Alankaar: A Dataset for Figurativeness Understanding in Bangla

Geetanjali Rakshit and Jeffrey Flanigan
Computer Science and Engineering Department
UC Santa Cruz
{grakshit,jmflanig}@ucsc.edu

Abstract

Bangla has a rich written literature, automatically making it replete with examples of creative usage of language. There have been limited efforts to computationally analyze creative text in the Bangla language due to a lack of resources. We present *Alankaar*, a dataset of 2500 manually annotated examples of text fragments in Bangla containing metaphors. We also provide automatic and manual English translations of these examples. Additionally, we provide 2500 examples of non-metaphorical text in Bangla. We use this dataset to build a metaphor identification system in Bangla. We also use it as a test bed for cross-lingual metaphor translation, finding that not all metaphors translate literally across languages and there are several cultural factors at play in the translation of metaphors. We hope this will advance the field in metaphor translation research and in grounding cultural nuances at work in the process of machine translation.

1 Introduction

"I have had my invitation to this world's ***festival****, and thus my life has been blessed."*

- Rabindranath Tagore

Bangla has a rich written literature, replete with examples of creative usage of the language. However, it is also considered a low-resource language (Alam et al., 2021). There have been limited efforts to computationally analyze creative text in the Bangla language. Since it is still considered low-resource, the availability of data resources is often limited even for many commonplace NLP tasks in Bangla. Studies in metaphor in the Bangla language have been minimal, and largely been analyzed from a linguistic perspective, on a handful of examples (Patowari, 2015).

আমার কণ্ঠে সেথায় সুর কেঁপে যায় ত্রাসনে
amaar kanthe sethaay sur kenpe jaai aasone

My voice trembles in terror.

Figure 1: An example of a metaphor in Bangla, and its corresponding English translation

We present *Alankaar*, which is (to the best of our knowledge) the first metaphor identification dataset in Bangla. It also includes parallel translations in English. Multilingual editions of existing datasets in English are often created by utilizing manual annotation or automatic machine translation systems, like in the case of Question Answering (TyDi (Clark et al., 2020), MLQA (Lewis et al., 2019), XQuAD (Artetxe et al., 2019)) or Natural Language Inference (X-NLI (Conneau et al., 2018)). We combine manual annotation alongside existing automatic resources in Bangla and English to create this dataset.

Clark et al. (2020) note that the ”Translationese” approach to multilinguality (where text is translated word-for-word, i.e., "as is") doesn't capture the individuality of the language. This is true to an even greater extent in the more creative uses of language. Particularly in the case of creative expressions of language such as poetry, much of the phonetic effects such as rhyme and prosody are automatically lost in translation. Sometimes, the creativity in semantics is also lost, because expressions do not *literally translate*. For example, the literal translation of "book worm" in English would be কৃমি (*"boi krimi"*) in Bangla, but the actual term used is "boi poka", which literally translates to "book-bug" or "book insect" in English. Not all expressions translate word-for-word, but

Proceedings of Recent Advances in Natural Language Processing,pages 998–1002
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-114

creative expressions exist where cultural elements often form an innate part of creativity. We highlight these cases in our dataset.

The contributions of the paper are the following:

- The first metaphor identification dataset in Bangla. (2500 metaphorical, 2500 literal examples).
- A bilingual version of this corpus (Bangla-to-English).
- Two baselines for metaphor identification in Bangla.
- Analysis of metaphors that do not translate word-for-word and the existence of cultural influences in metaphors.

Our code and data can be found at https://github.com/geetanjali-rakshit/Alankaar/.

2 Related Work

The task of metaphor identification has predominantly been studied in English with the VU Amsterdam Metaphor Corpus and the TOEFL Native Language Identification Corpus (Leong et al., 2020) as well as in the TroFi dataset (Birke and Sarkar, 2007), MOH-X (Mohammad et al., 2016), etc. Metaphor identification datasets are available in other high resource languages as well (Mohler et al., 2016), such as Farsi, Russian, Spanish (Sánchez Bayona, 2023), Chinese (Zhang et al., 2018; Lu and Wang, 2017). It is, however, a relatively understudied problem in low-resource languages. Smaller metaphor-related do exist in Italian (Bambini et al., 2016), German and French (Lönneker-Rodman, 2008). In situations where dataset sizes are small, transfer-learning-based approaches have been proposed for multilingual metaphor identification. Tsvetkov et al. (2014) use lexical semantic features of words in metaphorical constructions to perform a model transfer approach from English to Spanish, Farsi, and Russian. Aghazadeh et al. (2022) show that pre-trained language models (PLMs) encode metaphorical knowledge in middle layers in experiments in 4 languages: English, Spanish, Russian, and Farsi. Berger (2022) propose a transfer learning-based approach to metaphor detection from English to German. Khatun et al. (2020) created a dataset of 15000 idioms in Bangla, but to our knowledge, there is no metaphor identification dataset in Bangla.

3 Dataset Creation

The dataset comes from literary text (verses) in Bangla scraped from https://www.tagoreweb.in We split the verses by line. They are not complete sentences by themselves but text fragments. We filter out text that is shorter than 3 word tokens. This left us with 6332 examples in total. Each fragment is then manually annotated with the labels: metaphorical or literal. Two annotators (native speakers of Bangla) were trained to do this annotation task by reading 5 examples each of metaphorical and non-metaphorical text. The annotators had an inter-annotator agreement of 0.74 calculated by Cohen's κ. We discard all examples where the annotators disagreed, resulting in the final dataset, which we call *Alankaar* with statistics shown in table 1.

Number of metaphorical examples	2553
Number of non-metaphorical examples	2665
Avg length of examples	5.5 words

Table 1: Data Statistics in Alankaar

4 Bilingual Metaphor Corpus

We use automatic and manual methods to create the bilingual version of *Alankaar*. Figure 2 shows the overall pipeline for creating the bilingual version (Bangla-English) of our dataset. First, we use the Bangla-to-English machine translation model from Bhattacharjee et al. (2023) to translate the Bangla examples to English. This is a T5-based model that has state-of-the-art performance on Machine Translation on the BanglaNMT corpus (Hasan et al., 2020). Next, we obtain language-agnostic BERT embeddings (LaBSE) (Feng et al., 2020) for the Bangla-English example pairs and filter out the examples with a cosine similarity lower than or equal to 0.65. For these examples, we manually create English translations. We annotate the English translation for correctness on a scale of 1-5 for the remaining examples. We keep the translations with a correctness score greater than 3 and add them to the dataset, and for the rest, we create manual English translations. This gives us the bilingual version of Alankaar. We show an example of automatically generated and manually corrected English translation in table 2.

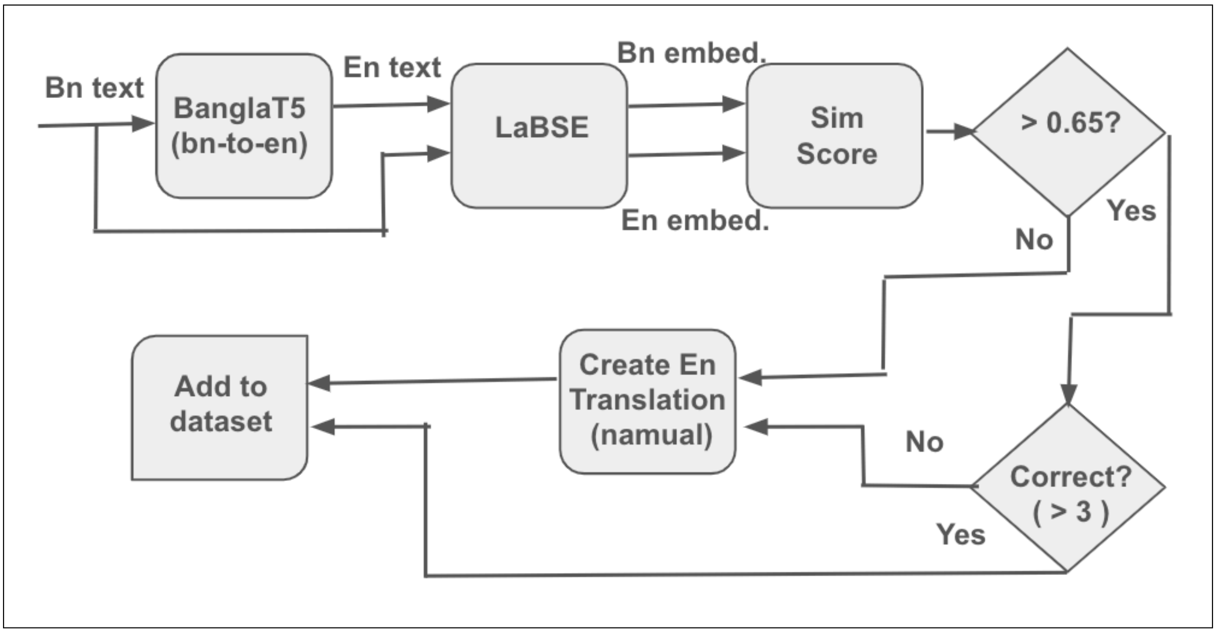

Figure 2: Pipeline for creation of Bilingual Metaphor Corpus

Bangla (Metaphorical)	বুকেচমক দিয়ে তাই তো ডাক *buuke chomok diye taai to daako*
English	Call with a shock in your chest
Corrected English	You call to me with a tremor in your chest

Table 2: An example of automatic and corrected English translation of a metaphorical example in Bangla.

5 Metaphor Identification

We provide two baselines for metaphor identification in Bangla with this dataset. We split the data into train and test (70:30), and perform 5-fold cross-validation. We train a baseline model, XLM-RoBERTa (Conneau et al., 2019), a multilingual pre-trained language model on the train split (referred to as XLM-R-Alankaar) and average over the results on the test splits. We also ran a state-of-the-art English metaphor identification model from Wachowiak et al. (2022) on the English translations. This is an XLM-RoBERTa model trained on the VUA Metaphor Corpus in English (referred to as XLM-R-VuA). The classification results are summarized in table 3. The in-domain model XLM-R-Alankaar has a better metaphor identification accuracy by 4.3 percentage points.

Model	Accuracy
XLM-R-Alankaar	76.3 %
XLM-R-VuA	72 %

Table 3: Classification Accuracy on Metaphor Identification on Alankaar

6 Metaphor Translation and Cultural Context

Not all metaphors are universal, and some may be culturally informed (Alonge, 2006). In this dataset, we mostly provide the English translation of the metaphorical context word-for-word. However, we observe that not all these translations are valid metaphors in English. We term these examples ***non-transferable*** metaphors. In table 4, we show an example that doesn't read like a complete metaphor in English without sufficient context. A possible way to address this issue is to use an equivalent or closest English version of the metaphor, instead of a "literal", word-for-word translation. We found 153 such examples in the dataset during the manual translation process. There are also metaphors in our dataset that are significantly related to cultural context and are hard to comprehend on their own in English. We refer to these as ***cultural*** metaphors. In the example in table 4, there is a reference to *flute* and *bakul* flowers, which are connected to religious, mythical and folk symbolism in the Indian subcontinent and do not

Type	Count	Example
Non-transferable	153	উজান বায়ে ফেরে যদি কে রয় সে আশায় *ujaan baaye fere jodi ke roy se ashaay* If anyone comes upstream in hope
Cultural	207	বকুলগুলি আকুল হয়ে বাঁশির গানে মুঞ্জরে *bakulguli aakul hoye baanshir gaane munjore* The bakul flower impatiently awaits the song of the flute.

Table 4: Examples of metaphors difficult to translate to English without sufficient context

directly translate in English. We came across 207 such cultural motifs in our dataset. One way to address these kinds of metaphors during translation is to provide more cultural background in the translated version, possibly as footnotes.

7 Conclusion and Future Work

We present the first metaphor identification dataset in Bangla (to the best of our knowledge). We hope this will advance research on figurative language in Bangla, focusing on metaphor identification and metaphor translation. In creating this dataset, all the text is collected from the same source. We intend to augment this dataset by adding more examples from other sources to make it more diverse. We would also like to expand the annotation scheme to label metaphoricity at the word level. While the overall goal is to create a high-quality metaphor resource in the Bangla language, this dataset can be used as a test set for any transfer-learning-based approaches for metaphor identification across languages. An important takeaway from doing the manual translations is that sufficient background information may be beneficial to appropriately handle cultural references in multilingual text.

Limitations

The proposed dataset captures a range of metaphors limited to the literary domain. It lacks examples of colloquial metaphors in Bangla. It may also be limited in examples of more modern metaphorical expressions.

References

Ehsan Aghazadeh, Mohsen Fayyaz, and Yadollah Yaghoobzadeh. 2022. Metaphors in pre-trained language models: Probing and generalization across datasets and languages. *arXiv preprint arXiv:2203.14139*.

Firoj Alam, Arid Hasan, Tanvirul Alam, Akib Khan, Janntatul Tajrin, Naira Khan, and Shammur Absar Chowdhury. 2021. A review of bangla natural language processing tasks and the utility of transformer models. *arXiv preprint arXiv:2107.03844*.

Antonietta Alonge. 2006. The italian metaphor database. In *LREC*, pages 455–460.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *arXiv preprint arXiv:1910.11856*.

Valentina Bambini, Chiara Bertini, Walter Schaeken, and Francesco Di Russo. 2016. Disentangling metaphor from context: an erp study. *Frontiers in psychology*, 7:171722.

Maria Berger. 2022. Transfer learning parallel metaphor using bilingual embeddings. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 13–23.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, and Rifat Shahriyar. 2023. BanglaNLG and BanglaT5: Benchmarks and resources for evaluating low-resource natural language generation in Bangla. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 726–735, Dubrovnik, Croatia. Association for Computational Linguistics.

Julia Birke and Anoop Sarkar. 2007. Active learning for the identification of nonliteral language. In *Proceedings of the Workshop on Computational Approaches to Figurative Language*, pages 21–28, Rochester, New York. Association for Computational Linguistics.

Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in ty pologically di verse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Tahmid Hasan, Abhik Bhattacharjee, Kazi Samin, Masum Hasan, Madhusudan Basak, M Sohel Rahman, and Rifat Shahriyar. 2020. Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for bengali-english machine translation. *arXiv preprint arXiv:2009.09359*.

Ayesha Khatun, Md Gulzar Hussain, Md Jahidul Islam, Sumaiya Kabir, and Md Mahin. 2020. An emperical framework of idioms translator from bengali to english: Rule based approach. In *2020 IEEE Region 10 Symposium (TENSYMP)*, pages 378–381. IEEE.

Chee Wee Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xianyang Chen. 2020. A report on the 2020 vua and toefl metaphor detection shared task. In *Proceedings of the second workshop on figurative language processing*, pages 18–29.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

Birte Lönneker-Rodman. 2008. The hamburg metaphor database project: issues in resource creation. *Language Resources and Evaluation*, 42:293–318.

Xiaofei Lu and Ben Pin-Yun Wang. 2017. Towards a metaphor-annotated corpus of mandarin chinese. *Language Resources and Evaluation*, 51:663–694.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the fifth joint conference on lexical and computational semantics*, pages 23–33.

Michael Mohler, Mary Brunson, Bryan Rink, and Marc Tomlinson. 2016. Introducing the lcc metaphor datasets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4221–4227.

Jyotirmoy Patowari. 2015. A comparative analysis of emotion conceptual metaphor in english and bangla. *Language in India*, 15(11):264–274.

Elisa Sánchez Bayona. 2023. Detection of everyday metaphor in spanish: annotation and evaluation.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 248–258.

Lennart Wachowiak, Dagmar Gromann, and Chao Xu. 2022. Drum up support: Systematic analysis of image-schematic conceptual metaphors. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 44–53.

Dongyu Zhang, Hongfei Lin, Liang Yang, Shaowu Zhang, and Bo Xu. 2018. Construction of a chinese corpus for the analysis of the emotionality of metaphorical expressions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 144–150.