The Mother Tongue Parallel Text Corpus of India: Design, Development, and Applications

Dr. Narayan Kumar Choudhary[1], Dr. Rejitha K. S.[2], Rajesha N.[3], Anand Jain[4], StephenFernandes[5], Writtik Bhattacharya[6]

[1]Central Institute of Indian Languages, Mysore, India
nchoudhary.ciil@gmail.com
[2]Central Institute of Indian Languages, Mysore, India
ksrejitha@gmail.com
[3]Central Institute of Indian Languages, Mysore, India
n.rajesha@yahoo.co.in
[4]Crisant Technologies, Mysore, India
anand@crisant.com
[5]Central Institute of Indian Languages, Mysore, India
stephen.fern13@gmail.com
[6]Central Institute of Indian Languages, Mysore, India
rittik.2000.12@gmail.com

Abstract

Parallel corpora serve as the fundamental resource for cross-linguistic studies, translation systems, and language teaching. This paper documents the corpus creation process, framework design, challenges, and mitigation plan while building Mother Tongue Parallel Text Corpus of India published by the Linguistic Data Consortium for Indian Languages (LDC-IL). The corpus covers 147 mother tongues and English, comprising 5,332 sentences that are systematically categorized sentences across 159 grammatical categories. The corpus provides an unprecedented foundation for linguistic research, language technology, and cultural studies in India.

Keywords: Parallel Corpus, Less resourced languages, Mother tongue, LDC-IL

1. Introduction

A parallel corpus is a collection of texts in a source language that are aligned with their corresponding translations into one or more target languages. It is constructed by selecting sentences from one source language and translating them into the target languages. In such a corpus, the source language influence can be seen in all language levels like vocabulary, syntax, and even in discourse of the target language texts. It can be subtle or striking depending on how well the translation aligns to the source. The translators try to preserve the meaning of the source while maintaining the naturalness in the target language. Therefore, translations may be influenced by the grammatical or stylistic features of the source language. Sometimes, the translation may contain phrases or idiomatic expressions that are artificial in the target language but common in the source language. These variations are not usually classified as errors or mistakes in the usual sense but they are considered as cross-linguistic transfer. These differences are valuable linguistic evidence of how languages interact in translation processes and it will help in comparative study.

The parallel corpus can be used as a fundamental resource in linguistic and computational domains. It will be useful in the fields of lexicography, translation studies, grammatical analysis, comparative study, and machine translation. It can be used in second language learning, where the learners get exposure to structural variations across languages. Parallel corpus helps to check a word is actually translated in a context, which is different from the dictionary. It is easy to monitor the syntactic construction and ways of translating them into different languages. It is a fundamental resource for the development of automated translation systems. It works as an authentic data in language learning and teaching. It serves as an excellent source in contrastive language study.

2. Linguistic Diversity in India

India is a linguistically diverse landmass, with varied geographical contexts and cultures. The Indian languages primarily belong to one of the four major language families, namely Indo-Aryan, Dravidian, Austro-Asiatic, and Tibeto-Burman. The 2011 Census of India records 121 major languages and 270 mother tongues.

The multilingual environment is a characteristic of Indian society. There are many factors behind the linguistic diversity in India. Since ancient times, many communities have invaded and migrated to India, making India a linguistically diverse country. Geographically separated communities became isolated from other communities and developed distinct regional languages. India's education policy supports multilingualism so children are proficient in their regional language along with the official languages. The Indian Constitution, however, does not declare any language as a national language, recognizing the importance of linguistic preservation and inclusivity.

The written comprehensive textual resources or literature are limited in most of these mother tongues. The Linguistic Data Consortium of Indian Languages (LDC-IL) is a national repository of linguistic data. LDC-IL initiated the creation of the Mother Tongue Parallel Text Corpus of 270 mother tongues. It can provide fundamental linguistic features of these languages.

Narayan Choudhary, Rejitha K. S., Rajesha N., Anand Jain, Stephen Fernandes, Writtik Bhattacharya, *The Mother Tongue Parallel Text Corpus of India: Design, Development, and Applications*. In: Z. Vetulani, P. Paroubek, I. Kardava (Eds.), Human Language Technologies as a Challenge for Computer Science and Linguistics – 2025, pp. 19–22. Poznań: Adam Mickiewicz University Press, 2025. © Authors, 2025. https://doi.org/10.14746/amup.9788323245414.3. Open Access chapter, distributed under the terms of the CC licence (BY-NC-ND, https://creativecommons.org/licenses/by-nc-nd/4.0/).

3. Methodology

The project began with a preliminary exercise of building a structure by extracting sentences from descriptive grammar books of Indian languages to accommodate the comprehensive linguistic features present in these languages, as well as some common phrases that are shared among Indian languages. For example, some languages do not have the negative imperative form, clefting, etc. But such structures occur naturally in the source language and when they are translated into the target language, those structures seem unnatural. Additionally, some words or senses do not exist in some languages, and in such cases words with near-by meanings in the target language or their explanations must be provided.

The corpus framework was constructed based on (Comrie & Smith, 1977), supported by the insights of (Abbi, 2001; Sivashanmugam & Thayalan, 2012; Bhattacharya et al., 2016). A total of 5,332 sentences were created that are categorized into 159 grammatical categories, based on their linguistic features.

The categories cover a wide range of phenomena; morphology, which includes tense, aspect, mood, voice case, gender markers; syntax, which includes word order, interrogatives, compounding, sentence types; lexicon, which includes culturally specific items with or without equivalents across languages. Rejitha K. S. and Narayan Kumar Choudhary (2025) document in detail the grammatical categories that should be considered when building the framework of the dataset.

The corpus is constructed in such a way that it can provide similarities and differences between languages. There are interrelations and conflicts at all levels of language. If sentences are constructed by including a syntactic category, an attempt has been made to include all its subcategories. For example, if the category is interrogative, all types of interrogative sentences such as yes-no question, Wh-question, tag question, alternative question, indirect question, etc. have been included.

The 5332 sentences were conceptualized in Indian languages but it is created in English. They were then translated into 22 scheduled languages of India. Consequently, these translations can serve as source sentences for translators of any non-scheduled language or mother tongue, who may be proficient in one of the scheduled languages but not in English, thereby facilitating corpus creation.

LDC-IL empanelled freelance translators and reviewers after conducting linguistic proficiency tests. Until two freelancers are empaneled for a language, work on that language was not taken up, as at least two freelancers are needed for translation and mutual review of each other's work. For easy management the languages were clustered based on linguistic and geographic proximities. 22 Linguists from LDC-IL managed these language clusters, preferably considering their own proficiency in a common language of communication with the target language freelancers. The online platform Trankit[1] is used to develop the mother tongue parallel text corpus. Trankit is a platform to carry out the translation straightforwardly. It is a simple and user friendly interface.

4. Results

The first phase of corpus creation has yielded coverage of the 147 Indian languages along with English. There are 5,332 sentences in each language and a total of 4,404,845 (~4.4 million words) tokens across the corpus. 413 translators/reviewers were involved in the process.

The Indian languages covered are 1.Anal, 2.Angami, 3.Apatani, 4.Are, 5.Assamese, 6.Awadhi, 7.Bagheli/Baghel Khandi, 8.Bagri Rajasthani, 9.Bagri, 10.Balti, 11.Bengali, 12.Bhadrawahi, 13.Bharmauri/Gaddi, 14.Bhojpuri, 15.Bilaspuri Kahluri, 16.Bodo/Boro, 17.Brajbhasha, 18.Bundeli/Bundelkhandi, 19.Chakru/Chokri, 20.Chambeali/Chamrali, 21.Chang, 22.Chhattisgarhi, 23.Chirr, 24.Chungli, 25.Churahi, 26.Coorgi/Kodagu, 27.Deori, 28.Dhundhari, 29.Dimasa, 30.Dogri, 31.Gangte, 32.Garhwali, 33.Garo, 34.Gujarati, 35.Gujari, 36.Gujjari/Gujar/Gojri, 37.Halabi, 38.Handuri, 39.Hara/Harauti, 40.Haryanvi, 41.Hindi Multani, 42.Hindi, 43.Irula/IrularMozhi, 44.Kabui, 45.Kachchhi, 46.Kangri, 47.Kannada, 48.Karbi/Mikir, 49.Kashmiri, 50.Khandeshi, 51.KhariBoli, 52.Khasi, 53.Khezha, 54.Khiemnungan, 55.Khortha/Khotta, 56.Kisan, 57.Kodava, 58.Kokbarak, 59.Kolami, 60.Koli, 61.Kom, 62.Konda, 63.Konkani, 64.Konyak, 65.Koya, 66.Kudubi/Kudumbi, 67.Kuki, 68.KurmaliThar, 69.Ladakhi, 70.Lepcha, 71.Liangmei, 72.Limbu, 73.Lotha, 74.Lyngngam, 75.Magadhi/Magahi, 76.Maithili, 77.Malayalam, 78.Malvi, 79.Manipuri, 80.Mao, 81.Mara, 82.Maram, 83.Marathi, 84.Maring, 85.Mech/Mechhia, 86.Mewari, 87.Mewati, 88.Miri/Mishing, 89.Mishmi, 90.Mizo, 91.Mongsen, 92.Monpa, 93.Mundari, 94.Muwasi, 95.Nawait, 96.Nepali, 97.Nimadi, 98.Nissi, 99.Nocte, 100.Odia, 101.Pahari, 102.Paite, 103.Palmuha, 104.Pania, 105.Paola, 106.Pawari/Powari, 107.Phom, 108.Pnar/Synteng, 109.Pochury, 110.Punjabi, 111.Purkhi, 112.Rai, 113.Rajasthani, 114.Reang, 115.Rengma, 116.Rongmei, 117.Sadan/Sadri, 118.Sambalpuri, 119.Sangtam, 120.Sanskrit, 121.Santhali, 122.Saurashtra/Saurashtri, 123.Sema, 124.Shina, 125.Sindhi, 126.Sirmauri, 127.Sugali, 128.Surjapuri, 129.Talgalo (Galo), 130.Tamil, 131.Tangkhul, 132.Telugu, 133.Thado/Thadou, 134.Tibetan, 135.Tikhir, 136.Tripuri, 137.Tulu, 138.Urdu, 139.Vaiphei, 140.Wagdi, 141.Wancho, 142.Yerava, 143.Yerukala/Yerukula, 144.Yimchungre, 145.Zeliang, 146.Zemi, 147.Zou.

The resultant dataset of the first phase is available in public domain at LDC-IL data portal[2] under the title *The Mother Tongue Parallel Text Corpus of India Vol. I.* (Dr.Narayan Kumar Choudhary, et al., 2025) The dataset is available free of cost for non-commercial users.

5. Discussion

The development of the Mother Tongue Parallel Text Corpus of India faced several challenges, both at the practical and the linguistic levels. India's vast linguistic diversity can pose challenges because there are many significant regional and dialectical variations. For instance, Hindi spoke in Delhi and Jharkhand shows variations in pronunciation, vocabulary and grammar. These challenges and the LDC-IL approach to addressing them are discussed here.

[1] www.trankit.in

[2] data.ldcil.org

5.1 Language Identification and Classification

LDC-IL planned to build a parallel corpus of 270 mother tongues as per 2011 census. But while implementing the project it is found that all the 270 names in the list are not actually mother tongues, sometimes it may be the name of a tribe or community. For example, Kodu is a community name but in census it is considered as a mother tongue. Hence it was dropped from the intended mother tongue list. There were instances where the language may have been marked both with super-class and sub-class. For example, mother tongue 'Ao' has two sub-classes namely Mongsen and Chungli, both were collected. It can be argued that Kodava and Coorgi are the same language but mentioned distinctly in the census, both mother tongues were collected because of the ambiguity. There are cases such as Poula which is noted as Paola in Census, communities were highly particular about correct spelling as it is not just a tool of communication but their identity. The data is collected under the name Poula as community identifies.

5.2 Script Concerns

Indian languages use many scripts. In contrast, for geographical, political, or historical reasons, several Indian languages employ more than one script. If the translators and reviewers agreed on a single script then that script was used for the corpus. For example, in Korava, one freelancer preferred Tamil script and the other preferred Malayalam script, so the work could not continue.

If Unicode keyboards are not available for certain scripts then corpus development for the mother tongue is postponed until a technological solution is found. There is an instance where the freelancers of Savara language preferred International Phonetic Alphabet (IPA) instead of any predominant script, and the mother tongue is never initiated.

5.3 Human Resource Constraints

The empanelment of freelancers from remote locations, who are computer-literate, as well as proficient in their mother tongue along with either English or one of the Indian scheduled languages, is a challenging task. The corpus creation process consists of two main stages namely translation and review, designed to ensure quality of the text. Each translated sentence is validated by a reviewer to identify and correct any translation errors.

If freelancers are not proficient in English, they will be given any of the 22 Indian scheduled languages in which they are proficient as a source language. For example, Paniya mother tongue from Kerala used Malayalam as a source language. So, the translation has happened from Malayalam to Paniya not from English to Paniya. While this layered translation approach increases accessibility, it increases the risk of translation deviations too.

5.4 Issue of Rationalization of Remuneration

Corpus creation is a laborious and cost-intensive affair. The payments need to be uniform for the work of translation and review. LDC-IL conducted the corpus building project in workshop mode. Remuneration was paid at a throughput rate per working day based on the number of words in the source language. Based on the work done internally for Hindi at LDC-IL, a throughput of 1,000 source words has been set for translation, and 1400 source words and its translation have been set for review. This throughput rates are same whenever English is the source language.

When using languages like Malayalam and Kannada as sources, these throughput figures cannot be applied because they are agglutinative in nature. In these languages word count will be less compared to English or Hindi because a single lexical item will be a combination of two or more words.

The Director of the Central Institute of Indian Languages constituted a committee to review and rationalize the applied throughput rates when any schedule language is used as source language instead of English.

The committee selected the translated equivalents of not less than 5,000 English words from each scheduled language. Then the resultant word count in each target language is used to calculate the expected daily throughput equivalent of 1,000 English words for translation and 1,400 words of English text for review.

Table 1 provides the calculations for the Translation Equivalent for 1,000 English Words for each scheduled language.

SL No.	Language	English Source (Words)	Target Language (Words)	Translation Equivalent (1,000 English Words)
1	Malayalam	5,009	3,339	667
2	Kannada	5,006	3,389	677
3	Telugu	6,844	4,661	681
4	Tamil	5,034	3,476	691
5	Bodo	5,034	3,783	751
6	Konkani	5,034	3,923	779
7	Assamese	5,034	4,004	795
8	Manipuri	5,034	4,008	796
9	Bengali	5,601	4,406	787
10	Odia	5,034	4,163	827
11	Sanskrit	5,034	4,090	812
12	Marathi	5,034	4,124	819
13	Nepali	5,034	4,198	834
14	Gujarati	5,034	4,664	926
15	Kashmiri	5,034	4,870	967
16	Maithili	5,034	4,937	981
17	Hindi	5,394	5,576	1,034
18	Urdu	5,036	5,160	1,025
19	Dogri	5,034	5,143	1,022
20	Sindhi	5,034	5,214	1,036
21	Santhali	5,034	5,296	1,052
22	Punjabi	5,002	5,371	1,074

Table 1.Translation equivalence for English words

Based on these calculations the committee recommended, a normalized daily throughput of 650 words while translating from Malayalam, Kannada, Telugu and Tamil; 750 words for Bodo, Bengali and Konkani; 800 words for Assamese, Manipuri, Odia, Sanskrit, Marathi, Nepali; 900 words for Gujarati; 950 words for Kashmiri and Maithili; whereas 1050 words for Santhali and Punjabi. Hindi, Urdu, Dogri, Urdu, Sindhi were set as same as English, 1000 words per working day.

Table 2 provides the calculations for the Translation Equivalent for 1,400 English words for each scheduled language.

SL No.	Language	English Source (Words)	Target Language (Words)	Review Equivalent (1,400 English Words)
1	Malayalam	5,009	3,339	933
2	Kannada	5,006	3,389	948
3	Telugu	6,844	4,661	953
4	Tamil	5,034	3,476	967
5	Bodo	5,034	3,783	1,052
6	Konkani	5,034	3,923	1,091
7	Assamese	5,034	4,004	1,114
8	Manipuri	5,034	4,008	1,115
9	Bengali	5,601	4,406	1,101
10	Odia	5,034	4,163	1,158
11	Sanskrit	5,034	4,090	1,137
12	Marathi	5,034	4,124	1,147
13	Nepali	5,034	4,198	1,168
14	Gujarati	5,034	4,664	1,297
15	Kashmiri	5,034	4,870	1,354
16	Maithili	5,034	4,937	1,373
17	Hindi	5,394	5,576	1,447
18	Urdu	5,036	5,160	1,434
19	Dogri	5,034	5,143	1,430
20	Sindhi	5,034	5,214	1,450
21	Santhali	5,034	5,296	1,473
22	Punjabi	5,002	5,371	1,503

Table 2.Review Equivalence for English words

For review, the committee recommended a normalized daily throughput of 950 words for Malayalam, Kannada, Telugu, Tamil; 1,050 words for Bodo, Konkani; 1,100 words for Assamese, Manipuri, Bengali, Sanskrit; 1,150 words for Odia, Marathi, Nepali, 1,300 words for Gujarati, 1,350 words for Kashmiri, Maithili; 1,450 words for Hindi, Santhali, Sindhi, and 1,500 words for Punjabi. The Urdu and Dogri were set at the same throughput rates as English 1,400 words for working day

A 10% tolerance limit is used for the normalization of throughput word count; if a value is within 10% of the next increment then it is rounded up; otherwise, it is rounded down to the nearest 50.

5.5 Linguistic Challenges

There are grammatical asymmetries between languages. Features such as honorificity is absent in Malayalam and grammatical genders are not found in many Dravidian and tribal languages. The absence of some of these features leads to syntax inconsistencies during translation.

Moreover, syntactic mismatches were found in translation. Structures such as passive voice or double negative are not common in some languages. When such constructions were translated, they appeared artificial to native speakers.

5.6 Socio-Cultural Factors

In some communities, speakers were reluctant to share linguistic knowledge, considering their language as sacred and not to be shared with others. It is difficult to get data from these types of communities.

5.7 Ease of Access Challenges

There is a scarcity of human resource who is both proficient in their mother tongue and computer literacy. This challenge becomes more in remote parts of India, where access to computers and the internet is limited. Hence, even those who have proficient in their mother tongue may not be participating in online that require technological needs. This hinders the representation of less resourced linguistic communities to be a part of technological development efforts.

6. Conclusion

The Mother Tongue Parallel Text Corpus of India represents a pioneering effort to build linguistic resources for India's diverse mother tongues. By combining linguistic documentation principles with modern translation and corpus-building methodologies LDC-IL has created a resource for linguistics, computational research, and language pedagogy. The corpus is not yet large enough for full-scale AI applications but it built a foundation for future resource augmentation.

References

Abbi, Anvita.(2001). *A Manual of Linguistic Field Work and Structures of Indian Languages*.Muenchen: Lincom Europa.

Bhattacharya, Krishna et al.(2016).*Language Documentation Handbook*. Mysore: CIIL.

Comrie, Bernard & Smith, Norval.(Eds.) (1977). *Lingua Descriptive Studies: Questionnaire.* Amsterdam: North-Holland Publishing Company.

Narayan Kumar Choudhary, Anand Jain, Neha Dixit, Prof.Shailendra Mohan.(2025). *The Mother Tongue Parallel Text Corpus of India Vol.I* CIIL, Mysore.978-93-48633-08-8

Sivashanmugam, C. &Thayalan, V. (2012).*(Comp.) Comprehensive Questionnaire for Tribal Studies*. Coimbatore: Dept. of Linguistics, Bharathiar University.

Rejitha K. S. and Narayan Kumar Choudhary. (ed.). (2025). *LDC-IL Corpus Insights*. CIIL, Mysore. 978-93-48633-33-0.