Lost in Translation? Testing AI Understanding of Figurative and Cultural Language

Group members-

1: Prabal Mehra
CCID:prabal2
Student Id: 1802423

2: Donna Mathew
CCID: donnamat
Student Id: 1770629

3: Mohammad Shahriar Hossain
CCID: mhossai6
Student Id: 1724709

Problem Statement

The ability of artificial intelligence systems to process and interpret natural language has improved significantly in recent years, driven by the development of neural machine translation (NMT) systems and large language models (LLMs). Despite these advances, there remain substantial gaps in how these systems handle figurative and culture-specific language. Idioms, proverbs, metaphors, and polysemous expressions often carry meanings that cannot be inferred directly from their literal components. For example, the English idiom “spill the beans” cannot be understood by analyzing the words “spill” and “beans” in isolation; rather, it conveys the culturally established meaning “to reveal a secret.” These expressions, which are abundant across world languages, continue to present challenges for computational systems that rely heavily on statistical associations and surface patterns.

This project aims to evaluate how effectively LLMs capture the semantic content of figurative and culturally embedded expressions across English and several South Asian languages, including Bengali, Malayalam, and Punjabi. The study will center on two complementary approaches. First, we will analyze model outputs qualitatively, categorizing errors into types such as literal renderings, partial preservation of meaning, and complete semantic failures. Second, we will employ quantitative techniques such as cosine similarity between sentence embeddings to measure how closely AI outputs align with gold-standard translations. This twofold strategy allows us to balance subjective human judgments of meaning with more systematic, reproducible metrics.

To firmly situate the project within computational semantics, we will incorporate resources and methods such as sense-annotated corpora, multilingual wordnets, and word sense disambiguation (WSD). By doing so, we can test whether LLMs achieve true semantic equivalence across languages or whether they falter when confronted with ambiguity and non-literal usage. We will also highlight the role of polysemy, since words with multiple senses provide a natural test bed for WSD and semantic alignment. For instance, the English word “bank” can refer to a financial institution or the side of a river, and translation systems must select the correct sense based on context.

Through this work, we aim to build both a detailed case study and a lightweight demonstration tool. The artifact will serve as an accessible example of how figurative language challenges remain unsolved in modern AI, while also providing a structured framework for comparing multiple models. Ultimately, the project will produce insights that are both practical in helping everyday users of translation tools, recognize where meaning is likely to be lost and theoretical, by advancing our understanding of semantic equivalence in multilingual settings.

Motivation

The motivation for this project arises from both practical needs and gaps in current research. While neural models excel at literal translation, they often break down when tasked with figurative or culturally bound expressions. Baziotis, Mathur, and Hasler show that idioms are often rendered literally by neural translation systems, leading to outputs that fail to capture intended meaning, and they even propose evaluation metrics designed to quantify these literal translation errors (Baziotis et al.). This finding suggests that idioms remain a crucial benchmark for probing whether models capture true semantics or merely align surface forms.

Recent work also underscores this difficulty at scale. The SemEval-2022 Task 2 introduced a multilingual benchmark for idiomaticity detection and sentence embedding, confirming that idiomatic expressions continue to pose challenges for both monolingual and multilingual systems (Tayyar Madabushi et al.). The availability of this benchmark demonstrates the importance of systematic evaluation and provides inspiration for our approach: testing figurative expressions across multiple languages and assessing where and why models succeed or fail.

At the same time, resources are becoming available that can enrich the scope of our analysis. The Multilingual Idioms & Proverbs dataset on Kaggle (AryanRahulTandon) and the ID10M idiom identification framework released by Babelscape demonstrate that the NLP community is investing in idiom-focused datasets. These resources can serve as starting points or supplementary evaluation material for our study, ensuring that our work is not limited to ad hoc examples but grounded in broader datasets.

Beyond idioms, the challenge extends to polysemy and metaphors, which require context-sensitive interpretation. The ELEXIS Final Report highlights that multilingual word sense disambiguation (WSD) and entity linking remain underdeveloped, particularly for low-resource languages such as Bengali, Malayalam, and Punjabi (Maru et al.). Without effective WSD, translation systems often default to the most frequent sense of a word, resulting in mistranslations that distort meaning. By aligning our project with computational semantics and leveraging resources such as wordnets and sense-annotated corpora, we directly address this problem.

The practical motivation is equally compelling. For language learners, incorrect translations of figurative expressions can lead to misunderstandings and frustration. For educators, mistranslations in multilingual teaching materials may obscure cultural knowledge or hinder learning. For casual users, such as travelers or consumers of translated media, literal renderings of idioms can cause confusion or miscommunication. At a broader level, these issues reduce trust in AI translation systems and highlight the need for human oversight when dealing with culturally embedded meaning.

By addressing both theoretical gaps and real-world concerns, this project seeks to demonstrate that figurative and culture-specific language is not just a corner case, but a central challenge for achieving semantic equivalence in multilingual AI.

Audience Definition

The audience for this project spans a wide range of stakeholders, reflecting both practical and research-oriented concerns.

Practical users include language learners, tourists, and everyday users of translation tools who frequently encounter figurative expressions. For a student using an application such as Duolingo, mistranslating an idiom may create confusion and hinder language acquisition. For tourists or professionals engaged in multilingual meetings, a literal translation of a proverb or metaphor can create communication breakdowns or cultural misunderstandings. Similarly, content creators and subtitle teams working on media localization rely on accurate rendering of idiomatic expressions to preserve humor, tone, and cultural resonance.

Educational stakeholders include teachers and curriculum designers who incorporate translation tools into language-learning environments. For this group, accurate handling of figurative language is crucial to ensure that students are not only learning vocabulary and grammar but also gaining cultural competence. Mistranslations risk reinforcing incorrect interpretations and diminishing the richness of the target language.

Research and technical stakeholders include computational linguists, NLP researchers, and AI developers. For them, figurative and culture-specific language represents a benchmark for evaluating semantic models. By focusing on idioms, proverbs, metaphors, and polysemy, and grounding our analysis in sense-annotated corpora, wordnets, and WSD, our project provides a clear link to ongoing research in computational semantics. The results can inform both theoretical models of meaning and practical approaches to improving translation systems.

By combining these perspectives, the project aims to provide insights that are accessible and relevant to multiple communities. It speaks to the everyday need for reliable translation, the educational importance of cultural awareness, and the research imperative of developing systems that go beyond literal word mappings to capture semantic equivalence.

Bibliography

1- AryanRahulTandon. “Multilingual Idioms & Proverbs.” *Kaggle*, 22 Dec. 2024, www.kaggle.com/datasets/aryanrahultandon/multilingual-idioms-indian?utm_source=chatgpt.com.

2- Babelscape. “Babelscape/ID10M: Data and Code for the Paper ‘ID10M: Idiom Identification in 10 Languages’ (NAACL 2022).” *GitHub*, github.com/Babelscape/ID10M. Accessed 24 Sept. 2025.

3-Singh, Harpreet. “Punjabi Grammer.” *Scribd*, Scribd, www.scribd.com/document/709731653/Punjabi-Grammer. Accessed 24 Sept. 2025.

4- Singh, Gurpal. “Muhavre Akhaan.” *Scribd*, Scribd, www.scribd.com/doc/271652841/MuHavre-AkhaAn. Accessed 24 Sept. 2025.

5- Liu, Chaoqun, et al. “Is Translation All You Need? A Study on Solving Multilingual Tasks with Large Language Models.” *Nanyang Technological University, Singapore, National University of Singapore; Shanda AI Research Institute*, NAACL, 2025, pp. 1–21.

6- Baziotis, Christos, Prashant Mathur, and Eva Hasler. “Automatic Evaluation and Analysis of Idioms in Neural Machine Translation.” Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL), 2023, pp. 3682–3700. aclanthology.org/2023.eacl-main.267.pdf. Accessed 24 Sept. 2025.

7- Maru, Michele, et al. D3.5: Multilingual Word Sense Disambiguation and Entity Linking Algorithms—Final Report. ELEXIS, 2022, elex.is/wp-content/uploads/ELEXIS_D3_5_Multilingual_Word_Sense_Disambiguation_and_Entity_Linking-Final_Report.pdf. Accessed 24 Sept. 2025.

8 - “Category:Malayalam Idioms - Wiktionary, The Free Dictionary.” *Wiktionary*, 22 May 2021, en.m.wiktionary.org/wiki/Category:Malayalam_idioms. Accessed 25 Sept. 2025.

9 - “Malayalam Proverbs.” *Scribd*, www.scribd.com/doc/66776656/Malayalam-Proverbs. Accessed 25 Sept. 2025.

10 - Tayyar Madabushi, Harish, et al. “SemEval-2022 Task 2: Multilingual Idiomaticity Detection and Sentence Embedding.” Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Association for Computational Linguistics, 2022, pp. 46–60. ACL Anthology, aclanthology.org/2022.semeval-2.6/. Accessed 25 Sept. 2025.

Team charter

Name	Skills/Expertise	Relevant Courses	Goal for course
Prabal Mehra	Multilingual (English,Punjabi), Python,SQL,Statistics	CMPUT 267, CMPUT 365,INTD 161,CMPUT 256,CMPUT 291	Learn NLP along with understanding how different LLMs work and why they differ.
Donna Mathew	Multilingual(English, Malayalam), Python, NLP, SQL	CMPUT 267, CMPUT 365,INTD 161, CMPUT 291	Use NLP techniques in a hands-on project in a meaningful, practical context
Mohammad Shahriar Hossain	Multilingual (English, Bengali), Python, NLP, Vector Embeddings, Computational Semantics	CMPUT 267, CMPUT 365, CMPUT 497, CMPUT 401, OM 420, OM 468	Apply NLP knowledge, Integrating AI (using LLMs in applications), Learn more about Computational Semantics

Communication Details

Main platform: Whatsapp + shared Google Doc.
Response time: Within 12 hours for async messages.
Progress updates: Weekly check-ins (Google Meet)/ in-person meetings.
Conflict resolution: Raise concerns early –> majority vote –> escalate to prof if needed.
Collaboration: All members contribute to dataset, analysis, and writing.