Insights from Romanized Manipuri Social Media Text: A Transliteration Corpus and Variation Analysis

Maisang Kamei Salice, Sanasam Ranbir Singh, Priyankoo Sarmah
Indian Institute of Technology Guwahati
North Guwahati, India
{s.maisang, ranbir, priyankoo}@iitg.ac.in

Abstract

This paper presents the first large-scale study of Romanized Manipuri, a low-resource Indic language widely used by native speakers on social media. Social media text is highly informal and often noisy, posing challenges for natural language processing tasks; therefore, normalization through back-transliteration is essential. We construct a Romanized Manipuri to Manipuri–Bengali script back-transliteration corpus from YouTube comments, capturing diverse informal writing styles and orthographic variations. The dataset is analyzed to examine variation patterns at two levels: character-level inconsistencies and pragmatic stylistic variations influenced by user writing behavior. We also compare social media romanization with formal transliteration conventions, including standardized romanization schemes and textbook-based systems. Furthermore, we evaluate Transformer model at both character and subword levels and conduct a detailed error analyses to identify key challenges affecting back-transliteration performance.

Keywords: Manipuri, low-resource language, Indic language, back-transliteration, social media text

1. Introduction

With the rapid growth of communication on social media platforms, multilingual speakers increasingly use Roman script to write their native languages, even when those languages have distinct writing systems. This practice of writing a language in another script is known as transliteration, while converting it back into the original script is termed back-transliteration. When the target script is the Roman (Latin) script, the process is more specifically called Romanization.[1] Romanization is primarily motivated by the convenience of standard QWERTY keyboards and the need for cross-lingual communication. Although standardized Romanization schemes such as ISO 15919[2] and ITRANS[3] exist, these are rarely followed in informal digital communication. As a result, social media user-generated Romanized text displays substantial inconsistency in the representation of phonemes, syllables, and words. Such irregularities introduce significant challenges for Natural Language Processing (NLP) tasks including transliteration, machine translation, and information retrieval.

These challenges are not unique to Manipuri but are shared across many low-resource languages that rely on informal Romanized writing in digital communication. Similar orthographic variability and noisy transliteration patterns have been documented for Arabizi (Bies et al., 2014), Romanized Assamese (Baruah et al., 2024a), Singlish (Liwera and Ranathunga, 2020), and Romanized Bangla (Fahim et al., 2024) in recent back-transliteration and social media studies. The transliteration task, which involves preserving pronunciation, is particularly sensitive to complex phoneme–grapheme correspondences. Thus, when Romanized text is inconsistent or noisy, back-transliteration becomes inherently ambiguous, as a single Romanized word may correspond to multiple valid native-script forms.

The Manipuri language exemplifies these challenges and serves as the focus of this study. Although Manipuri is a low-resource language with limited formal digital presence, it exhibits a distinctive hyper-local popularity. In localized online spaces—such as Manipuri-centric YouTube comment sections and community forums—Romanized Manipuri is the predominant mode of written communication. This occurs despite the absence of a standardized Romanization convention and despite the language traditionally being written in Manipuri-Bengali and Meitei Mayek scripts.

This localized concentration of informal Romanized text provides a rich but noisy source of linguistic data. However, the language's overall digital footprint remains relatively small compared to high-resource languages such as English on global platforms. The lack of standardization has resulted in extensive phonological substitutions, dialectal influences, and user-driven stylistic variations, leading to multiple Romanized spellings for the same native-script word.

Existing transliteration research for Manipuri has primarily relied on clean and canonical datasets. For example, multilingual transliteration benchmarks such as (Madhani et al., 2022), which include

[1] https://en.wikipedia.org/wiki/Romanization
[2] https://en.wikipedia.org/wiki/ISO_15919
[3] https://www.aczoom.com/itrans/online/

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1878–1888
11–16 May 2026.

Manipuri, are designed and evaluated using standardized, noise-free data. Such approaches typically assume near one-to-one grapheme–phoneme correspondences—an assumption that does not hold for informal Romanized social media text. Despite the growing prevalence of Romanized Manipuri in online communication, there has been no systematic study examining its linguistic variability or the resulting transliteration complexity in real-world settings.

In contrast to prior work on transliterations, our corpus captures the natural variation present in social media user-generated content. In particular, this study highlights:

•**Pragmatic and stylistic variations**, including vowel elision, digit-based substitutions (e.g., the use of symbol '*2*' to represent phonologically similar syllables or to indicate repetition), and character repetition in stretched words (e.g., "*neirhhhhhh*", a stylized variant of the canonical form "*neirehe*",'it's cool! / it's awesome!'), which are widely observed in social media writing but remain challenging for existing transliteration models.

•**Phonemic inconsistency**, characterized by non-uniform mappings between native graphemes and Romanized forms (e.g., the native phoneme /p/ realized as 'p', 'b', or 'f'), reflecting substantial variability in user spelling conventions.

The contributions of this work are threefold:

•**Corpus Construction:** We curate and annotate a Romanized-Manipuri to Manipuri-Bengali script back-transliteration corpus derived from YouTube comments, providing a word-level parallel dataset.[4]

•**Variation Analysis:** We conduct an in-depth linguistic analysis of Romanization patterns, identifying character-level and pragmatic–stylistic phenomena (e.g., elision, digit usage, code-mixing) that are widely observed in Romanized social media text across low-resource languages.

•**Model Evaluation and Error Analysis:** We evaluate Transformer model (Vaswani et al., 2017) using both character-level and subword (BPE) representations (Gage, 1994), and conduct detailed error analyses to identify systematic weaknesses arising from the real-world writing variability.

2. Related Studies & Language Background

Transliteration Approaches and Roman-to-Indic Systems Transliteration systems are commonly categorized into three types based on the source of linguistic information: (i) phonetic-based (pivot), (ii) spelling-based (direct), and (iii) hybrid approaches (Karimi et al., 2011; Kaur and Singh, 2014; Nair and Ahammed, 2021; Mammadzada, 2023).

Early Roman-to-Indic transliteration systems primarily relied on statistical sequence models and intermediate representations. For instance, Bhat et al. (2014) proposed a back-transliteration system based on a second-order Hidden Markov Model trained using a structured perceptron, where Roman input is first mapped to WX notation before being converted to the target script. More recent work has shifted towards neural sequence-to-sequence and Transformer-based architectures. Multilingual datasets such as Dakshina (Roark et al., 2020) and Aksharantar (Madhani et al., 2022) provide large-scale benchmarks for South Asian languages, with Aksharantar also including Manipuri. However, these datasets mainly contain canonical transliteration pairs and do not capture the informal spelling variations commonly found in social media text. Ahmed et al. (2011) examined how phonological properties of Indian scripts contribute to variation and ambiguity in transliteration. Using Romanized input for Hindi, Bangla, and Telugu, their study of input method editors showed that phonological variation, spelling ambiguity, and code-mixing are major sources of transliteration errors.

Machine Transliteration in Manipuri Early Manipuri transliteration research focused on canonical script conversion between Bengali and Meitei Mayek. Nongmeikapam et al. (2011); Nongmeikapam and Bandyopadhyay (2012) developed rule-based systems, while Singh (2012) compared rule-based and statistical approaches for bidirectional transliteration. Laitonjam et al. (2018) explored neural models for English loanword and named-entity back-transliteration into Manipuri-Bengali, and Laitonjam and Singh (2022) proposed hybrid multi-source encoder–decoder architectures combining phoneme and grapheme inputs. However, these studies assume standardized input and do not address informal Romanized usage common in social media communication.

Transliteration in Informal and Social Media Text Noisy Romanized text in digital communication has received increasing attention. Romanized Arabic (Arabizi), characterized by digit substitutions and orthographic variation, is a widely studied example (Bies et al., 2014). Related work on dialectal Arabic orthography normalization highlights challenges in processing non-standard user-generated text (Habash et al., 2012).

In South Asia, Chakma and Das (2014) analyzed code-mixed Romanized social media text involving English, Bengali, and Hindi, while Vathsala and Holi (2020) demonstrated neural approaches for tweet transliteration. For Singlish (Sinhala in Roman script), Liwera and Ranathunga (2020) combined statistical and rule-based methods on YouTube data. Romanized Assamese social media text has also been studied, including analyses of spelling

[4] Dataset will be available on request.

The 24 Consonants	
15 indigenous	**9 exotic**
/p/, /p^h/, /t/, /t^h/, /c/, /k/, /k^h/, /m/, /n/, /ŋ/, /s/, /h/, /l/, /w/, /j/	/b/, /b^h/, /d/, /d^h/, /g/, /g^h/, /j/, /j^h/, /r/
The 12 Vowels	
6 monophthongs	**6 diphthongs**
/i/, /e:/, /a/, /a:/, /u/, /o:/	/ai/, /au/, /a:i/, /a:u/, /o:i/, /ui/

Table 1: Categorization of Manipuri Phonemes. The notation follows (Singh, 1975). A full phonetic chart is provided in Table 3 of the cited source.

variation (Baruah et al., 2024b) and neural back-transliteration models (Baruah et al., 2024a).

Despite these advances, systematic investigation of noisy Romanized Manipuri remains limited. Our work connects prior research on Manipuri transliteration with studies on informal Romanized social media text by introducing a substantial parallel corpus, systematically categorizing variation patterns, and analyzing model performance under realistic noisy conditions.

Language Background Manipuri is the lingua franca and official language of Manipur in Northeast India and one of the 22 scheduled languages of the Indian Constitution. It is also an additional official language in four districts of Assam. Smaller speech communities are found in Assam, Mizoram, Tripura, Bangladesh, and Myanmar[5]. Belonging to the Tibeto-Burman branch of the Sino-Tibetan family, Manipuri is typologically distinct from the Indo-Aryan languages of mainland India. Manipuri is highly agglutinative and its roots are mostly monosyllabic, though disyllabic forms also occur. Phonologically, it comprises 38 phonemes represented by 55 graphemes(table 1) in the Manipuri-Bengali script (Singh et al., 2007). The language remains a low-resource language compared to Indic languages like Hindi or Bengali, lacking governmental support, inadequate documentation, and limited access to technological resources(Pal et al., 2023). Although digital platforms like regional news portals exist, computational resources are scarce, and even Google Translate offers only limited support. Thus, despite its official status, Manipuri continues to be digitally underrepresented within the Indic language technology landscape.

3. Dataset

Data Collection. The dataset was compiled using the YouTube Data API by extracting comments and live chat messages from videos retrieved through 66 targeted search queries related to Manipuri-language content. These included 39 popular Manipuri YouTube channel names and 27 general keywords, ensuring broad topical coverage. All retrieved posts written in Romanized Manipuri were collected, spanning three thematic domains—*Entertainment* (43.9%), *Culture & Knowledge* (31.8%), and *News & Public Affairs* (24.2%). Data collection was conducted between July-September 2022, covering content published from 2019-2022. Although most posts were in Manipuri, a few were written entirely in Hindi, English, or other regional languages and were excluded. However, *code-switched* and *code-mixed* posts (e.g., combining Manipuri with English or Hindi) were retained.

Annotation Process. Annotation was conducted through a custom-built online platform hosted on a local server, where each annotator was provided secure login access and presented with one Romanized sentence at a time for back-transliteration into the Manipuri-Bengali script. A total of 27 native Manipuri speakers from Manipur participated. To ensure language proficiency, each annotator completed an initial qualification set of at least 20 sentences, which were manually reviewed by the first author, a native Manipuri speaker. Only those demonstrating sufficient competence were allowed to continue.

Quality control was implemented in two stages. In the first phase, all submitted annotations were manually verified by the first author; minor errors were corrected, while substantially incorrect submissions were rejected and returned for revision. In cases where typographical errors or highly stylized Romanized forms introduced ambiguity, annotators relied on sentence-level context to determine the most plausible native-script equivalent, following predefined annotation guidelines. In the second phase, validated sentence-level annotations were segmented into Roman–native word pairs and independently reviewed by a hired Manipuri linguist to ensure spelling consistency and reliability. Such ambiguous cases were further reviewed during validation to maintain uniformity across annotations. Annotators were compensated at INR 2 per sentence, and the linguist at INR 1 per word pair. Although multiple independent annotations were not collected for formal inter-annotator agreement computation, this two-stage validation process served as a quality assurance mechanism.

Lexical Distributions. The lexical structure of the dataset (in the native script) was examined using Zipf's Law (Zipf, 1949) and Heaps' Law (Heaps, 1978). The Zipfian slope is approximately –1.07, close to the theoretical value of –1.0, while the Heaps' Law relation $V = 27.023N^{0.738}$ yields $\beta = 0.738$, within the typical range for natural languages (0.4–0.8). These results confirm that the corpus exhibits natural lexical behavior.

Dataset Statistics. The overall dataset statistics are summarized in Table 2. The dataset consists of aligned Roman–native script word pairs obtained

[5] https://en.wikipedia.org/wiki/Meitei_language.

Description	Count
Unique Roman-script words	62,090
Unique native-script words	44,632
Unique Roman–native word pairs	73,661

Table 2: Dataset statistics

	Romanized Manipuri word	Native-script
Same Meaning variations	“natte”(161),“ntte”(150),“nte”(65), “matte”(7),“natty”(6),“ntea”(2), “ntee”(2),“ntdey”(1),“nutte”(1),etc.	“নৎতে”(/nat-te:/, ‘not’), “নত্তে”(/nat-te:/, ‘not’)
Different Meanings Variations	“ahanba”(134), “ahnba”(25), “ahnb”(4)	"অহন্বা"(/a-han-ba:/, ‘new one’), "অহান্বা(/a-ha:n-ba:/, ‘initial one’)

Table 3: Few examples from the dataset showing transliteration variations

from the annotated corpus.

4. Variations in Romanization

The dataset shows extensive variation—both across Romanized Manipuri words and in their native-script equivalents. Table 3 illustrates examples of such inconsistencies. The study of these variations are performed from two different aspects: (i) Character-Level Variations, involving inconsistencies in grapheme-to-phoneme representation, and (ii) Pragmatic and Stylistic Variations, arising from informal writing practices such as shortcuts or expressive spellings.

4.1. Character-Level Variations

Ph.	Gr.	Occ.	#Vr.	Roman Character Mappings
/p/	প	9920	3	‘p’: 99.17%, ‘b’:0.75%, ‘f’:0.07%
/p^h/	ফ	4549	4	‘f’:71.52%, ‘ph’:28.47%, ‘p’:0.52%, ‘b’:0.21%
/b/	ব	20864	3	‘b’:99.24%, ‘p’:0.36%, ‘v’:0.17%, ‘bh’:0.17%, ‘n’:0.02%, ‘f’:0.01%
/b^h/	ভ	119	3	‘v’:37.96%, ‘b’:32.41%, ‘bh’:29.62%
/t/	ত	11583	2	‘t’:99.51%, ‘d’:0.49%
	ট	112	1	‘t’:100%
	ৎ	4308	2	‘t’:99.77%, ‘d’:0.23%
/t^h/	থ	5336	2	‘th’:98.86%, ‘t’:1.14%
	ঠ	1	1	‘th’:100%
/d/	দ	19178	2	‘d’:99.14%, ‘t’:0.6%, ‘dh’:0.23%
/d^h/	ধ	97	2	‘dh’:70.88%, ‘d’:29.11%,
/c/	চ	4710	6	‘ch’:81.36%, ‘c’:1.9%, ‘j’:0.52%, ‘chh..’:0.13%, ‘sh’:0.06%, ‘xh’:0.03%
/j/	জ	5954	9	‘j’:94.81%, ‘s’:1.56%, ‘z’:1.45%, ‘y’:0.39%, ‘g’:1.38%, ‘ch’:0.19%, ‘jh’:0.17%, ‘x’:0.017%,‘c’:0.017%
/j^h/	ঝ	0		
/k/	ক	17237	4	‘k’:98.27%, ‘c’:0.88%, ‘g’:0.44%, ‘ck’:0.41%(coda)
/k^h/	খ	7312	3	‘kh’:96.11%, ‘k’:2.64%, ‘g’:1.24%
/g/	গ	11294	2	‘g’:98.59%, ‘k’:0.9%, ‘gh’:0.54%
/g^h/	ঘ	6	2	‘gh’:80%, ‘g’:20%
/m/	ম	16885	2	‘m’:99.58%, ‘n’:0.42%
/n/	ন	30651	2	‘n’:99.51%, ‘l’:0.48%
	ণ	70	2	‘n’:99.53%, ‘l’:1.47%
	ঞ	6	1	‘n’:100%
/ŋ/	ঙ	5758	4	‘ng’:97.21%, ‘g’:1.29%, ‘n’:1%, ‘l’:0.5%
	ং	14159	3	‘ng’:96.76%, ‘g’:1.95%, ‘n’:1.29%
/s/	স	3405	7	‘s’:86.9%, ‘sh’:8.2%, ‘j’:3.26%, ‘z’:0.19%, ‘ch’:0.8%,‘c’:0.64% ‘x’:0.03%
	ছ	18	2	‘s’:80%, ‘ch’:20%
	শ	11665	6	‘s’:88%, ‘sh’:8.86%, ‘j’:2.47%, ‘ch’:0.31%, ‘c’:0.3%, ‘x’:0.05%
	ষ	22	2	‘s’:70%, ‘sh’:30%,
/h/	হ	10797	1	‘h’:100%
/w/	ৱ	1641	2	‘w’:98.17%, ‘y’:1.83%
/r/	র	19705	3	‘r’:99.9%,‘l’:0.09%,‘rh’:0.005%
	ড়	16	1	‘r’:100%,
/y/	য়	8955	2	‘y’:97.93%, ‘j’:2.07%
/l/	ল	13360	3	‘l’:98.96%, ‘n’:0.97%, ‘r’:0.07%

Table 4: Statistics of Romanization Variations in Manipuri Consonant Phonemes

Ph.	Gr.	Occ.	#Var.	Roman Character Mappings
/i/	ই	10851	6	‘i’:76.99%, ‘e’:18.58%, ‘ee..’:2.18%, ‘ii..’:1.76%, ‘ae’:0.31%, ‘ea’:0.16%
	ি	29474	7	‘i’:87.98%, ‘e’:7.19%, ‘ii..’:1.89%, ‘ee’:1.45%, ‘y’:1.26%, ‘ae’:1.22%, ‘ea’:0.09%
	ঈ	18	2	‘e’:95%, ‘ee..’:5%
	ী	4044	7	‘i’:76.01%, ‘e’:17.37%, ‘ee’:4.19%, ‘ii..’:1.8%, ‘y’:0.41%, ‘ae’:0.16%, ‘ea’:0.05%
/e:/	এ	484	4	‘a’:60.37%, ‘e’:25.56%, ‘ee’:7.41%, ‘aa..’:6.67%
	ে	19278	8	‘e’:77.08%, ‘a’:5.33%, ‘ee’:5.16%, ‘ey’:5.03%, ‘ae’:2.53%, ‘y’:2%, ‘ay’:1.12%, ‘ea’:1.75%
/a:/	আ	19	2	‘aa..’:58.33%, ‘a’:41.67%
	া	41235	2	‘a’:97.95%, ‘aa..’:2.05%
/a/	অ	2983	2	‘a’:97.24%, ‘aa..’:2.75%
/o:/	ও	6508	3	‘o’:96.72%, ‘oo’:2.28%, ‘wo’:1%
	ো	14275	2	‘o’:98.12%, ‘oo’:1.88%
/u/	উ	1171	5	‘u’:91.95%, ‘o’:3.35%, ‘au’:1.71%, ‘uu..’:1.54%, ‘oo’:1.44%
	ু	14140	5	‘u’:97.75%, ‘o’:0.81%, ‘uu..’:0.74%, ‘oo’:0.61%, ‘w’:0.08%
	ূ	4	1	‘u’:100%

Table 5: Statistics of Romanization Variations in Manipuri Vowel-Monophthong Phonemes

Ph.	Gr.	Occ.	#Var.	Roman Character Mappings
/ai/	আই	2	2	‘ai’:50%, ‘i’:50%
	াই	3766	4	‘ai’:79.56%, ‘y’:19.1%, ‘ay’:0.76%, ‘aai’:0.58%
	ায়	1709	4	‘ai’:59.78%, ‘y’:40.01%, ‘ay’:0.2%, ‘aai’:0.06%
/ai/	ঐ	304	9	‘ei’:77.35%, ‘ai’:5.57%, ‘aei’:4.88%, ‘e’:4.18%, ‘a’:3.83%, ‘eai’:2.09%, ‘ay’:1.04%, ‘ey’:0.7%, ‘ae’:0.35%
	ৈ	5695	5	‘ei’:93.14%, ‘ai’:5.03%, ‘eii..’:1.5%, ‘aei’:0.19%, ‘aii..’:0.13%
/oːi/	ওই	1479	5	‘oi’:96.88%, ‘oei’:0.99%, ‘woi’:0.91%, ‘oii..’:0.61%, ‘oy’:0.61%
	োই	3097	4	‘oi’:98.81%, ‘oy’:0.95%, ‘oe’:0.21%, ‘oei’:0.04%
	োয়	59	2	‘oi’:99.1%, ‘oy’:0.9%
/ui/	উই	23	4	‘ui’:73.68%, ‘oui’:10.53%, ‘uii..’:10.53%, ‘uei’:5.26%
	ুই	283	6	‘ui’:88.36%, ‘oi’:3.88%, ‘uii..’:3.02%, ‘ue’:2.16%, ‘uei’:1.3%, ‘wi’:1.29%
/aːu/	াউ	75	6	‘w’:33.33%, ‘ao’:25.4%, ‘ou’:23.81%, ‘ow’:9.52%, ‘au’:4.76%, ‘aw’:3.17%
	াও	4160	6	‘ao’:54.23%, ‘w’:27.14%, ‘ou’:11.16%, ‘au’:3.99%, ‘aw’:2.33%, ‘ow’:1.14%
/au/	ঔ	3	1	‘ou’:100%
	ৌ	7926	6	‘w’:50.16%, ‘ou’:36.66%, ‘ao’:5.85%, ‘au’:4.61%, ‘aw’:1.51%, ‘ow’:1.2%

Table 6: Statistics of Romanization Variations in Manipuri Vowel-Diphthong Phonemes

From the dataset, we extracted all mappings between native-script graphemes and their Roman representations, organized by phonemic structure. A single native grapheme often corresponds to multiple Roman letters or sequences. While a few show consistent one-to-one mappings (e.g., ‘ঠ’(/t^h/) → ‘th’, ‘ঞ’(/n/) → ‘n’), such cases are rare. Tables 4, 5, and 6 present the phoneme-level transliteration variations for consonants, monophthongs, and diphthongs, along with their distributions. These variations are not random but systematic—most follow regular phonological or orthographic patterns that can be grouped into distinct categories as follows.

1. **Variations with Phonetically Similar Letters:** Besides the dominant Roman forms, several phonemes show frequent alternate representations with similar pronunciations. The examples below list such phonemes along with their dominant and alternate Roman forms and sample words.
   - /p^h/(ফ) →‘f’,‘ph’: “হৌরক্ফম্” (‘source; origin’) →“hourakfam”, “haorakpham”
   - /b^h/(ভ) →‘v’,‘bh’: “সেভা” (‘service’) →“seva”, “sebha”
   - /j/(জ) →‘j’,‘z’: “ওইজশনু” (‘let it be’) →“oijasanu”, “oizasnu”
2. **Voiced–Voiceless Plosive Variations:** Voiceless plosive phonemes are also romanized us-

ing their roman forms' voiced counterparts.e.g.:
• /p/(প) →'p','b' : "অপাক্পা" ('broad; wide') →"apakpa", "apakb"
• /t/(ত) →'t','d': "অতোপ্পগা" ('with other; with someone else') →"atoppaga", "adoppaga"
• /k/(ক) →'k','g': "কিহোম" ('pineapple') →"kihom", "gihom"
• /c/(চ) →'ch','j': "হিংচাবি" ('a female demon') →"hingchabi", "hinjabi"
Similarly, voiced plosive phonemes are also romanized using their roman forms' voiceless counterparts: /b/(ব) →'b','p'; /d/(দ) →'d','t'; /g/(গ) →'g','k'; /j/(জ) →'j','ch'.

3. **Variations in similar places of articulation:** Certain phonemes are alternatively romanized using letters produced at similar places of articulation. The phonemes and their common examples are listed below:
• /r/(র) →'r','l': য়াদরে ('it is not working') →"yadare", "yadale"
• /n/(ন) →'n','l': হোরেন্ ('later') →"horen", "horel"
• /y/(য়) →'y','j': য়াওদাবা ('not present') →"yaodaba", "jaodaba"
• /b/(ব) →'b','f': ফজবম্ ('something that is good or beautiful') →"fajabam", "fazafam"
• /p^h/(ফ) →'f','ph','b': চাফম্ ('place to eat') →"chafam", "chapham", "chabam"
• /k^h/(খ)→'kh','g':নুংঙাইখ্রেদা ('(I) feel good/happy/content') →"nungaikhreda", "nungaigreda".

4. **Alternate Roman Forms for Aspirated Plosives:** Aspirated consonants are often written in their unaspirated roman form: /p^h/(ফ)→'f','ph','p'; /b^h/(ভ)→'bh','b'; /t^h/(থ)→'th','t'; /d^h/(ধ)→'dh','d'; /k^h/(খ)→'kh','k'; /g^h/(ঘ)→'gh','g'. Conversely, unaspirated plosives are occasionally romanized with the aspirated form: /p/(প)→'p','f'; /b/(ব)→'b','bh'; /t/(ত)→'t','th'; /d/(দ)→'d','dh'; /k/(ক)→'k','kh'; /g/(গ)→'g','gh'.

5. **Dilemma of /ŋ/(ঙ/ং(only finalizer)):** The phoneme /ŋ/ often appears at syllable boundaries—ending one syllable and beginning the next. In such cases, users typically represent both occurrences with a single "ng" in Roman script instead of writing them separately. e.g., চংঙকপা(/caŋ-ŋak-pa:/, 'coming in')→"changakpa".
In other instances, the phoneme may be represented by either 'n' or 'g' alone. e.g.: /ŋ/ →'ng', 'g', 'n': ঙসি(/ŋasi/, 'today')→"ngac", "gasi", "nasi".

6. **Other Unusual Roman Mappings:** From Table 4, we observe several rare consonant–Roman mappings that occur with very low probabilities. Key cases are summarized below.
i). **/ch/(চ)–xh:** Found only twice, likely a typo due to proximity of 'c' and 'x' on the keyboard or a personal writing style. e.g.: ইচন্(/i-can/, 'sister')→"exhan"; শকচৈ(/sak-cai/, 'sings (respectful form)') →"sakxhei".
ii). **/s/(স)–x:** Mostly appears in English-influenced words where 'x' represents the cluster ক্স (/ks/). Rarely maps directly to /s/. e.g.: "নেক্সতা"(/ne:ks-ta:/, 'in/at the next (codemixed)') →"nextta"; "তেক্সি"(/te:k-si/, 'taxi (english word)') →"taxi"; rare: পুন্সিনা(/pun-si-na:/, 'by life; as life') →"pungxina".
iii). **/j/(জ)–x:** Used to represent the cluster ক্জ, reflecting English /gz/ sound as in "exam." e.g.: "এক্জাম্দা"(/e:k-ja:m-da:/, 'in exam (codemixed)') →"examda".
iv). **/j/(জ)–c:** The letter 'c' often represents the syllable /ji/ (জি, সি, শি), e.g.: "তৌশনজি"(/tau-san-ji/, 'lets just do') →"twsnc"; or directly maps to /j/, e.g.: "ফজদে"→"facade".
v). **/j/(জ)–s or /s/(স,শ)–j:** /s/ and /j/ are occasionally interchanged when used as affixes, without altering meaning. e.g.: "ঙাংজিন্লে"(/ŋa:ŋ-jin-le:/, '(someone) is verbally interrupted (while speaking); (something) reddening') →"ngangsinle"; "লৈরিশিদি"(/lai-ri-ʃi-di/, 'it exists / it is there') →"leirijidi".
vi). **/s/(শ)–x:** The use of 'x' for /sh/(শ), found in 9 instances, appears stylistic, likely influenced by Pinyin, where 'x' denotes a similar sound. e.g.: "শক্লম্মু"(/sak-lam-mu/, '(please) sing') →"xhklmmu".
vii). **/m/(ম)–n:** Occurs 23 times, arising from both typing errors (e.g.:"অচুম্বা"(/a-cum-ba:/, 'right; true') →"achunba") and natural phonetic alternation (/m/→/n/)(e.g.: "মথংদা"(/ma-t^haŋ-da:/, 'in/at the next') →"nathngda").
viii). **/ng/(ঙ)–l or /l/(ল)–ng:** Reflects elision at syllable boundaries. When a syllable ending in /ng/ is followed by one beginning with /l/, the /l/ sound is often dropped in speech. e.g.: তানিংলক্রে(/ta:-niŋ-lak-tre:/, '(I) do not want to listen/hear'), তানিংঙক্রে(/ta:-niŋ-ŋak-tre:/) →"taningaktre".
ix). **/w/(ৱ)–y:** A frequent and deliberate orthographic substitution (20 instances). e.g.: "তৌৱি"(/tau-wi/, '(I) do') →"touyee".

Key findings with respect to vowels from Table 5 and 6:

1. **Broader and Less Skewed Patterns in Vowel Romanization:** Vowel phonemes also show dominant Roman forms, but their secondary variants occur more frequently (≈20%), reflecting English-influenced spelling overlap. e.g.: 'a' → /e:/, /a/, /a:/; 'e' → /i/, /e:/, /a/; 'i' → /a:i/, /i/, /i:/; 'o' → /o:/, /u/; 'u' → /u/, /u:/, /a/.
2. **Intuitive Spelling in Vowels:** Social media users often rely on perceived sound when romanizing and also uses English-influenced spellings—e.g., representing monophthongs with two letters and diphthongs with three. For instance, the transliteration variations of /e:/ are

'ae', 'ay', 'ey', etc. and /ai/ are 'aei', 'eai', 'ei', etc.

3. **Vowel Stretching for Phonetic Emphasis:** Writers often repeat vowel letters to express elongation or emphasis. e.g.: "faaatjei" for "ফাজৈ"(/fa-jei/, 'nice; beautiful').
4. **Prominent Use of 'w' and 'y' for Syllable-Final Diphthongs:** The graphemes w and y are frequently employed to represent syllable-final diphthongs, particularly in open CV syllables, for the diphthongs listed below.
   • **'w'** → /a:u/ (33.33% for ৌউ: 24 instances; 27.14% for ৌও: 1129 instances) and /au/ (50.16%: 3975 instances).
   • **'y'** → /a:i/ (19.1% for ৌই: 719 instances; 40.01% for ৌয়: 683 instances).
5. **Inconsistent Representation of the Phoneme /a/ (IPA: /ə/):** The phoneme /a/ lacks a dedicated native grapheme, leading to inconsistent representation in informal or noisy texts. e.g., the words—"youhallase" and "youhllase", differ in the presence of 'a' in the <hal> and <hl> syllables, respectively. Both forms correspond to the same native word—য়ৌহল্লসে(/yau-hal-la-se:/, 'let's get it delivered/reached').

4.2. Pragmatic & Stylistic Variations

Another major source of variation stems from pragmatic writing behaviors, stylistic choices, and occasional performance errors. The corpus reveals several common types of behaviors as discussed below:

1. **Character omission or elision:** Vowels and/or consonants are often omitted to shorten spellings while keeping the words intelligible. e.g.: "ঐঙোন্দা"(/ai-ŋo:n-da:/, 'to me')→"eind", "aionda", "eigonda", "eingond", etc.;
2. **Syllable substitution by an English alphabet or digits:** Substitution of a syllable in a word with an English alphabet letter/digit which resembles the syllable sound. e.g.: "সিদা"(/si-da:/, 'here')→"cda", "sida", etc.
3. **Loanwords, Code-Mixed Words, and Named Entities:** The corpus also contains loanwords, code-mixed words, and named entities, which are often spelled phonetically or idiosyncratically, contributing further to variation in Romanization. e.g.:
   •Loanword: "বিদিও"(/bi-di-o:/,'video') → "vedio", "vdo", "vdeo",etc.
   •Code-Mixed Word: "লাস্টতা" (/la:st-ta:/, meaning "at last") → "lastta", "lasta"
   •Named Entity: "মণিপুর"(/ma-ni-pur/, 'Manipur') → "mnipur", "maniipur", "manupur", "mainpur",etc.
4. **Stretched words:** Word spellings stretched or extended to convey emotional intensity or emphasize tone of expression. e.g.: "নৈরেহে"(/nai-re:-he:/, 'that's cool! / that's awesome!')→"neirhhhhhh", "neireheeee", etc.
5. **Use of Digits for Repetition, Sound, or Numerical Meaning:** Digits are often used creatively in Romanized Manipuri—to mark reduplication (e.g.: "ফজফজবা"(/fa-ja-fa-ja-ba:/, 'beautiful ones') → "fj2ba", "faja²ba"), replace similar-sounding syllables (e.g.: "ফোতো"(/p^ho:-to:/, 'photo') → "4to", "foto", etc.), or represent actual numbers (e.g.: "অমদা"(/a-ma-da:/) → "1da", "1mda", "amada", etc.).
6. **Space-Separated Words:** Some users split words into sub-units; e.g.,"চাওখৎতবা"(/ca:u-k^hat-ta-ba:/, 'underdeveloped')→"chau khat ta ba".
7. **Other Unintentional mistakes:** i) Wrong character order. e.g.,"মিওই"(/mi-o:i/, 'person; human') → "meioo". ii) Accidentally adding extra characters. iii) Replacement of a character by its nearby character in the qwerty keyboard. iv) Words modified automatically by the spell-check or autocorrect function on the user's device. e.g., "মথেলেশ"(/ma-t^he:l-se:/, 'this side dish') → "math else", and v) Other typos.

5. Comparison with Formal Romanization Conventions

We compared social media romanization patterns with formal transliteration conventions across two categories: (1) Documented standard schemes such as ITRANS, PH[6] , BH[7] , and ISO 15919, and (2) De facto conventions found in published texts such as Manipuri Dictionary (Webonary)[8] , Manipuri Grammar (Primrose)[9] , Dictionary of Manipuri Medical Terms (Sharma)[10] , and NCERT Bhasha Sangam Manipuri[11] . The two groups differ notably. Standard schemes emphasize linguistic precision through case sensitivity and diacritics (e.g., dot, tilde, semicolon), whereas book-based conventions use simpler, symbol-free representations—more suitable for practical use. While consonant mappings are mostly consistent, small variations exist: for example, 'ভ' appears as 'bh' or 'v' in books but only 'bh' in formal schemes; 'চ' as 'ch' in books versus 'c'/'ch' in standards; and 'জ' as 'j'/'z' in books versus 'j'/'y' in standards. Similarly, vowel diphthongs in books occasionally use the caret (ˆ) to indicate intensity, unlike symbol-heavy standard

[6] https://phtranslator.sourceforge.net
[7] https://www.baraha.com/help/Keyboards/ben-phonetic.htm
[8] https://www.webonary.org/manipuri/browse/
[9] https://archive.org/details/amanipurigramma00primgoog/page/n16/mode/2up
[10] https://archive.org/details/dli.language.1325/mode/2up?view=theater
[11] https://ncert.nic.in/ebsb/12_Manipuri.pdf

Text Source	Cons.	Vowel-Mo.	Vowel-Di.
Dictionary	1.16	1.78	7.22
Grammar	0.27	1.94	8.10
Medical terms	0.40	1.61	6.50
NCERT	1.13	1.55	6.44

Table 7: KL divergence between social media text and formal text sources' romanization patterns, with formal sources as baseline.

schemes. Overall, book-based conventions better reflect realistic, user-adopted romanization, while standards like ISO 15919 remain largely theoretical. Hence, we use the book conventions as the base for quantitative comparison. We measured the deviation of social media romanization from this base using Kullback–Leibler (KL) divergence: $D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}$, where P is the romanization distribution in our corpus and Q is that from formal texts. Results (Table 7) show that vowel diphthongs have the highest divergence, followed by monophthongs, while consonants show the lowest—indicating that vowel romanizations in social media are significantly more variable, whereas consonant mappings remain relatively stable due to closer phonetic alignment.

Similar patterns of orthographic variability have also been observed in Romanized social media text of Assamese. For instance, studies on Romanized Assamese (Baruah et al., 2024b) report greater divergence from standardized romanization for vowel representations compared to consonants. Although their analysis uses similarity measures rather than KL divergence, the findings are consistent with our observation that vowel representations exhibit higher variability in informal Romanized writing.

6. Model Evaluation

The model evaluation aims to empirically examine how the high variation and user-generated, noisy nature of Romanized Manipuri text affects back-transliteration performance. Specifically, we evaluate the model's ability to generalize across diverse and irregular patterns in the dataset, in order to understand how such variability influences learning and to identify the model's strengths and limitations. To this end, we employ the Transformer architecture (Vaswani et al., 2017), a state-of-the-art architecture for sequence-to-sequence modeling. The model is trained entirely from scratch to observe how a standard, robust architecture performs when exposed solely to this noisy and irregular dataset.

6.1. Evaluation Setup

Dataset. The back-transliteration task follows a direct spelling-based mapping strategy, where each Romanized source word is paired with its corresponding Manipuri-Bengali target word. The data is divided into training, validation, and test sets in an 80%–10%–10% ratio.

Model Architecture. We employ an encoder–decoder Transformer architecture implemented using the Hugging Face Transformers library. Two variants are trained: a character-level model and a subword-level model. For both variants, a shared tokenizer is trained on the concatenated source (Romanized) and target (Manipuri–Bengali script) corpora, resulting in a unified vocabulary used by both the encoder and the decoder. Preliminary experiments on the character-level variant indicated that using separate source and target tokenizers resulted in slightly lower performance; therefore, we adopt the shared tokenizer configuration for all reported experiments. The character-level tokenizer contains 111 tokens, while the subword-level tokenizer uses Byte Pair Encoding (BPE) with a vocabulary size of 3,000. Each model comprises six encoder and decoder layers, four attention heads, a hidden size of 256, a feed-forward dimension of 512, and GELU activation (Hendrycks and Gimpel, 2023), with an attention dropout of 0.1 and a maximum sequence length of 128. The models are trained in a purely sequence-to-sequence manner without the use of any external lexicon or dictionary resources, with predictions generated directly from the learned character or subword mappings in the training data.

Training Details. Training is performed using the Seq2SeqTrainer from Hugging Face with a batch size of 16, a learning rate of 5×10^{-4}, and the AdamW optimizer over 101 epochs. The best checkpoint (epoch 26) is selected based on validation loss, with the top three checkpoints retained.

Evaluation Metrics. Performance is evaluated using Exact Match (EM)(Rajpurkar et al., 2016), Character Accuracy (CA), and the Bilingual Evaluation Understudy (BLEU)(Papineni et al., 2002) score. EM measures the proportion of words that exactly match their reference forms. CA, derived from the Character Error Rate (CER)(Jurafsky and Martin, 2025), is computed as $\mathrm{CA} = (1 - \mathrm{CER}) \times 100\%$, where CER accounts for character-level substitutions, deletions, and insertions. BLEU measures the precision of n-grams in the system output relative to the reference. For this back-transliteration task, BLEU is computed at the character level, where each character is treated as a token, and n-gram precision is calculated up to 4-grams.

6.2. Performance Result and Error Analysis

Evaluation on the test set shows that the character-level Transformer model outperforms the subword-

Model with Parameter size	EM(%)	CA(%)	BLEU(%)
Char.-level model(8.16M)	72.76	93.15	87.85
Sub.-level model(9.65M)	66.92	90.53	84.05

Table 8: Test set performance of the character-level and subword-level Transformer models.

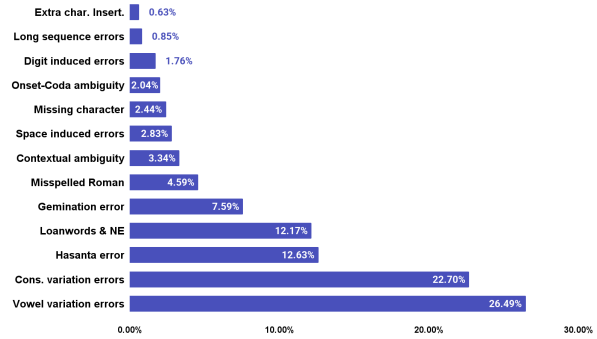

Figure 1: Quantitative Error Analysis

level variant in all metric(Table 8). A detailed analysis of the test set predictions reveals that the model performs well on words whose spellings are close to their canonical forms, as well as on shortened words that omit or elide vowels. It also handles short words effectively and performs well on words that employ common syllable substitutions, such as using the letter 'c' for the syllables 'সি' or 'শি'. In addition, the model shows strong performance on squared words-forms that use the digit '2' to indicate repetition.

Conversely, an examination of the erroneous predictions highlights several underlying sources of back-transliteration errors. Figure 1 presents the overall distribution of these error types, which are discussed in detail below.

1. **Vowel variation errors:** Many mismatches occur with the presence of vowels which maps to multiple native phonemes (see Tables 5 and 6). e.g.: "kalakcheiya" → ref.[12]:"কলকৈচেয়ে"(/ka-lak-cei-ye:/, '(I) envy that') → hyp.[13]:"কলকৈচেয়া"(/ka-lak-cei-ya:/; an invalid word) — here 'a' is realized as /e:/ in the ref. but as /a:/ in the hyp.
   A frequent source of this confusion is the 'ো' diacritic (/ə/), which has no consistent Roman representation and is often inconsistently rendered with 'a'. e.g.: "katando" → ref.: "কতন্দো"(/ka-tan-do:/, 'that lazy person') → hyp.: "কাতন্দো"(/ka:-tan-do:/; deviates from the correct spelling and pronunciation, though still intelligible).
2. **Consonant Variation Errors:** Roman consonant letters with high back-transliteration variability—i.e., those that map to multiple phonemes—are a major source of errors. e.g.: "ecai" → ref.: "ইশৈ"(/i-sai/, 'song') → hyp.: "ইচাই"(/i-ca:i/). This confusion arises from the multiple native-script correspondences of the Roman letter 'c', which can represent several phonemes: 'c' → /c/ (চ), /j/ (জ), /k/ (ক), and /s/ (স, শ).
3. **'Hasanta'-related errors:** In the Manipuri–Bengali script, each consonant inherently carries the vowel /a/ (IPA: /ə/); For e.g., 'ক' is pronounced /ka/. The Hasanta diacritic suppresses this vowel, forming a consonant-ending syllable, e.g., english word "take" → "তেক্". However, its application is inconsistent—sometimes omitted for heavier or paused articulation—making Hasanta usage unpredictable and causing irregular back-transliteration outputs.
4. **Loanword and Named-Entity errors:** Errors often occur when transliterating borrowed or mixed-language words, especially when pronunciation diverges from spelling. English borrowings pose greater difficulty due to irregular orthography, while Indo-Iranian loanwords are comparatively easier. English-Manipuri code-mixed word e.g.: "minutese" → ref.:"মিনৎসে"(/mi-nat-se:/, 'that/this minute') → hyp.:"মিনুতেশে"(/mi-nu-te:-se:/); e.g. of Loanword from Hindi: "chowkidar" → ref.:"চাউকিদার"(/ca:o-ki-da:r/, 'watchman; gatekeeper')→"চাওকিদরা"(/ca:o-ki-da-ra:/).
5. **Gemination Errors:** Doubled consonants in Romanized words are not consistently mapped to conjunct or geminated consonants in the native script, leading to back-transliteration inconsistencies. Two main cases are observed: (i) only one of the doubled consonant letters is transliterated, and (ii) both are transliterated separately but not combined into a geminate form. e.g.: Case (i): "ummy" → ref.: উম্মি(/um-mi/, 'keeping something in the mouth') → hyp.: উমাই(/u-ma:i/). Case (ii): "nammi" → ref.: নম্মি(/nam-mi/, 'smelling something') → hyp.: নমমি(/na-ma-mi/);
6. **Misspelled Roman words:** The model often reproduces misspelled or shortened Roman inputs, leading to mismatches with the correct reference. In the e.g., 'miounda" → ref.:"মিঙোন্দা"(/mi-ŋo:n-da:/, 'to other(s)') → hyp.:"মিওন্দা"(/mi-on-da:/), "miounda" represents an informal spoken pronunciation used by native speakers for the word "mingonda"(/mi-ŋon-da:/).
7. **Contextual Ambiguity:** Some words allow multiple valid transliterations, with the correct form depending on context. For instance, "pichabg" may be পিজবগী(/pi-ja-ba-gi/) (meaning:"due to feeding") or পিজবগা(/pi-ja-ba-ga:/) ("meaning:on feeding"), where the final 'g'—a shortened form of 'ga', 'gi' or 'ge'—requires surrounding context to resolve. This causes valid but ground-truth-mismatched predictions.
8. **Space induced errors:** Users often insert or

[12]ref. denotes the gold target
[13]hyp. denotes the model's prediction

omit spaces inconsistently in Romanized words, even when unnecessary for forming a valid native word. This creates ambiguity in the dataset, making it difficult for the model to decide when to preserve or remove spaces. E.g., of word with space inserted: "marak tagesu" → ref.:"মরক্তগীশু"(/ma-rak-ta-gi-su/, 'also from among/between') → hyp.:"মরক্ তশু"(/ma-rak ta-su/, 'also among; also in between'); E.g., of word with space omitted: "echilnaosingda" → ref.:"ইচিল্ ইনাওশিংদা"(/e-cil e-na:o-siŋ-da:/, 'to brothers and sisters') →"ইচিল্নাওশিংদা"(/e-cil-na:o-siŋ-da:/, a shortened form intelligible to native speakers and conveying the same meaning as the previous form).

9. **Missing character errors:** A recurring error pattern is omission of medial characters, whereby the model generates a shorter, simplified hypothesis from the source word. For e.g.,"thdokrrni" → ref.:"থাদোকররনি"(/t^ha:-do:k-ra-ra-ni/, 'may be released; likely to be released') → hyp.:"থাদোরনি"(/t^ha:-do:-ra-ni/, an invalid word). Here, it appears that the middle grapheme 'kr' is omitted during back-transliteration by the model.
10. **Onset–Coda Ambiguity:** Certain Roman consonants and digraphs introduce ambiguity at syllable boundaries, making it unclear whether they function as codas of the preceding syllable or onsets of the following one, which can lead to back-transliteration errors. For instance:
    • The Roman grapheme 'k' may act either as the coda of a syllable or the onset of the next, affecting vowel realization. e.g., skaga — ref.: শক্অগা (/sak-a-ga/, 'by singing') — hyp.: শকাগা (/sa-ka-ga/, an invalid word).
    • The digraph 'ng' may represent either a single nasal phoneme /ŋ/ or a consonant sequence (/n/-/g/).
    • The Roman grapheme 'h' may function as an independent onset or as part of an aspirated consonant.
11. **Digit-Induced Errors:** Although the model performs well in predicting repeated words denoted by the digit "2", it struggles with cases where digits represent literal numbers or are used as syllable substitutes. Its performance is notably weaker for digits other than "2," likely due to their limited presence in the training dataset. e.g.:
    •"2umamdaida" → ref.: "তুম্মন্দাইদাই"(/tum-mam-da:i-da:/, 'when about to sleep; when about to be exhausted') → hyp.: "অনন্দাইদা"(/a-nam-da:i-da:/, an invalid word);
    •"50gi" → ref.: "৫০ গী"(numeric figure in native script /gi/, 'for 50') → hyp.: "অনিগী"(/a-ni-gi/, 'for two').

While our empirical analysis is grounded in Manipuri, many of the observed variation patterns and corresponding model error types—such as vowel instability, digit-based substitutions, phonologically motivated grapheme variation, and expressive stretching—closely resemble phenomena documented in other Romanized social media contexts. This suggests that the challenges in back-transliteration arise from general patterns of informal Romanized writing rather than from unique features of a specific language.

7. Conclusion and Future Work

This study presented the first large-scale analysis of Romanized Manipuri social media text. A Romanized Manipuri to Manipuri-Bengali script back-transliteration corpus was constructed and analyzed using Zipf's and Heaps' laws, confirming natural language-like lexical behavior. Variation analysis revealed two major sources: character-level inconsistencies, where vowels showed broader and less skewed distributions than stable consonants, and pragmatic-stylistic variations shaped by user writing habits. Comparison with formal transliteration schemes showed that vowel diphthongs deviate most from standardized forms. Evaluation of the Transformer model demonstrated that character-level models outperform subword-level ones, though challenges persist due to noise, digit use, and inconsistent spellings, with most errors linked to vowel grapheme mismatches. More broadly, this study highlights systematic sources of variation and transliteration errors that are likely to arise in Romanized social media text for low-resource languages.

In future work, we plan to evaluate pretrained back-transliteration and multilingual language models on the dataset and explore domain-adaptive fine-tuning and error-aware training to improve back-transliteration robustness for Manipuri social media text. Incorporating lexicon-based constraints or dictionary resources may also help improve prediction accuracy.

8. Acknowledgements

This dataset was developed at the Open Source Intelligence Lab, Indian Institute of Technology Guwahati, and was partially supported by the Ministry of Electronics & Information Technology, Government of India. We thank the annotators for their contributions to the annotation process.

References

Umair Z Ahmed, Kalika Bali, Monojit Choudhury, and Sowmya VB. 2011. Challenges in designing input method editors for Indian languages:

The role of word-origin and context. In *Proceedings of the Workshop on Advances in Text Input Methods (WTIM 2011)*, pages 1–9, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Hemanta Baruah, Sanasam Ranbir Singh, and Priyankoo Sarmah. 2024a. Assamesebacktranslit: Back transliteration of romanized assamese social media text. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1627–1637.

Hemanta Baruah, Sanasam Ranbir Singh, and Priyankoo Sarmah. 2024b. Transliteration characteristics in romanized assamese language social media text and machine transliteration. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(2):1–36.

Irshad Ahmad Bhat, Vandan Mujadia, Aniruddha Tammewar, Riyaz Ahmad Bhat, and Manish Shrivastava. 2014. Iiit-h system submission for fire2014 shared task on transliterated search. In *Proceedings of the 6th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '14, page 48–53, New York, NY, USA. Association for Computing Machinery.

Ann Bies, Zhiyi Song, Mohamed Maamouri, Stephen Grimes, Haejoong Lee, Jonathan Wright, Stephanie Strassel, Nizar Habash, Ramy Eskander, and Owen Rambow. 2014. Transliteration of arabizi into arabic orthography: Developing a parallel annotated arabizi-arabic script sms/chat corpus. In *Proceedings of the EMNLP 2014 workshop on Arabic natural language processing (ANLP)*, pages 93–103.

Kunal Chakma and Amitava Das. 2014. Revisiting automatic transliteration problem for code-mixed romanized indian social media text. *Social-India*, 2014:42.

Md Fahim, Fariha Shifat, Fabiha Haider, Deeparghya Barua, Md Sourove, Md Ishmam, and Md Bhuiyan. 2024. BanglaTLit: A benchmark dataset for back-transliteration of Romanized Bangla. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14656–14672.

Philip Gage. 1994. A new algorithm for data compression. *C Users Journal*, 12(2):23–38.

Nizar Habash, Mona Diab, and Owen Rambow. 2012. Conventional orthography for dialectal Arabic. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 711–718, Istanbul, Turkey. European Language Resources Association (ELRA).

Harold Stanley Heaps. 1978. *Information Retrieval: Computational and Theoretical Aspects*. Academic Press, New York.

Dan Hendrycks and Kevin Gimpel. 2023. Gaussian error linear units (gelus).

Daniel Jurafsky and James H. Martin. 2025. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd edition. Online manuscript released August 24, 2025.

Sarvnaz Karimi, Falk Scholer, and Andrew Turpin. 2011. Machine transliteration survey. *ACM Computing Surveys (CSUR)*, 43(3):1–46.

Kamaljeet Kaur and Parminder Singh. 2014. Review of machine transliteration techniques. *International Journal of Computer Applications*, 107(20):13–16.

Lenin Laitonjam, Loitongbam Gyanendro Singh, and Sanasam Ranbir Singh. 2018. Transliteration of english loanwords and named-entities to manipuri: Phoneme vs grapheme representation. In *2018 International Conference on Asian Language Processing (IALP)*, pages 255–260. IEEE.

Lenin Laitonjam and Sanasam Ranbir Singh. 2022. A hybrid machine transliteration model based on multi-source encoder–decoder framework: English to manipuri. *SN Computer Science*, 3(2):125.

WMP Liwera and L Ranathunga. 2020. Combination of trigram and rule-based model for singlish to sinhala transliteration by focusing social media text. In *2020 From Innovation to Impact (FITI)*, volume 1, pages 1–5. IEEE.

Yash Madhani, Sushane Parthan, Priyanka Bedekar, Gokul Nc, Ruchi Khapra, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M Khapra. 2022. Aksharantar: Open indic-language transliteration datasets and models for the next billion users. *arXiv preprint arXiv:2205.03018*, pages 40–57.

Sabina Mammadzada. 2023. A review of existing transliteration approaches and methods. *International Journal of Multilingualism*, 20(3):1052–1066.

Jayashree Nair and Riyaz Ahammed. 2021. English to indian language and back transliteration with

phonetic transcription for computational linguistics tools based on conventional transliteration schemes. In *2021 Fourth International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, pages 1–6.

Kishorjit Nongmeikapam and Sivaji Bandyopadhyay. 2012. A transliteration of crf based manipuri pos tagging. *Procedia Technology*, 6:582–589.

Kishorjit Nongmeikapam, Ningombam Herojit Singh, Sonia Thoudam, and Sivaji Bandyopadhyay. 2011. Manipuri transliteration from bengali script to meitei mayek: A rule based approach. In *International Conference on Information Systems for Indian Languages*, pages 195–198. Springer.

Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, Pankaj Kundan Dadure, and Sandeep Kumar Dash. 2023. Findings of the wmt 2023 shared task on low-resource indic language translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 682–694.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Brian Roark, Lawrence Wolf-Sonkin, Christo Kirov, Sabrina J. Mielke, Cibu Johny, Isin Demirsahin, and Keith Hall. 2020. Processing South Asian languages written in the Latin script: the Dakshina dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2413–2423, Marseille, France. European Language Resources Association.

Inder Singh. 1975. *Manipuri phonetic reader*. Central Institute of Indian Languages, Mysore.

Leihaorambam Sarbajit Singh, Kabita Thaoroijam, and Pradip Kumar Das. 2007. Written manipuri (meiteiron)–phoneme to grapheme correspondence. *Language in India Strength for Today and Bright Hope for Tomorrow Volume 7 : 6 June 2007*.

Thoudam Doren Singh. 2012. Bidirectional bengali script and meetei mayek transliteration of web based manipuri news corpus. In *Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing*, pages 181–190.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

MK Vathsala and Ganga Holi. 2020. Rnn based machine translation and transliteration for twitter data. *International Journal of Speech Technology*, 23(3):499–504.

George Kingsley Zipf. 1949. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley Press, Cambridge, MA.