ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25dec1243

Regional Language to Bangla Joypuhat, Rajshahi, Bangladesh

Md. Abu Horaira Sarder[1]

[1]Daffodil International University
Dhaka, Bangladesh

Publication Date: 2026/01/02

Abstract: Bangladesh and its adjoining regions exhibit extensive linguistic diversity, comprising numerous regional languages and dialects that remain underrepresented in digital communication systems. The absence of standardized translation frameworks for these regional varieties poses substantial barriers to information accessibility, knowledge dissemination, and inclusive technological development. This study proposes an NLP-based computational model for systematically translating regional languages into Standard Bangla, thereby addressing the linguistic gap between informal spoken varieties and formal written Bangla. The research methodology encompasses corpus development, data annotation, text normalization, tokenization, phonological mapping, and the application of machine-learning and sequence-to-sequence translation architectures. A parallel dataset consisting of region-specific lexical items, syntactic structures, and semantic patterns was constructed to train and evaluate the system. Experimental evaluation indicates that the proposed model achieves promising translation accuracy while preserving semantic integrity and contextual meaning. The findings highlight the system's potential to support language standardization, promote linguistic inclusivity, and facilitate broader digital participation among speakers of marginalized dialects. The study further advances localized NLP research in Bangladesh and provides a foundation for future extensions to educational technology, governmental communication platforms, and multilingual AI systems.

How to Cite: Md. Abu Horaira Sarder (2025) Regional Language to Bangla Joypuhat, Rajshahi, Bangladesh. *International Journal of Innovative Science and Research Technology*, 10(12), 2253-2258. https://doi.org/10.38124/ijisrt/25dec1243

I. INTRODUCTION

Bangladesh’s linguistic landscape is characterized by a rich tapestry of regional languages and dialects that differ substantially from Standard Bangla. Among these, the regional dialects spoken in Joypurhat, a district in the Rajshahi Division, exhibit unique phonetic patterns, lexical structures, and cultural expressions deeply rooted in local heritage. Although Joypurhat belongs to the broader North Bengal linguistic zone, its speech varieties reflect distinct influences shaped by geographical proximity, community practices, and historical interactions. These linguistic features often diverge noticeably from Standard Bangla, making communication challenging in formal, educational, and administrative settings. Despite the widespread use of local dialects in daily interactions, most digital communication platforms, educational resources, government services, and media content in Bangladesh primarily rely on Standard Bangla. As a result, speakers of the Joypurhat dialect, particularly those from rural and semi-urban communities, may face barriers to accessing essential information, interpreting official documents, or using digital services. This linguistic gap becomes even more pronounced in sectors such as healthcare communication, disaster alerts, e-governance, and academic learning, where clarity and comprehension are crucial. Current Natural Language Processing (NLP) technologies in Bangladesh remain highly underdeveloped for regional dialects, including those of the Rajshahi region. The limited availability of linguistic corpora, the lack of standardized orthography, and the low-resource nature of these dialects pose significant challenges for machine translation and data-driven modeling. Consequently, the development of computational tools capable of translating regional dialects such as that of Joypurhat into Standard Bangla is critically essential for enhancing linguistic inclusivity and ensuring equitable access to information. This research aims to develop an NLP-based translation framework to convert the regional dialect of Joypurhat into Standard Bangla. The study involves collecting authentic speech samples and text data from local speakers, developing a parallel corpus, performing linguistic normalization, and applying machine-learning and deep-learning translation techniques. By capturing unique phonological features and context-dependent expressions used in Joypurhat, the proposed system seeks to produce translations that are both accurate and culturally meaningful. The significance of this work extends beyond language translation. It contributes to preserving regional linguistic identity, reducing communication barriers, and promoting digital inclusivity for marginalized dialect communities. Moreover, the study provides a foundation for future research on low-resource language processing in Bangladesh, enabling advances in speech recognition, educational tools, localized AI systems, and accessible public information platforms. Ultimately, this research aspires to bridge the gap between regional dialects and Standard Bangla, thereby supporting more inclusive communication and social participation in districts such as Joypurhat and across the Rajshahi region.

II. METHODOLOGY

This section describes the methodology used to develop a Regional Language-to-Bangla Translation System for the Joypurhat dialect of the Rajshahi Division, Bangladesh. The proposed approach follows a data-driven experimental framework that employs Natural Language Processing (NLP) and Neural Machine Translation (NMT) techniques to handle low-resource, phonetic, and non-standard dialectal text.

➢ *Research Design*

The study employs an experimental research design that integrates corpus creation, machinelearning-based translation modeling, and quantitative evaluation. The primary objective is to translate expressions from the Joypurhat regional dialect into fluent, semantically correct Standard Bangla.

➢ *Dataset Collection and Corpus Development*

A parallel corpus was created by collecting dialect expressions from native speakers of the Joypurhat region through informal conversations and daily spoken usage. Each dataset entry comprises a sentence in the Joypurhat dialect and its manually verified Standard Bangla translation.

Example:

- Regional Text: কোটে আচু
- Bangla Text: কোথায় আছিস

The dataset was stored in CSV format and represents authentic phonetic and conversational language usage.

➢ *Data Preprocessing*

Due to the informal and inconsistent nature of the dialect, extensive preprocessing was applied:

- Text Cleaning: Removal of punctuation, noise, and extra whitespace

- Unicode Normalization: Ensuring consistent Bangla encoding

- Phonetic Normalization: Handling multiple spellings of the same dialect word

- Tokenization: Custom tokenization preserving dialect-specific morphemes such as "চু" and "লু"

➢ *Feature Representation*

To convert text into numerical representations, the following techniques were used:

- TF-IDF Vectorization: Applied for cosine similarity computation and linguistic analysis

- Word Embeddings: Word2Vec/FastText embeddings to capture semantic similarity

- Contextual Embeddings**:** Transformer-based embeddings (mBERT/XLM-R) for contextual understanding

➢ *Translation Model Architecture*

Three neural architectures were implemented and compared:

- Seq2Seq with Attention: Encoder–decoder model suitable for short dialect phrases

- Transformer-based NMT: Self-attention mechanism enabling robust handling of noisy dialect input

- Adversarial Neural Machine Translation: Generator–discriminator framework to improve fluency and naturalness Among these, the Transformer-based model demonstrated the best overall performance.

➢ *Model Training*

The dataset was divided into 70% training, 15% validation, and 15% testing sets. Training configurations included:
Optimizer: Adam
Loss Function: Cross-Entropy
Regularization: Early stopping to prevent overfitting

➢ *Evaluation Metrics*

Model performance was evaluated using both automatic and human-based metrics.

- *Automatic Metrics*

BLEU
ROUGE
METEOR
Cosine Similarity

- *Translation Quality Score (TQS)*

TQS=0.4A+0.3F+0.3CSTQS = 0.4A + 0.3F + 0.3CSTQS=0.4A+0.3F+0.3CS where AAA denotes adequacy, FFF denotes fluency, and CSCSCS denotes cosine similarity.

- *Human Evaluation*

Native Bangla speakers evaluated translation quality with respect to meaning preservation, grammatical correctness, and fluency.

➢ *Experimental Analysis*

Comparative analysis showed that Transformer-based and Adversarial models outperform Seq2Seq models, particularly in handling phonetic spelling variations and preserving semantic meaning in low-resource dialect translation.

➢ *Tools and Implementation*

- Programming Language: Python
- Libraries: TensorFlow / PyTorch, Scikit-learn, NLTK
- Platform: Google Colab Data Format: CSV

➢ *Ethical Considerations*

All data were collected with informed consent. No personal or sensitive information was included. The study aims to preserve linguistic identity while ensuring respectful and ethical use of regional language data.

III. RESULTS

This section presents the experimental results of the proposed Regional Language-to-Bangla Translation System, evaluated on the Joypurhat dialect dataset. Performance is assessed using standard automatic metrics and human evaluation to measure accuracy, semantic preservation, and fluency.

➢ *Dataset and Experimental Setup*

Experiments were conducted on a curated parallel corpus comprising expressions in the Joypurhat regional dialect and their corresponding Standard Bangla translations. The dataset contains short, phonetic, and meaning-based expressions commonly used in daily communication. The corpus was split into 70% training, 15% validation, and 15% testing sets.

➢ *Translation Performance (BLEU)*

Table 1 BLEU Scores for Different Models

Model	**BLEU-1**	**BLEU-2**	**BLEU-4**
Seq2Seq	0.61	0.48	0.34
Transformer	0.74	0.66	0.52
Adversarial Transformer	**0.81**	**0.73**	**0.59**

➢ *Observation*

Transformer-based models outperform the Seq2Seq baseline. Incorporating adversarial learning further improves translation accuracy, particularly for noisy dialect inputs.

➢ *ROUGE and METEOR Evaluation*

Table 2 ROUGE and METEOR Scores

Model	**ROUGE-1**	**ROUGE-L**	**METEOR**
Seq2Seq	0.58	0.53	0.55
Transformer	0.70	0.65	0.68
Adversarial Transformer	**0.76**	**0.71**	**0.73**

➢ *Interpretation*

Higher ROUGE and METEOR scores indicate improved content preservation and semantic alignment for dialect-to-Bangla translation.

➢ *Semantic Similarity Analysis*

Cosine similarity between generated translations and reference Bangla sentences was computed using TF-IDF representations.

Table 3 Average Cosine Similarity

Model	Cosine Similarity
Seq2Seq	0.72
Transformer	0.81
Adversarial Transformer	**0.87**

- *Observation:*

The adversarial Transformer achieves the highest semantic similarity, confirming strong meaning preservation.

- *Translation Quality Score (TQS)*

The overall translation quality was measured using the Translation Quality Score (TQS): TQS=0.4A+0.3F+0.3CSTQS = 0.4A + 0.3F + 0.3CSTQS=0.4A+0.3F+0.3CS where AAA denotes adequacy, FFF fluency, and CSCSCS cosine similarity.

Table 4 TQS Evaluation

Model	Adequacy	Fluency	Cosine	TQS
Seq2Seq	0.78	0.73	0.72	0.74
Transformer	0.90	0.88	0.81	0.87
Adversarial Transformer	**0.94**	**0.92**	**0.87**	**0.92**

- *Interpretation*

The adversarial Transformer attains a high-quality translation score (TQS > 0.90), indicating excellent adequacy and fluency.

- *Qualitative Evaluation*

Human evaluation by native speakers indicates that **92%** of translations preserve meaning and **89%** are fluent in Standard Bangla. Dialect-specific expressions are accurately interpreted in most cases.

- *Error Analysis*

Primary error sources include (i) rare dialect expressions with limited training coverage, (ii) overgeneralization of frequent phonetic patterns, and (iii) occasional loss of pragmatic nuance in longer inputs. These errors are reduced in the adversarial model compared to baselines.

IV. CONCLUSION

This research sought to address a critical linguistic and technological gap in Bangladeshi Natural Language Processing (NLP): the absence of computational tools and datasets for translating the Joypurhat regional dialect into Standard Bangla. The Joypurhat dialect, rooted in the Rajshahi Division, represents a distinct linguistic identity shaped by rural communication patterns, cultural expressions, and oral storytelling traditions. Despite its widespread use among thousands of native speakers, the dialect remained largely undocumented in digital form and absent from existing national-level NLP resources. This study presents the first structured digital dataset and translation framework for this dialect, thereby contributing both academically and socially to the field of regional language processing. A significant achievement of this work is the creation and curation of a high-quality parallel corpus comprising expressions in the Joypurhat dialect and their Standard Bangla equivalents. The dataset captures meaningful examples, including highly phonetic forms, simplified verb structures, short conversational utterances, and region-specific vocabulary that are rarely represented in mainstream Bangla corpora. Through preprocessing steps such as normalization, tokenization, dialect-specific cleaning, and semantic alignment, the dataset was transformed into a reliable resource suitable for machine learning and linguistic analysis. This dataset not only supports model training but also serves as an essential record of linguistic heritage.

The methodological framework proposed in this research integrates Region Detection and Dialectto-Bangla Translation, enabling a comprehensive understanding of the patterns of the Joypurhat dialect. Multiple neural architectures, including Seq2Seq with Attention, Transformer-based NMT, and Adversarial NMT, were evaluated to determine the most effective approach for this low-resource, highly phonetic dialect. The experimental results show that the Adversarial Transformer model consistently outperformed traditional models, demonstrating superior fluency, semantic accuracy, and robustness when handling noisy phonetic variations such as "কেিুলু", "ঘুমোচু", and "ক োটে আচু". The high scores across BLEU, ROUGE, METEOR, Cosine Similarity, and the Translation Quality Score (TQS) confirm the strength of this approach. This research also highlights the significant potential for real-world applicability. The developed system can facilitate communication across educational, administrative, and healthcare settings. For example, it can help rural students understand academic Bangla content, enable government officers to interpret dialect-based queries more accurately, and assist healthcare providers in understanding patient speech. Moreover, integrating the translation system into mobile apps, chatbots, and digital government platforms can advance the national vision of an inclusive "Digital Bangladesh," where technology accommodates all linguistic variations.

Beyond its immediate utility for translation, the work makes an essential contribution to cultural preservation and linguistic diversity. Documenting the Joypurhat dialect in a machine-readable format ensures that its unique expressions, phonetic styles, and semantic patterns are preserved for future generations. It also encourages academic interest in other underrepresented dialects of Bangladesh, such as Bogura, Naogaon, and Rangpur variants, expanding the scope of dialectal NLP research.

However, the research also acknowledges several limitations. The current dataset is primarily limited to short, phrase-level expressions and does not include longer contextual sentences, natural conversations, or spoken audio samples. Additionally, the dataset does not yet distinguish dialectal variation across the five Upazilas of Joypurhat (Sadar, Akkelpur, Kalai, Khetlal, Panchbibi), each of which may exhibit subtle phonetic or lexical differences. Addressing these gaps in future work will significantly enhance model generalizability and improve real-world translation quality. Looking forward, this thesis opens numerous avenues for expansion. Future researchers can develop speech-to-text systems, region-specific microdialect models, cross-dialect transfer learning, context-aware translation, and multimodal dialect analysis. There is also strong potential to establish a nationwide Bangla Dialect Translation Framework that enables real-time translation across all major dialects of Bangladesh. Such advancements will not only strengthen national NLP infrastructure but also empower rural and linguistically marginalized communities. In conclusion, this research establishes a computational foundation for the Joypurhat dialect, providing both linguistic documentation and an effective machine translation model. It demonstrates that even low-resource and phonetically inconsistent dialects can be modeled effectively using advanced neural architectures. Most importantly, it shows that linguistic inclusivity can be achieved through thoughtful data collection, methodological precision, and community-centered technological innovation. This thesis therefore stands as both a technological achievement and a contribution to cultural preservation, paving the way for future advancements in Bangla dialect processing and regional language technology.

REFERENCES

[1]. I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3104–3112, 2014.

[2]. D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2015.

[3]. M. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1412–1421, 2015.

[4]. A. Vaswani *et al*., "Attention is all you need," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5998–6008, 2017.

[5]. I. Goodfellow *et al*., "Generative adversarial nets," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2672–2680, 2014.

[6]. G. Lample, A. Conneau, L. Denoyer, and M. Ranzato, "Unsupervised machine translation using monolingual corpora only," in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2018.

[7]. P. Koehn, *Statistical Machine Translation*. Cambridge, U.K.: Cambridge Univ. Press, 2010.

[8]. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, pp. 4171–4186, 2019.

[9]. A. Conneau *et al*., "Unsupervised cross-lingual representation learning at scale," in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 8440–8451, 2020.

[10]. M. Lewis *et al*., "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proc. ACL*, pp. 7871–7880, 2020.

[11]. K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. ACL*, pp. 311–318, 2002.

[12]. C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. ACL Workshop on Text Summarization Branches Out*, pp. 74–81, 2004.

[13]. S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop*, pp. 65–72, 2005.

[14]. C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[15]. M. Hasan and M. S. Islam, “Bangla language processing: A survey,” *Journal of Information and Communication Technology*, vol. 19, no. 2, pp. 123–145, 2020.

[16]. S. A. Chowdhury and M. J. Alam, “Neural machine translation for Bangla–English using transformer architecture,” *Int. J. Comput. Appl.*, vol. 176, no. 10, pp. 1–7, 2019.

[17]. A. Bhattacharjee, M. S. Rahman, and S. Sarker, “BanglaBERT: Language model pretraining for Bangla language processing,” *arXiv preprint arXiv:2101.00204*, 2021.

[18]. M. J. Islam, M. T. Taher, and S. Paul, “Vashantor: A multilingual benchmark dataset for Bangla regional dialect translation,” *arXiv preprint arXiv:2303.XXXXX*, 2023.

[19]. A. Rahman and M. S. Islam, “Computational challenges in Bangla dialect processing,” *Dhaka Univ. J. Linguistics*, vol. 15, no. 1, pp. 45–60, 2022.

[20]. A. H. Author, “Regional Language to Bangla: Joypurhat dialect dataset,” Self-compiled dataset, Rajshahi Division, Bangladesh, 2025.