A Large-scale Evaluation of Neural Machine Transliteration for Indic Languages

Anoop Kunchukuttan, Siddharth Jain, Rahul Kejriwal
Microsoft India, Hyderabad
{ankunchu,sija,rakejriw}@microsoft.com

Abstract

We take up the task of large-scale evaluation of neural machine transliteration between English and Indian languages, with a focus on multilingual transliteration to utilize *orthographic similarity* between Indian languages. We create a corpus of 600K word pairs mined from parallel translation corpora and monolingual corpora, which is the largest transliteration corpora for Indian languages mined from public sources. We perform a detailed analysis of multilingual transliteration and propose an improved multilingual training pipeline for Indic languages. We analyse various factors affecting transliteration quality like language family, transliteration direction and word origin.

1 Introduction

Transliteration is an essential technology for multilingual and cross-lingual capabilities in NLP applications to handle named entities, support cross-script input methods. Transliteration between English and Indic languages is important since English is widely used in the Indian subcontinent. Indic languages are written in different scripts from various writing systems. We focus on languages using scripts derived from the ancient Brahmi script. Their character sets are very different from the Latin script - making transliteration non-trivial.

These scripts are *abugida* scripts, where the basic unit is the *akshar* which consists of one or more consonants along with a vowel diacritic (Daniels and Bright, 1996). They exhibit a high degree of grapheme-to-phoneme correspondence. There is a large overlap in the logical character sets of these scripts, though the visual appearance of the characters varies. The languages utilizing these scripts are said to exhibit *orthographic similarity* on account of various shared characteristics (Kunchukuttan et al., 2018a).

We undertake a systematic, large-scale evaluation of neural machine transliteration for 10 major Indic languages from 2 major language families (Indo-Aryan and Dravidian languages) spoken by more than a billion speakers. Other than *BrahmiNet* (Kunchukuttan et al., 2015) and *Dakshina* (Roark et al., 2020), no other previous work has explored a wide range of Indic languages; Dakshina only explores transliteration into Indic languages. Our major contributions are:

- For a large-scale evaluation, we mine 600K transliteration pairs across 10 languages from publicly available parallel and monolingual sources. This is much larger than existing corpora like MSR-NEWS (Banchs et al., 2015), Brahminet (Kunchukuttan et al., 2015), Dakshina (Roark et al., 2020) and other small datasets (Banchs et al., 2015; Kunchukuttan et al., 2018b; Gupta et al., 2012; Khapra et al., 2014). The BrahmiNet and Dakshina datasets span multiple languages; BrahmiNet is small and Dakshina by design consists mostly of Indian origin words.
- From the mined corpus, we create a high-quality, manually validated testset annotated with foreign and Indian origin words.
- We propose various improvements to the multilingual transliteration system proposed by Kunchukuttan et al. (2018a) for Indian languages, and suggest a *recipe* for building multilingual transliteration systems for Indic languages.
- We present an evaluation of transliteration systems according to various factors like language family, word origin and transliteration direction.

2 Mining transliteration corpus

This section explains our transliteration mining methods (from parallel and monolingual corpora) and presents an analysis of the mined corpus. We mine transliteration corpora from English to 10 In-

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 3469–3475
April 19 - 23, 2021. ©2021 Association for Computational Linguistics

Language	pa	hi	bn	or	gu	mr	kn	te	ml	ta
Word pair count ($\times$1000)	55.3	157.7	65.4	34.7	65.5	38.0	24.7	77.4	31.1	57.1
Mining Accuracy	81.2	NA	76.7	NA	93.0	89.0	87.1	86.2	82.3	77.9

Table 1: Statistics on mined transliteration corpora

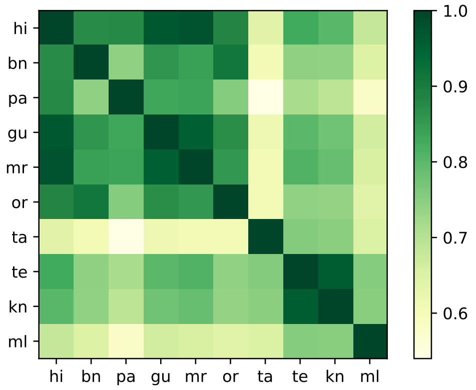

Figure 1: Orthographic Similarity: Indic languages

dian languages from 2 major language families: (a) Indo-Aryan branch of Indo-European family (Hindi, Marathi, Gujarati, Bengali, Odia, Punjabi), (b) Dravidian family (Kannada, Telugu, Malayalam, Tamil).

2.1 Mining from Parallel Translation Corpus

While very little transliteration corpora exists, a reasonable amount of parallel translation corpora between English and Indian languages are available in the public domain.

Method. Alignments between words in parallel sentences from two languages can be discovered from parallel translation corpora. The aligned words can either be translations or transliterations. We use the unsupervised method suggested by Sajjad et al. (2012) to mine transliteration pairs from these word alignments by distinguishing translation and transliterations. For morphologically rich languages, the approach can discover partial transliterations also. For instance, the English-Marathi pair word (*station*, स्टेशनावर [sTeshanaavara]). The Marathi word includes the locative case marker. To remove such transliteration pairs, we identify morphological variants by clustering together Marathi words corresponding to the same English word in the candidate transliteration pairs. We only retain the pair with the root word.

Mining Details. We mined transliteration pairs from English to Indian language parallel translation corpora from different sources (7.4 million sentence pairs across all languages, see Appendix A for details). We use the Moses transliteration mining module (Durrani et al., 2014) implementation of Sajjad et al. (2012) to mine transliteration pairs using the default settings.

2.2 Mining from Monolingual Corpora

Monolingual text corpora often have borrowed words from other languages (particularly English). We mine such transliteration pairs using only the vocabularies in the source and target languages.

Method. We first train initial transliteration models using available data in both directions ($L_e \rightarrow L_x$,$L_x \rightarrow L_e$) and build vocabularies for both languages (L_e, L_x). Given words in L_e, we identify the most promising transliteration candidates from L_x and then re-score these candidates. The scoring is based on edit-distance between Double Metaphone[1] representations of the words, which we found works well in practice. We consider scores in L_e as well as L_x. We use ITRANS[2] conversion from Indic scripts to Latin in order to be able to compute Double Metaphone representations on the Indic language side. Note that the phonetic nature of Indic scripts enables conversion of Indic scripts to Double Metaphone that is sufficient for transliteration mining. Thus, the score for a candidates pair $s(e, x)$ is $E(e, T_{XE}(x)) + E(x, T_{EX}(e))$, where E is the edit-distance function and T_{xy} denotes transliteration from x to y. As mentioned above, the strings are converted to Double Metaphone representation prior to edit-distance computation. Finally, we prune the generated pairs based on a chosen threshold of scores.

Mining Details. We use monolingual vocabulary from the AI4Bharat IndicNLP dataset (Kunchukuttan et al., 2020) and the OSCAR corpus (Ortiz Suarez et al., 2019) for Indic languages. For English, we use the AI4Bharat IndicCorp dataset (Kakwani et al., 2020) which contains crawls from English newspapers from India - this helps mining

[1] https://en.wikipedia.org/wiki/Metaphone#Double_Metaphone

[2] https://en.wikipedia.org/wiki/ITRANS

Indian named entities.

2.3 Characteristics of the Mined corpora

Corpora Statistics. Across 10 languages, we mined ~373k and ~339k transliteration pairs from the parallel translation and monolingual corpora respectively. The final train set of 606k word pairs was created after deduplicating and creating train, test and dev splits (See Table 1 for a summary of the mined corpus). We estimate that the training set has 55% non-Indian origin words and 45% Indian origin words.

Quality of the mined corpus. We evaluated the quality of mined transliterations via crowdsourcing. We used an internal, managed crowd-sourcing platform to validate the testsets and retained the transliteration pairs judged as correct in the final testset. The testset for every language had transliterations for 1500 English words and all their mined transliterations. This manual evaluation also gave us an estimation of the transliteration mining quality. We asked native-speaker judges for each language to report whether the pair is a transliteration or not. Our guidelines specified that pairs should be marked as valid if the pair is phonetically equivalent and are canonical spellings. In case no canonical spelling exists, the judges may mark the pairs solely on only phonetic equivalence. To control for quality, we used 3 judges per pair and used majority-voting for establishing correctness of a transliteration pairs. We added honey-pot pairs to tasks to filter out judges spamming our task.

Table 1 also shows the transliteration mining accuracies (average accuracy of 84.18%). An analysis of the errors revealed that an overwhelming majority involved wrong/missing/extra inflections (plurals and/or Indic casemarkers). These word pairs are also partial transliterations which are useful for learning transliteration models.

Test Set Creation. The test and dev sets were created by selecting 1500 English words each that are common across all language corpora along with their transliterations. We ensure that the test and dev set do not have any overlap with training set across languages. The testset were verified via crowdsourcing. The test set contains 928 foreign origin words and 572 Indian origin words.

Study of orthographic similarity. Following Kunchukuttan and Bhattacharyya (2020), we estimate the orthographic similarity between languages using the n-way parallel testset. For every language pair, it is the average Longest Common Subsequence Ratio (LCSR) (Melamed, 1995) between word pairs in the test set (See Figure 1) and follows linguistic genealogy. Tamil and Malayalam are most divergent to other languages. Punjabi is also divergent to other languages, possibly on account of: (a) some of its special characters like *tippi* and *addak*, (b) little use of conjunct consonants unlike other Indian languages.

3 Analysis: Multilingual Transliteration

We study multilingual transliteration models with the intent of identifying factors that improve multilingual models. First, we describe our baseline multilingual model and then introduce different variants to improve the baseline model.

Baseline Multilingual model (Kunchukuttan et al., 2018a). It is a character-level, attention-based, encoder-decoder model with all the model components shared amongst all the languages. We train joint EX (multi-target, English to Indian languages) and XE models (multi-source, Indian languages to English) separately. For EX models, we append a special target language token to the input sequence (Johnson et al., 2017).

Language Partitioning. To understand the role of orthographic similarity, we investigate two language groupings: (a) all the Indic languages are jointly trained, (b) Indo-Aryan and Dravidian languages are separately trained.

Vocabulary. Indic languages use a variety of scripts with a high overlap in the logical character set, but assigned unique characters in the Unicode character set. We investigate if transfer learning works better with a combined vocabulary by mapping logically equivalent characters across scripts for better transfer learning. We use the IndicNLP Library (Kunchukuttan, 2020) for mapping all Indic scripts to the Devanagari script, thus combining the vocabularies of all languages. We experiment with two configurations: (a) disjoint vocabularies (*i.e.,* different scripts), (b) combined vocabularies (*i.e.,* same script). Combining the vocabularies reduces the vocabulary significantly as the number of scripts reduces from 9 to 1.

Source language tag. In spite of the high degree of orthographic similarity between Indian languages, there are few cases of language-specific variations. For instance, the Malayalam script overloads a few characters with multiple sounds, Bengali pronunciation of the *aa* vowel differs, *etc.* To make the model sensitive to these language-

Experiment	Indo-Aryan (IA)						Dravidian (DR)				Average		
	pa	hi	bn	or	gu	mr	kn	te	ml	ta	IA	DR	IND
X to E TRANSLITERATION													
FOREIGN WORDS													
bilingual	48.53	52.29	50.85	48.84	43.81	52.93	56.95	50.58	54.9	38.28	49.54	50.18	49.80
all Indic	57.45	65.23	55.78	60.81	56.74	65.52	64.46	59.18	60.61	42.54	60.26	56.7	58.83
by family													
different scripts	61.27	65.23	59.64	60.58	59.18	66.19	63.23	58.28	61.02	44.29	62.02	56.7	59.89
same script	58.82	66.79	57.58	62.21	59.26	67.18	60.54	58.28	58.98	44.04	61.97	55.46	59.37
+source tag	60	64.50	60.36	60.93	59.10	68.07	62.56	57.48	63.88	45.04	**62.16**	**57.24**	**60.19**
INDIAN WORDS													
bilingual	68.01	73.18	68.02	74.79	71.14	81.29	77.63	74.29	71.9	45.99	72.74	67.45	70.62
all Indic	68.91	77.65	73.13	77.60	74.95	81.88	78	74.19	70.37	44.89	75.69	66.86	72.16
by family													
different scripts	71.49	78.32	73.85	79.07	75.16	82.96	79.5	76.11	75.18	49	76.81	69.95	74.06
same script	69.58	77.32	73.54	80.29	74.52	82	74.88	73.19	74.47	50.6	76.21	68.28	73.04
+source tag	71.16	78.44	72.92	79.56	75.05	84.03	78.88	76.21	75.18	50.5	**76.86**	**70.19**	**74.19**
E to X TRANSLITERATION													
FOREIGN WORDS													
bilingual	74.24	68.36	80.14	72.02	75.21	73.77	73.45	74.23	65.09	67.22	73.96	70	72.37
all Indic	75.83	76.92	81.29	75.43	80.68	79.97	77.4	78.55	68.53	73.22	**78.35**	74.43	76.78
by family													
different script	76.55	73.44	81.14	74.29	79.45	77.15	78.67	80.36	71.39	75.73	77.00	**76.54**	76.82
same script	77.99	74.17	82.29	74.43	80.27	77.57	77.68	79.11	71.24	73.5	77.79	75.38	**76.83**
INDIAN WORDS													
bilingual	78.61	71.69	75.34	79.65	76.47	80.33	75	78.9	72.45	76.26	77.01	75.65	76.47
all Indic	82.83	79.22	85.04	83.42	85.16	87.11	78.51	81.85	77.96	80.9	**83.80**	79.81	**82.20**
by family													
different script	81.34	77.96	81.54	80.19	82.49	84.94	79.86	81.18	76.58	80.22	81.41	79.46	80.63
same script	83.11	79.64	84.77	81.27	84.22	85.62	79.86	83.33	77.55	81.72	83.10	**80.62**	82.11

Table 2: Multilingual Transliteration results (%accuracy). The *all Indic* experiment trains with a common script.

specific variations in XE models, we add an special source language token in the input sequence.

Addressing divergence between Tamil and other Indic scripts. The Tamil script is highly under-specified and has fewer characters than sounds in the English language (unlike other Indic scripts). When training a multilingual model, there is an inconsistency in learnt mappings between Tamil and other Indic scripts. We address this issue by training a Tamil-specific multilingual model for Dravidian languages where all characters from other scripts are mapped to the closest character in the Tamil script via deterministic rules using the IndicNLP library.

3.1 Experimental Setup

We use *Marian* (Junczys-Dowmunt et al., 2018) to train our transliteration models. We use 128 LSTM units for encoder and decoder (1 layer for bilingual models and 2 layers for multilingual models). The encoder uses a bidirectional LSTM. The input embeddings are also 128 units in size. These hyperparameters were decided based on a parameter sweep on the dev set. We use a batch size of 100 sequences and early stopping with patience=100. We use beam-search for decoding (beam size=4).

3.2 Results and Discussion

Table 2 shows the top-1 accuracy of the different models. For translation into Indian languages, multiple references are available.

Bilingual models. Bilingual results show some trends about Indic transliteration. Transliteration is more difficult for non-Indian origin words than Indian origin words. EX direction accuracies are higher than XE direction. Tamil transliteration accuracy is the least in the XE direction (due to deficient orthography), while Malayalam has the least accuracy in the EX direction (possibly due to overloading of some characters).

Impact of multilingual training. In the XE direction, multilingual systems provide significant gains over bilingual systems (~20%). Most gains come from improved accuracy on non-Indian words. The multilingual system is better at generating canonical spelling in contrast to phonetically-equivalent incorrect spellings. More gains are observed for Indo-Aryan languages compared to Dravidian languages. Tamil benefits the most among Dravidian languages, but it still lags behind other languages. In the EX direction, accuracy improves by 6-7% for both Indian and non-Indian words. Both Indo-Aryan and Dravidian languages benefit from multilingual training.

Effect of language family. We do not observe any major advantage in training the two language families together. Hence, in subsequent experiments we train separate models. There are some differences in the spelling conventions between these languages. Thus, it seems a reasonable conservative choice to train separate model for the two families in the face of data not bringing any clarity.

Effect of source language tag. Adding source language tag improves XE transliteration accuracy 6-7% for some languages with divergent spelling conventions (Bengali, Malayalam and Tamil).

Effect of script conversion. It significantly reduces the vocab size (reducing number of scripts from 9 to 1), but results in just a small drop in accuracy. For the XE direction, the drop in accuracy is recovered by using the source language tag.

Mapping to Tamil Script. The table below show that this simple approach improves the Tamil transliteration accuracy by 2-5 points on non-Indian as well as Indian words in both directions.

Experiment	ta-en		en-ta	
	foreign	indic	foreign	indic
hiscript	44.04	50.6	73.5	81.72
tascript	**47.37**	**53.3**	**78.8**	**83.9**

4 Conclusion

We present a study of transliteration between English and Indic languages. We mine a 600k parallel transliteration corpus having a good coverage of Indian and non-Indian origin words as well as create a manually validated testset. We recommend the following recipe for Indic multilingual transliteration: (a) all training data in the same script, (b) separate models for IA and DR languages, (c) source language tags for XE transliteration. Multilingual training significantly improves non-Indian word transliteration. Our results validate previous results on benefits of multilingual transliteration on a wider set of languages and larger datasets. We also improve Tamil to English transliteration by representing multilingual data in Tamil script. More details about the corpus is available at https://github.com/anoopkunchukuttan/indic_transiteration_analysis.

References

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Rafael E. Banchs, Min Zhang, Xiangyu Duan, Haizhou Li, and A. Kumaran. 2015. Report of NEWS 2015 Machine Transliteration Shared Task. In *Proceedings of the Fifth Named Entities Workshop*.

Peter T Daniels and William Bright. 1996. *The world's writing systems*. Oxford University Press on Demand.

Nadir Durrani, Barry Haddow, Philipp Koehn, and Kenneth Heafield. 2014. Edinburgh's phrase-based machine translation systems for WMT-14. In *Proceedings of the ACL 2014 Ninth Workshop on Statistical Machine Translation*.

Kanika Gupta, Monojit Choudhury, and Kalika Bali. 2012. Mining Hindi-English transliteration pairs from online Hindi lyrics. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2459–2465, Istanbul, Turkey. European Language Resources Association (ELRA).

Barry Haddow and Faheem Kirefu. 2020. PMIndia – A Collection of Parallel Corpora of Languages of India. *arxiv 2001.09907*.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of EMNLP*.

Mitesh M Khapra, Ananthakrishnan Ramanathan, Anoop Kunchukuttan, Karthik Visweswariah, and Pushpak Bhattacharyya. 2014. When transliteration met crowdsourcing: An empirical study of transliteration via crowdsourcing using efficient, non-redundant and fair quality control. In *LREC*, pages 196–202. Citeseer.

Anoop Kunchukuttan. 2020. The IndicNLP Library. https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf.

Anoop Kunchukuttan and Pushpak Bhattacharyya. 2020. Utilizing language relatedness to improve machine translation: A case study on languages of the Indian Subcontinent. *arXiv preprint arXiv:2003.08925*.

Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. AI4Bharat-IndicNLP Corpus: Monolingual Corpora and Word Embeddings for Indic Languages. *arXiv preprint arXiv:2005.00085*.

Anoop Kunchukuttan, Mitesh Khapra, Gurneet Singh, and Pushpak Bhattacharyya. 2018a. Leveraging orthographic similarity for multilingual neural transliteration. *Transactions of the Association for Computational Linguistics*, 6:303–316.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018b. The IIT Bombay English-Hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Anoop Kunchukuttan, Ratish Puduppully, and Pushpak Bhattacharyya. 2015. Brahmi-Net: A transliteration and script conversion system for languages of the Indian subcontinent. In *Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies: System Demonstrations*.

I Dan Melamed. 1995. Automatic evaluation and uniform filter cascades for inducing n-best translation lexicons. In *Third Workshop on Very Large Corpora*.

Pedro Javier Ortiz Suarez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora*, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Shantipriya Parida, Ondrej Bojar, and Satya Ranjan Dash. 2018. Odiencorp: Odia-english and odia-only corpus for machine translation. In *Proceedings of the Third International Conference on Smart Computing and Informatics*.

Matt Post, Chris Callison-Burch, and Miles Osborne. 2012. Constructing parallel corpora for six Indian languages via crowdsourcing. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 401–409, Montréal, Canada. Association for Computational Linguistics.

Loganathan Ramasamy, Ondřej Bojar, and Zdeněk Žabokrtský. 2012. Morphological Processing for English-Tamil Statistical Machine Translation. In *Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages*.

Brian Roark, Lawrence Wolf-Sonkin, Christo Kirov, Sabrina J. Mielke, Cibu Johny, Işin Demirşahin, and Keith Hall. 2020. Processing South Asian languages written in the Latin script: the Dakshina dataset. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC)*, pages 2413–2423.

Hassan Sajjad, Alexander Fraser, and Helmut Schmid. 2012. A statistical model for unsupervised and semi-supervised transliteration mining. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *arxiv 1907.05791*.

Parth Shah and Vishvajit Bakrola. 2019. Neural Machine Translation System of Indic Languages - An Attention based Approach. In *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)*. IEEE.

Shashank Siripragrada, Jerin Philip, Vinay P. Namboodiri, and C V Jawahar. 2020. A multilingual parallel corpora collection effort for Indian languages. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3743–3751, Marseille, France. European Language Resources Association.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

A Corpora used for Transliteration mining

Table 3 shows the language-wise parallel corpora statistics and Table 4 lists the various parallel corpora used for transliteration mining. Table 3 also shows the vocab sizes of monolingual corpora used for each language for the monolingual approach.

Language		ParInfo	MonoInfo
Punjabi	(pa)	535,796	476K
Hindi	(hi)	1,586,775	511K
Bengali	(bn)	444,593	501K
Oriya	(or)	116,492	490K
Gujarati	(gu)	557,342	524K
Marathi	(mr)	732,093	539K
Kannada	(kn)	450,139	479K
Telugu	(te)	664,670	596K
Malayalam	(ml)	708,266	596K
Tamil	(ta)	1,640,920	599K
English	(en)	-	781K
Total		7,437,086	6092K

Table 3: Information on copora used for transliteration mining. **ParInfo**: Parallel Translation Corpora Size per English-Indian language pair. WikiMatrix contributes around 1.7 millions pairs across languages. **MonoInfo**: Vocabulary size of monolingual corpora per language.

Source	Citation
CVIT-Mann ki Baat	(Siripragrada et al., 2020)
CVIT-PIB	(Siripragrada et al., 2020)
IITB en-hi v2.0	(Kunchukuttan et al., 2018b)
MTurk Corpora	(Post et al., 2012)
JW300	(Agić and Vulić, 2019)
MTEnglish2Odia	
NLPC-Uom Corpus	
OdiEnCorp 1.0	(Parida et al., 2018)
OPUS	(Tiedemann, 2012)
PMIndia	(Haddow and Kirefu, 2020)
UFAL-en-ta-v2	(Ramasamy et al., 2012)
Urs Tarsadia Corpus	(Shah and Bakrola, 2019)
Wikimatrix	(Schwenk et al., 2019)
Wikititles	

Table 4: Parallel Translation Corpora used for mining transliterations. All download URLs can be obtained from `https://github.com/AI4Bharat/indicnlp_catalog`