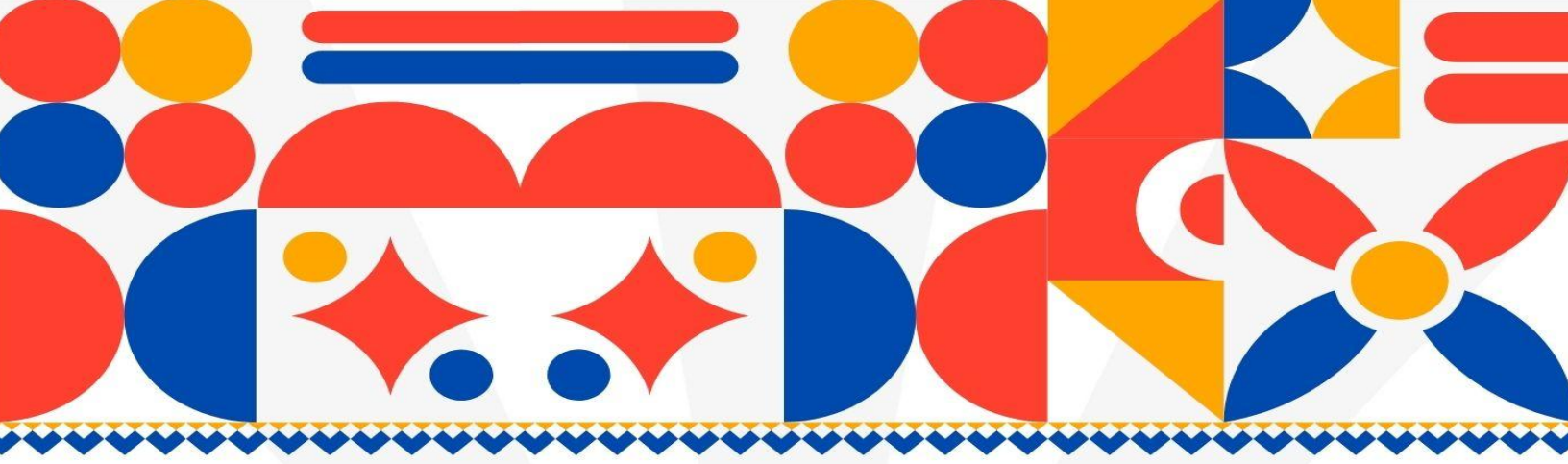




Surjapuri Parallel Text Corpus: Linguistic Features and Structures



SURJAPURI 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:

Ankita Tiwari

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

Surjapuri Parallel Text Corpus: Linguistic Features and Structures

Authors: Ankita Tiwari, Dr. Rejitha K. S., Dr. Narayan Choudhary, Prof. Shailendra Mohan

Contributors: Sanowar Jahan, Zoofishan Khanam and Khushnuma Begum

e-ISBN: 978-81-69175-27-2

CIIL Publication No.: 1727

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: Ankita Tiwari

TABLE OF CONTENTS

SURJAPURI PARALLEL TEXT CORPUS	5
1.1 LDC-IL PARALLEL TEXT CORPUS.....	5
1.2 CHALLENGES IN CREATING A PARALLEL CORPUS.....	5
1.3 SURJAPURI: A BRIEF INTRODUCTION	6
1.4 SURJAPURI: A BRIEF LINGUISTIC OVERVIEW	8
1.5 SUMMARY OF SURJAPURI PARALLEL CORPUS	8
1.6 REFERENCE	10

SURJAPURI PARALLEL TEXT CORPUS

Ankita Tiwari

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 SURJAPURI: A BRIEF INTRODUCTION

Surjapuri (ISO 639-3: sjp)¹ is an Indo-Aryan mother tongue grouped under Hindi in official Census of India records. It is primarily spoken in Kishanganj, Bihar and extends to neighboring regions such as Katihar, Purnia, and Araria, as well as in parts of the Mithila region of Bihar. The largest concentration of Surjapuri speakers is located in Kishanganj, Thakurganj, Pothia, Bahadurganj, Dighalbank, Kochadhaman, Barsoi, and Balrampur. It is alternatively known as Kishanganjia due to its association with the Kishanganj region. Surjapuri shares several structural features and vocabulary items with neighboring languages such as Assamese, Bangla, Bhojpuri, and Maithili. The effect of Bangla and Maithili is so strong that according to Grierson, "the dialect in question is, in the main, Bengali with an admixture of Hindi, but it is written in the Kaithi character of Bihar, in which Hindi and not Bengali is written." Surjapuri was written in the Kaithi script, which was widely used in Bihar for administrative and literary purposes. However, Kaithi is no longer in common use, and it has been replaced by the Devanagari script in contemporary writing practices.²

Grierson reported this Mother Tongue as Siriapuria and Kishangunjia during his survey of LSI. Interestingly, in 1961 Census Siripuria was replaced by 'Surjapuri' (26,947). The speakers of Surjapuri are well versed in Hindi and Urdu languages. Historically, it is said to be closely related to Kamta and Rajbanshi languages (2006 M. Toulmin).

¹ <https://iso639-3.sil.org/code/sjp>

² P Perumal Samy, A Linguistic Description of Surjapuri Mother Tongue Spoken in Bihar State

According to the most recent 2011, Census of India data, Surjapuri is classified as a mother tongue under the Hindi language category. A total of 22,56,228 individuals reported Surjapuri as their mother tongue in the Census.³

1.4 SURJAPURI: A BRIEF LINGUISTIC OVERVIEW

Surjapuri possesses a pronominal system that marks distinctions of person, number, and case, while gender marking is limited and mainly confined to the third person singular. Personal pronouns appear in the nominative case when functioning as subjects and shift to oblique forms in other grammatical roles such as accusative, genitive, and locative, usually in combination with postpositions. Unlike many other Indo-Aryan languages, Surjapuri does not employ honorific distinctions in its pronoun forms.

From a morphological perspective, Surjapuri demonstrates agglutinative characteristics, primarily through the addition of suffixes to indicate grammatical categories such as case, tense, aspect, and number. Inflectional morphology plays a more significant role than derivational processes. New words are formed mainly through suffixation, prefixation, and compounding.

Surjapuri follows a Subject–Object–Verb (SOV) word order. Grammatical relations, especially spatial and relational meanings, are expressed through postpositions, which are often attached to genitive-marked forms. Verbs generally agree with the subject in person and number.⁴

1.5 SUMMARY OF SURJAPURI PARALLEL CORPUS

The Surjapuri Parallel Text Corpus has been developed using Hindi as the source language. This Corpus connected with English and 146 mother tongues of India. Those mother tongues are Anal, Angami, Apatani, Are, Assamese, Awadhi, Bagri Rajasthani, Bagri, Bagheli/Baghel Khandi, 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, Gujari, Halabi, Handuri, Hara/Harauti, Haryanvi, Hindi

³ <https://censusindia.gov.in/nada/index.php/catalog/42561>

⁴ https://grokipedia.com/page/Surjapuri_language

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, Talgalo, Tamil, Tangkhul, Telugu, Thado, Tibetan, Tikhir, Tripuri, Tulu, Urdu, Vaiphei, Wagdi, Wancho, Yerava, Yerukala/Yerukula, Yimchungre, Zeliang, Zemi, Zou.

The Surjapuri dataset comprises 5,332 sentences/phrases, systematically structured according to 159 grammatical categories. The Surjapuri portion of the corpus contains 30,399 words and a total of 1,41,982 characters. Moreover, the corpus is aligned with 146 Indian mother tongues in addition to English. Altogether, the multilingual parallel corpus encompasses 4,404,845 words (over 4.4 million tokens) and 23,374,289 characters (approximately 23.3 million).

1.6 REFERENCE

- [1] S. P. Srivastava & Perumalsamy. (2021), Surjapuri (A Linguistic Description of Surjapuri Mother Tongue Spoken in Bihar State By P Perumal Samy). Language Division, Office of the Registrar General India
- [2] <https://censusindia.gov.in/nada/index.php/catalog/42561>
- [3] <https://iso639-3.sil.org/code/sjp>
- [4] https://grokipedia.com/page/Surjapuri_language
- [5] Rejitha K. S. and Narayan Kumar Choudhary. (ed.). (2025). LDC-IL Corpus Insights. CIIL, Mysore. 978-93- 48633-33-0.