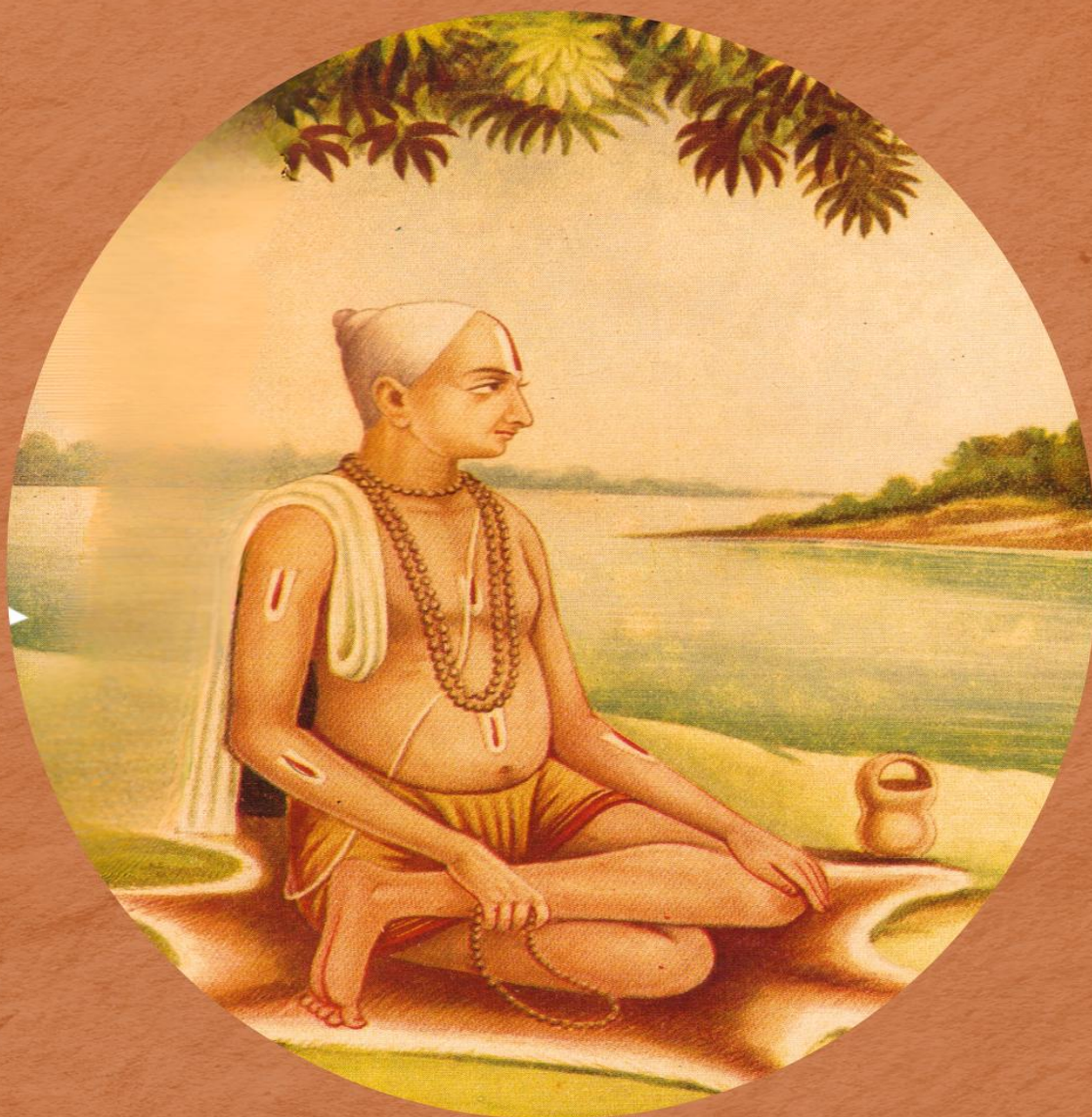


AWADHI PARALLEL TEXT CORPUS: LINGUISTIC FEATURES AND STRUCTURES



AWADHI 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. Satyaendra Kumar Awasthi

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, मैसूर-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

Awadhi Parallel Text Corpus: Linguistic Features and Structures

Authors: Dr. Satyaendra Kumar Awasthi, Dr. Rejitha K. S., Dr. Narayan Choudhary, Prof. Shailendra Mohan

Contributors: Dr. Sumedha Shukla, Vinod Kumar Tiwari

e-ISBN: 978-81-69099-42-4

CIIL Publication No.: 1681

First published: AD 2026 March :: Phalgunā 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: Teena Verma

Table of Contents

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

AWADHI PARALLEL TEXT CORPUS

Dr. Satyaendra Kumar Awasthi

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](#)..

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 AWADHI: A BRIEF INTRODUCTION

Awadhi is an Indo-Aryan language that is categorized as a dialect within the Eastern Hindi group. In fact, all dialects are independent languages with their own linguistic structure, culture, literature, and history. Therefore, Awadhi is also an independent language with its own literature and history. In Uttar Pradesh, districts like Amethi, Allahabad, Ayodhya, Bahraich, Barabanki, Gonda, Lucknow, Pratapgarh, Rae Bareilly, Sitapur, Sultanpur, Unnao etc. are mentioned as the Awadhi region. Along the northern boundary of Uttar Pradesh adjoining Nepal, Awadhi is widely spoken and used in literary compositions. Apart from the Terai region of Nepal, sizeable Awadhi-speaking communities are also found in countries such as Fiji and Suriname, and in this way Awadhi has exerted influence on Hindi as well. According to the Census of India 2011, about 3,850,906 people reported Awadhi as their mother tongue.

Awadhi is one of the major languages of the Eastern Hindi group. From a literary point of view, Awadhi is a rich language which has a very vast history. The celebrated epic ‘Ramcharitmanas’ composed by Tulsidas stands as an outstanding example of Awadhi literature. Similarly, the renowned epic ‘Padmavat’ written by Malik Muhammad Jayasi was composed in this language.

Awadhi is not only a literary language but also a living language of folk culture. Wedding songs, Sohar, Kajri, Birha, Phaag, proverbs, idioms, and folk tales are still prevalent in rural life. For example, Sohar (birthday song) and Phaag (Holi song) reflect the most melodious form of Awadhi. The simplicity of emotions and naturalness of expression in folk poetry are its distinctive hallmarks. Among modern literary figures, Chandrabhushan Trivedi ‘Ramai Kaka’, Banshidhar Shukla, and Balabhadra Prasad Dixit are held in high esteem for their contribution to Awadhi literature.

Awadhi possesses a rich tradition of ethical and folk literature. The didactic verses of Ghagh are still heard in rural life, and Rahim is regarded as an important poet in this tradition. Its folk literature includes labour songs, songs associated with different communities, festival songs, and compositions connected with women’s customs and rituals.

1.4 AWADHI PARALLEL CORPUS

Awadhi is a well-formed Indo-Aryan language with a clear grammatical structure and good linguistic features. Awadhi is a folk-based language with a well-organized grammar, featuring clear gender and number forms, case system, a rich pronoun system, and a simple SOV sentence structure. Despite its close relationship to Hindi, it maintains its distinct identity through its independent form, phonetics, and popular usage.

Here are some features of Awadhi-

- All vowels are nasalized in Awadhi.
- There are only two genders in Awadhi—masculine and feminine.
- Sound change or suffix is used to create the feminine form.
Example. gəḍḍḥa (donkey)→ gəḍḍḥi: (mare or female donkey), kuṭṭəva (dog) → kuṭṭija (bitch), gəgəra (pot)→ gəgəri: (tiny pot)
- Gender affects both pronouns and adjectives—
Example. həmar gḥəru (our house)/həmari məciḥa (our stool) , bəṛa lərika (elder son) /bəṛi: biṭəwa (elder daughter)
- Verbs are not affected by gender with animate objects —
Example. ram jaṭi həj (Ram goes), si:ṭa jaṭi həj (Sita goes),
- Sometimes the gender of verbs also changes according to inanimate nouns —
Example. kurəsi: pəri: həj (the chair is lying), ṭəkḥəṭu pəra həj (the bed is lying)
- In Awadhi, the primary marker for forming the plural is the suffix **-n** (न). It is added to the singular noun to indicate plurality.
Example: lərika (boy) → lərikən (boys)
- The palatal /sh/ and the cerebral /sh/ are not pronounced in Awadhi; instead, the dental /s/ is used. Sometimes, the /kḥ/ is also used in place of the cerebral /sh/. For example- bḥaṣa (language) - bḥasa / bḥakḥa, wiṣ (poison) - bis / bikḥ.
- Conversion of /v/ to /b/: If /v/ is at the beginning of words, it becomes /b/: e.g., van (forest) - ban, vər (groom) – bər.

Awadhi 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 AWADHI PARALLEL CORPUS

The development of Awadhi Parallel Corpus, Hindi was used as the source language which was translated into Awadhi. This Awadhi parallel text corpus connected with English and 146 mother tongues of India. Those mother tongues are Anal, Angami, Apatani, Are, Assamese, 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, 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, Yerava, Yerukala/Yerukula, Yimchungre, Zeliang, Zemi, Zou.

There are 5,332 sentences/phrases in Awadhi which are systematically structured based on 159 grammatical categories. Awadhi has 32,081 word count and 141,965 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] LangLex. (n.d.). Awadhi: Mother tongue profile (2011 Census report). Retrieved October 24, 2023, from <https://langlex.com/cens/MTPProfile.php?mtname=Awadhi>
- [2] Office of the Registrar General & Census Commissioner, India. (2018). C-16: Population by mother tongue (Table). Ministry of Home Affairs, Government of India. <https://language.census.gov.in/showMTSIdata>
- [3] Verma, D. (1954). Hindi bhasha ka itihās [History of the Hindi language]. Hindustani Academy.
- [4] Rejitha K. S. and Narayan Kumar Choudhary. (ed.). 2025. LDC-IL Corpus Insights. Central Institute of Indian Languages, Mysore. 978-93-48633-33-0.