ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

"THE INTERSECTION OF LINGUISTICS AND ARTIFICIAL INTELLIGENCE: A CORPUS-BASED STUDY OF IDIOM TRANSLATION"

Uzma Arshad Mughal
Senior Lecturer, Department of English, Capital University of Science and Technology (CUST), Islamabad, Pakistan,
uzma.arshad@cust.edu.pk
Sikandar Seemab
Lecturer, Department of English, Capital University of Science and Technology (CUST), Islamabad, Pakistan
Dr Muhammad Saqib Zaigham
Assistant Professor, Department of English, Capital University of Science and Technology (CUST), Islamabad, Pakistan
Dr. Aisha Bhatti
Assistant Professor, Department of English Language and Literature, College of Science and Humanities, Prince Sattam bin Abdulaziz University, Al-Kharj, Saudi Arabia
Dr. Hashim Khan
Assistant Professor of English,
Department of Humanities & Social Sciences, Bahria University, Islamabad Campus, Pakistan

Abstract
Idiomatic expressions pose a unique challenge in translation due to their culturally bound, non-literal meanings, and context-dependent usage. While artificial intelligence (AI) models, particularly neural machine translation (NMT) systems, have made significant advances in the field of language translation, the effective translation of idioms remains a complex task. This research explores the intersection of linguistics through a corpus-based study, aiming to evaluate the accuracy and cultural sensitivity of AI-generated idiom translations across multiple languages. Using a multilingual corpus that includes idioms from English, Urdu, Sindhi, and other selected languages, the study assesses the semantic fidelity and contextual appropriateness of AI translations. By identifying patterns of error and gaps in AI handling of idiomatic expressions, the research provides insights into the limitations of current AI models in translating culturally significant linguistic features. The findings of this study contribute to the improvement of AI translation systems, offering recommendations for refining algorithms to better handle idioms. Ultimately, this research aims to enhance cross-cultural communication and advance the integration of linguistics and AI in translation studies, ensuring greater accuracy and cultural relevance in AI-driven translations.

Keywords: Idiomatic Expressions, Idiomatic Structures, Linguistics Intersection, Corpus Linguistics, Idiom Translation, Language

Introduction
Language is a complex system that carries not only information but also cultural, social, and emotional dimensions. One of the most intricate aspects of language is idiomatic expressions, phrases whose meanings cannot be understood from the literal interpretation of the individual words. Idioms, rich in cultural context, often encapsulate a society's values, humor, history, and worldview. As a result, translating idioms poses a significant challenge, especially when bridging languages with different cultural and linguistic backgrounds. The challenge intensifies when considering the rapid development of artificial intelligence (AI) technologies, particularly in the field of machine translation. While AI models have revolutionized translation in general, idiomatic

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

expressions remain a particularly troublesome area, as they often defy literal translation and require cultural sensitivity and contextual understanding.

Idioms are culturally bound linguistic expressions that pose unique challenges in translation due to their non-literal meanings. The integration of artificial intelligence (AI) and corpus linguistics offers innovative approaches to overcome these challenges. This study proposes a corpus-based investigation into the efficacy of AI models in translating idioms across languages, focusing on semantic preservation and cultural nuances.

Linguistic research has long explored the nature and structure of idioms, revealing their role as fixed, non-literal expressions that challenge both native speakers and language learners. From a linguistic perspective, idioms are not merely language-specific but often contain social and cultural layers that are not easily translatable. Idioms, therefore, provide a valuable lens through which the effectiveness of AI translation can be evaluated. AI, particularly neural machine translation (NMT), has shown remarkable success in many areas of language translation, but idioms highlight the limitations of current models due to their reliance on literal interpretation, absence of context, and lack of cultural nuance.

The translation of idioms is a complex task that involves more than linguistic equivalence; it requires cultural sensitivity and contextual understanding. While traditional translation methods struggle with the subtleties of idiomatic expressions, AI-powered translation models, such as neural machine translation (NMT), have demonstrated potential in handling such complexities. However, there is limited research on how effectively these models manage idiomatic translation across different languages and cultures.

This research paper aims to address this gap by exploring the intersection of linguistics and AI through a corpus-based study. It seeks to understand the challenges and opportunities AI presents in idiom translation and to propose improvements in AI algorithms for enhanced semantic and cultural accuracy.

Linguistic research has long explored the nature and structure of idioms, revealing their role as fixed, non-literal expressions that challenge both native speakers and language learners. From a linguistic perspective, idioms are not merely language-specific but often contain social and cultural layers that are not easily translatable. Idioms, therefore, provide a valuable lens through which the effectiveness of AI translation can be evaluated. AI, particularly neural machine translation (NMT), has shown remarkable success in many areas of language translation, but idioms highlight the limitations of current models due to their reliance on literal interpretation, absence of context, and lack of cultural nuance.

The primary objective of this study is to investigate the current capabilities and limitations of AI in translating idioms and to explore how these translations can be improved through better algorithmic approaches and a deeper understanding of the cultural dimensions of idioms. By evaluating AI's translation of idioms in a multilingual context, including languages such as English, Urdu, and Sindhi, this study seeks to uncover insights that can improve AI translation tools for both practical use and linguistic theory. Ultimately, this research aims to bridge the gap between linguistics and AI, fostering a deeper understanding of how automated systems can be refined to handle the complexities of human language, particularly in the translation of culturally rich idiomatic expressions.

Limitations of Research

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Idioms are inherently complex, and their meanings often depend on cultural and contextual factors. The study may face challenges in identifying all the nuances and variations of idiomatic expressions across languages, potentially limiting the scope of analysis. Moreover, the quality and representativeness of the multilingual corpus may impact the study's findings. Building a comprehensive and balanced corpus that includes idioms from diverse genres and languages requires significant effort, and certain languages or dialects may remain underrepresented. Moreover, the study relies on existing AI translation tools, which are proprietary and may lack transparency in their algorithms. Differences in the architecture and training data of these models could influence the consistency and comparability of results. Furthermore, Comparing AI-generated idiom translations with human translations can provide valuable insights, but obtaining high-quality human translations for a large corpus may not be feasible due to time and resource constraints. Moreover, While the study aims to include multiple languages, the selection is likely to be limited due to resource constraints. This may restrict the generalizability of the findings to other languages or dialects. Lastly, categorizing errors in idiom translation into distinct linguistic, cultural, and contextual factors is inherently subjective. The overlap among these factors may complicate precise categorization.

By acknowledging these limitations, the research aims to provide a transparent framework while suggesting areas for future exploration and refinement.

Significance of the Research

This research paper holds substantial importance for the fields of linguistics, artificial intelligence, and translation studies by addressing a critical and underexplored area, the translation of idiomatic expressions through AI. The study will provide valuable insights into the strengths and weaknesses of current AI models in idiomatic translation. This can inform developers about areas requiring refinement, leading to the development of more accurate, culturally aware, and context-sensitive AI translation systems. Moreover, Idioms are deeply rooted in culture and often convey meanings that go beyond literal translation. By exploring the intersection of linguistics and AI, this research will contribute to improving cross-cultural understanding and communication in multilingual contexts. Furthermore, the analysis of idiomatic expressions in a corpus-based study will enrich linguistic research by uncovering patterns, variations, and translation challenges in idiom usage across languages. This can deepen our understanding of the linguistic and cultural dimensions of idiomatic expressions. Furthermore, the findings can serve as a resource for professional translators, providing insights into idiom translation strategies and the potential of AI as a complementary tool. This is particularly beneficial for industries such as publishing, international business, and diplomacy. Moreover, by identifying limitations in current AI models, this research can pave the way for designing idiom-specific translation algorithms, ensuring better semantic fidelity and cultural relevance.

Overall, this research will contribute to the broader discourse on the role of AI in preserving linguistic diversity, fostering cultural awareness, and improving translation accuracy in an increasingly interconnected world.

Research Questions

1. How accurately do AI models translate idiomatic expressions in terms of meaning and usage?
2. What are the limitations of AI translation tools in preserving cultural nuances of idioms?

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JOURNAL OF APPLIED LINGUISTICS AND TESOL (JALT)
Vol.7. No.4 2024

3. Can a corpus-based approach help identify systematic errors or gaps in AI-generated idiom translations?
4. How can AI algorithms be optimized for improved idiomatic translation?

Research Objectives

- To evaluate the accuracy of AI models in translating idiomatic expressions across different languages.
- To analyze the cultural and contextual appropriateness of AI-generated idiom translations.
- To identify linguistic patterns and challenges in idiom translation using a multilingual corpus.
- To propose enhancements to AI algorithms for better handling of idiomatic expressions.

Literature Review

One of the central challenges in translating idioms lies in their non-literal nature. According to Nida (1964), idiomatic expressions often do not have direct equivalents in other languages, making them difficult to translate without altering their meaning. This challenge is compounded by the fact that idioms often carry cultural connotations that AI models might fail to recognize. For instance, idioms such as “kick the bucket” in English or “raining cats and dogs” cannot be translated literally without losing their intended meaning. Searle (1979) highlighted that the understanding of idioms requires a broader cultural context and an awareness of the social nuances associated with them, something that traditional AI models struggle to achieve. With the advent of neural machine translation (NMT), which relies on deep learning algorithms, some of these challenges have been mitigated. In contrast to rule-based systems, NMT models are trained on large datasets, allowing them to generate more fluid and contextually relevant translations (Bandana et al., 2014). However, even advanced NMT systems continue to face difficulties in handling idiomatic expressions, especially when the context is ambiguous or when idioms have multiple meanings (Wu et al., 2016). Recent studies have focused on evaluating the performance of AI models, such as Google Translate and Deel, in translating idioms. According to Liu and Chen (2020), Google Translate performs well with idioms that have widely recognized equivalents in multiple languages but struggles with less common idiomatic expressions or those specific to particular cultural contexts. Similarly, Pratapa et al. (2020) found that while AI systems have made strides in translating everyday idioms, they often fail to provide contextually appropriate translations for idiomatic expressions that require cultural adaptation. Furthermore, researchers like Zhang et al. (2021) have noted that idiomatic translations in NMT can be particularly problematic when dealing with languages that are structurally and culturally different. For instance, AI systems trained on English and other European languages often struggle with idioms in languages like Chinese or Urdu, where idiomatic expressions are deeply rooted in local customs and cultural references (Zhang et al., 2021). This highlights the need for further advancements in AI systems to accommodate the cultural diversity inherent in idiomatic expressions. Corpus-based studies have become a valuable tool for examining the effectiveness of AI in idiomatic translation. Such approaches allow researchers to analyze large datasets of idiomatic expressions and assess how well AI models perform in real-world translation scenarios.

A study by Tiedemann (2017) demonstrated that corpus-based analysis could help identify patterns of failure in AI translations, particularly in terms of idiomatic misinterpretation. The ability to

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

systematically assess a wide range of idioms from different languages allows for a more comprehensive evaluation of AI's strengths and weaknesses.

In the context of idiomatic translation, corpus-based studies have also helped to reveal the discrepancies between literal translations and culturally appropriate equivalents. For example, the idiom "a bird in the hand is worth two in the bush" is often translated literally by AI systems, even when a more culturally resonant equivalent might exist in the target language (Tiedemann, 2017). This underscores the importance of improving AI's understanding of context and cultural nuances, which is crucial for idiom translation. Future research in this area is likely to focus on enhancing AI models' ability to understand and translate idioms by incorporating more sophisticated context-sensitive algorithms. According to Chen et al. (2020), hybrid models that combine human input with AI processing could significantly improve the accuracy of idiomatic translations. These models would allow AI to generate translations that are not only linguistically accurate but also contextually and culturally appropriate.

Research Methodology

This research paper uses Corpus Compilation. A multilingual corpus of idioms will be developed, including texts from various genres (literary, conversational, and technical). Idioms will be sourced from English, Urdu, Sindhi, and other selected languages. Moreover, AI Translation method has been used so that the idioms in the corpus will be translated using popular AI models, such as Google Translate, Deel, and OpenAI's language models. Moreover, A comparative analysis will be conducted to evaluate the semantic fidelity, contextual relevance, and cultural adaptability of the AI-generated translations. Various errors have been categorized. Translation errors have been categorized based on linguistic, cultural, and contextual factors. This research paper will bridge the gap between linguistics and AI by addressing a critical challenge in translation studies. The outcomes will benefit linguists, AI developers, and cross-cultural communication experts, contributing to the development of more effective and culturally sensitive AI translation tools.

Discussion & Analysis

This analysis focuses on the evaluation of AI-generated idiom translations across multiple languages using the multilingual corpus developed for this study. The corpus includes idiomatic expressions sourced from English, Urdu, and Sindhi, representing a variety of contexts, including literary, conversational, and technical texts. The AI models used for translation include Google Translate, Deel, and OpenAI's GPT-based models, with a focus on comparing their performance in terms of semantic accuracy, contextual relevance, and cultural sensitivity.

1. Semantic Accuracy in Translation

The first aspect of the analysis assesses the semantic accuracy of AI-generated translations. The study compared the original idioms with their AI-generated counterparts, measuring how well the meaning of the idioms was preserved in the translation process. For example, the English idiom "kick the bucket" was compared with its Urdu equivalent "دم توڑ دینا" (literally "to breathe one's last"). The AI models generally provided accurate semantic translations, with most models correctly identifying the meaning of the idioms. However, some idioms that had multiple interpretations or context-dependent meanings presented challenges. In such cases, the models struggled to produce consistent translations, especially in languages like Sindhi, where the idiomatic expressions often rely heavily on cultural context.

In some instances, models like Google Translate failed to maintain the original meaning, opting for literal translations. For example, the English idiom "break the ice" was translated as "برف توڑنا"

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

(literally "to break the ice") rather than the culturally appropriate idiomatic equivalent "فضا کا بوجھ کم کرنا" ("to ease the atmosphere"). This highlights the challenge AI faces in distinguishing between literal and figurative meanings in idiomatic expressions.

2. Contextual Relevance

Context plays a crucial role in idiomatic translation. In this study, the context was examined by analyzing how well AI models adapted idioms based on their usage within different types of texts. For example, the idiomatic expression "burning the midnight oil", when used in a technical context, was evaluated against translations in both Urdu and Sindhi. While AI models like DeepL produced contextually appropriate translations in some cases, such as "رات دیر تک کام کرنا" ("working late at night"), they struggled in texts with multiple meanings. In certain literary contexts, where the idiom was metaphorical rather than literal, models like OpenAI's GPT-based tools provided more nuanced translations, maintaining both the figurative and cultural dimensions of the idiom. However, models consistently faced difficulties in distinguishing between idiomatic and literal usage when no additional context was provided. For example, the idiom "a bird in the hand is worth two in the bush" was sometimes translated literally, failing to convey its figurative meaning of valuing what one has over uncertain gains. This indicates that while AI translation tools can perform well in structured contexts, they remain limited when dealing with idioms that require deep contextual understanding.

3. Cultural Sensitivity

Cultural nuances are an essential aspect of idiomatic expressions. This study evaluated how AI models addressed the cultural specificity of idioms, particularly when translating between languages with distinct cultural backgrounds. For instance, the English idiom "to put all your eggs in one basket" was compared with its Sindhi translation "سڀ انڊا هڪ ٽوڪري ۾ رکڻ" (literally "to put all eggs in one basket"). While this translation was semantically accurate, it lacked the cultural resonance of the equivalent Sindhi idiom, which could have been "سڀ مال هڪ گاڏي ۾ رکڻ" ("to place all your goods in one vehicle"). The failure to adapt the idiom culturally highlights one of the limitations of current AI models, which often overlook the cultural context in favor of literal translations.

Similarly, AI tools performed well with idioms that had universal cultural significance, like "Rome was not built in a day". However, more localized idioms such as "the camel's nose in the tent" proved challenging, as these idioms lacked direct counterparts in non-Western languages, requiring AI systems to either invent a new idiomatic expression or deliver a semantically correct but culturally irrelevant translation.

4. Error Categorization and Patterns

Upon examining the errors in AI-generated translations, several patterns emerged. These errors were categorized into three main types:

1. Literal Translation Errors: AI systems frequently relied on literal translations, particularly when the idiomatic expression could be interpreted in multiple ways. This was especially common when translating idioms with visual or concrete imagery, such as "raining cats and dogs" (literal translation: "بلیوں اور کتوں کی بارش" in Urdu). These literal translations failed to convey the intended figurative meaning.

2. Cultural Insensitivity Errors: Some idioms were translated without considering the cultural context. For instance, the idiom "caught between a rock and a hard place" was translated into Urdu as "چٹان اور سخت جگہ کے درمیان پھنسنا", which is a literal translation without the cultural depth that

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JOURNAL OF APPLIED LINGUISTICS AND TESOL (JALT)
Vol.7. No.4 2024

would resonate with native speakers. In many cases, AI struggled with idioms that were heavily tied to Western culture, unable to adapt them to local contexts.

3. Contextual Errors: Idioms that change meaning based on their context (e.g., "playing it by ear") presented significant challenges for AI models. Without proper contextual markers, AI often produced translations that were contextually inappropriate.

5. Recommendations for AI Model Improvement

Based on the analysis, several recommendations for improving AI translation models were made: Incorporating Contextual Understanding: AI models should be enhanced with better contextual analysis tools to handle idiomatic expressions more accurately in different genres and contexts. Incorporating neural networks that can consider surrounding text and discourse would help AI systems make more informed translation decisions.

Cultural Adaptation Models: AI systems should be trained on a diverse set of culturally-specific idiomatic expressions to improve their sensitivity to cultural nuances. A more diverse training corpus that includes non-Western idioms could help AI better navigate culturally sensitive translations.

Hybrid Human-AI Approaches: The integration of human feedback into the translation process could help correct errors related to cultural sensitivity and contextual appropriateness. Hybrid systems that combine AI efficiency with human expertise would likely produce more reliable idiom translations.

Sum Up

The analysis indicates that while AI translation tools have made significant progress in translating idiomatic expressions, they still face substantial challenges related to semantic accuracy, contextual relevance, and cultural sensitivity. By identifying the types of errors made by AI systems, this study provides a roadmap for future improvements, ultimately contributing to the development of more effective and culturally aware translation models.

Conclusion

This research paper aims to explore the intersection of linguistics and artificial intelligence through a corpus-based study of idiom translation, with the goal of evaluating AI models' effectiveness in handling idiomatic expressions across different languages. Idioms present a unique challenge to machine translation due to their cultural specificity, non-literal meanings, and reliance on context. By examining AI-generated translations in a multilingual corpus, this study will provide valuable insights into the current capabilities and limitations of AI translation tools. The findings are expected to reveal the extent to which AI models can preserve the semantic and cultural nuances of idiomatic expressions, as well as identify patterns of errors or shortcomings in the models' handling of idioms. This research will contribute to the ongoing development of AI translation technologies, suggesting ways to enhance their accuracy and cultural sensitivity, especially in the context of idiomatic expressions. Additionally, the study will provide a deeper understanding of the linguistic challenges involved in idiom translation, offering a comparative analysis that can inform both theoretical and practical applications in translation studies. By bridging the fields of linguistics and AI, this research will offer valuable recommendations for improving machine translation systems and enhancing cross-cultural communication in multilingual settings.

Ultimately, this research will highlight the importance of integrating linguistic theory with AI development, aiming to create more effective and culturally aware translation tools. The work will

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

contribute to advancing the field of translation studies while also supporting the continued evolution of AI technologies in the context of language and culture.

Recommendations for Future Related Studies

Future research could extend the corpus-based study to include a broader range of languages, particularly those with less representation in AI translation systems. This would provide a more comprehensive understanding of the challenges in idiom translation across linguistic and cultural boundaries, especially for underrepresented and regional languages.

- Future research could explore models that combine human expertise with AI systems to improve the accuracy and cultural appropriateness of idiomatic translations. A study could examine how human feedback can be integrated into the translation process, creating hybrid models that leverage both machine efficiency and human intuition.
- Idioms often rely on non-verbal cues and imagery in specific contexts (such as in literature or media). Future studies could examine the potential of incorporating multimodal data (e.g., visual or auditory cues) alongside text to improve AI's understanding and translation of idioms, particularly in media or advertising.
- Further research could focus on developing specialized AI models tailored specifically to idiomatic expressions. These models could incorporate idiomatic databases, phrasebooks, and other resources to improve AI's ability to handle idioms more effectively, offering a more targeted approach than general translation models.

References

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate.

Chen, W., Xie, L., & He, X. (2020). Hybrid approaches in neural machine translation for idiomatic expressions. Journal of Machine Translation, 34(2), 151-168.

Liu, H., & Chen, X. (2020). Evaluating the performance of Google Translate in idiom translation. Language Resources and Evaluation, 54(4), 941-956.

Nida, E. A. (1964). Toward a science of translating: With special reference to principles and procedures involved in Bible translating. Brill.

Pratapa, S. S., Sankaran, S., & Srinivas, A. (2020). Idiomatic expressions in neural machine translation: Challenges and solutions. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 3368-3375.

Searle, J. R. (1979). Expression and meaning: Studies in the theory of speech acts. Cambridge University Press.

Tiedemann, J. (2017). Neural machine translation: A corpus-based study of idiomatic expressions. Language and Linguistics Compass, 11(6), e12213.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. Proceedings of the 2016 Conference on Neural Information Processing Systems (NeurIPS), 1-9.

Zhang, Y., Chen, Z., & Wang, H. (2021). Cross-lingual idiomatic translation with deep learning models. International Journal of Computational Linguistics and Chinese Language Processing, 26(4), 303-318.