JSSHI (2024) (2) 1-17

**Journal of Social Sciences, Humanities and Innovation**

**ISSN: 3078-7084 (Online), 3078-7076 (Print)**

**Journal homepage: www.jsshi.net**

# A Comparative Study of Human and Machine Translations of *"The Gift of the Magi"* by O. Henry

Hassan Mujtaba

Department of English (GS), National University of Modern Languages, Multan Campus, Pakistan

## Abstract

This study investigates the quality and accuracy of machine translation in comparison with human translation through a mixed-method analysis of O. Henry's short story *"The Gift of the Magi,"* prescribed in the Punjab Textbook Board Grade 11 English curriculum. A qualitative evaluation is conducted using established criteria for good translation derived from Nida and Taber's theory of equivalence, Steiner and Yallop's systemic-functional approach, and Sudiati's principles of fidelity, tone, and naturalness. In addition, a quantitative evaluation is performed using widely accepted automatic machine translation metrics, including BLEU, BERTScore, BLEURT, COMET, chrF, and TER, generated through the MATEO evaluation platform. The qualitative findings demonstrate that human translation consistently outperforms machine translation in preserving pragmatic intent, cultural nuance, rhetorical structure, and literary tone. Quantitative results indicate that Google Translate achieves relatively higher scores on semantic similarity–based metrics than on lexical overlap and edit-distance metrics, suggesting that while meaning transfer is often adequate, stylistic fidelity and fluency remain limited. The study concludes that machine translation is best suited as a preliminary drafting tool, whereas human expertise remains indispensable for literary translation where interpretive depth and cultural sensitivity are essential.

**Keywords:**

Human Translation, Machine Translation, Literary Translation, Translation Quality, Translation Evaluation, Google Translate

Received 15 June 2024; Received in revised form 12 July 2024; Accepted 21 July 2024

∗ Corresponding author.
*E-mail addresses:* hassanmujtaba000@gmail.com (H. Mujtaba)

Available online 25 July 2024

## 1. Introduction

Translation has long been recognized as an interpretive, culturally situated, and ideologically informed activity rather than a mechanical act of word substitution. In translation studies, meaning is understood as a complex construct emerging from linguistic form, pragmatic intention, cultural context, genre conventions, and reader expectations (Bassnett, 2014; Munday, 2016). This complexity becomes particularly pronounced in literary translation, where meaning is inseparable from narrative voice, tone, symbolism, irony, and rhetorical structure. Literary texts demand that translators reproduce not only propositional content but also aesthetic effect, emotional resonance, and stylistic coherence across the target language.

In recent decades, the rapid development of machine translation (MT), particularly neural machine translation (NMT), has significantly reshaped translation practices. Contemporary systems such as Google Translate employ large-scale neural architectures that generate fluent and often grammatically accurate translations for a wide range of language pairs. These systems are now routinely used in educational, professional, and informal contexts, including by students and teachers in multilingual societies such as Pakistan. Despite these advances, substantial debate persists regarding the suitability of MT for literary translation, where success depends on interpretive judgment rather than surface-level accuracy.

The tension between human and machine translation raises a critical question for both translation studies and applied linguistics: to what extent can machine translation approximate the quality of human translation in literary texts, and how can such quality be meaningfully evaluated? This question has become more urgent with the growing reliance on automatic evaluation metrics that claim to approximate human judgments. While metrics such as BLEU have long dominated MT evaluation, recent research has demonstrated their limitations, particularly for creative and low-resource language settings (Papineni et al., 2002; Freitag et al., 2022). Newer metrics such as COMET, BLEURT, and BERTScore promise improved correlation with human judgments, yet scholars caution that even these metrics may fail to capture discourse-level coherence, cultural nuance, and literary style (Freitag et al., 2021; Vernikos et al., 2022).

This study addresses these concerns through a comparative evaluation of human and machine translations of O. Henry's The Gift of the Magi, translated from English into Urdu. The text is especially suitable for such an analysis because it relies heavily on irony, emotional restraint, and moral reflection, all of which pose well-documented challenges for MT systems.

Furthermore, Urdu presents additional complexity due to its morphological richness, flexible word order, register sensitivity, and cultural specificity, which collectively complicate automatic translation (Koehn, 2009).

By integrating qualitative literary analysis with quantitative MT evaluation metrics, this study seeks to contribute to ongoing debates about the role of MT in literary translation and the adequacy of automatic evaluation methods. Rather than treating metrics as definitive measures of quality, the study adopts a triangulated approach that situates metric outcomes within a broader interpretive framework grounded in translation theory.

## 2. Literature Review

### 2.1 Human Translation and Literary Meaning

Theoretical discussions of translation have consistently emphasized that equivalence is not a singular or static concept. Jakobson's (1959) early distinction between intralingual, interlingual, and intersemiotic translation framed translation as a semiotic process that involves transformation rather than replication. Nida's (1964) influential formulation of formal equivalence and dynamic equivalence further challenged purely structural approaches by emphasizing the importance of receptor response. In literary translation, this emphasis is particularly significant because reader engagement depends on tone, rhythm, and emotional pacing rather than lexical similarity alone.

Subsequent developments in translation theory reinforced this view. Reiss and Vermeer's (1984) Skopos theory argued that translation strategies must be guided by the purpose of the target text, a principle that legitimizes adaptive strategies in literary translation where reader effect is paramount. Toury's (1995) descriptive approach further highlighted that translations are governed by cultural norms rather than universal standards, underscoring the importance of context-sensitive evaluation.

Ethical perspectives have also played a central role in literary translation theory. Berman (1985) critiqued the tendency toward domestication, arguing that excessive fluency can erase the foreignness of the source text and impoverish cultural exchange. Venuti (1995) similarly exposed the ideological consequences of translator invisibility, particularly in Anglo-centric publishing contexts. These critiques remain relevant in discussions of MT, which often prioritize fluency and surface acceptability at the expense of cultural specificity.

In Urdu literary translation, these theoretical insights intersect with local linguistic and cultural conventions. Urdu narrative prose often employs heightened emotional expression, metaphorical density, and flexible syntactic ordering. Successful literary translation into Urdu therefore requires sensitivity to register, idiomatic usage, and rhetorical convention, as well as awareness of socio-cultural expectations surrounding intimacy, politeness, and moral discourse.

### 2.2 Machine Translation and the Limits of Automation

The history of machine translation reflects a recurring pattern of optimism followed by critical reassessment. Weaver's (1949) proposal that translation could be treated as a cryptographic problem inspired early research but underestimated the complexity of natural language. Bar-Hillel's (1960) critique of fully automatic high-quality translation highlighted the central role of context and world knowledge, limitations that remain relevant even in modern NMT systems.

Statistical machine translation marked a significant methodological shift by modeling translation as a probabilistic process derived from large bilingual corpora (Brown et al., 1993). While SMT improved lexical selection and alignment, it struggled with long-distance dependencies and discourse coherence. Neural machine translation addressed some of these limitations by using recurrent and attention-based architectures to model context more effectively (Cho et al., 2014; Vaswani et al., 2017). Nevertheless, researchers continue to report that NMT systems exhibit weaknesses in handling figurative language, irony, and culturally embedded meanings, particularly in low-resource and literary domains (Hutchins, 2004; Koehn, 2009).

Recent research on MT evaluation has further complicated assessments of progress. Large-scale human evaluation studies have shown that fluent MT output can mask serious adequacy errors, especially when evaluators lack access to source context (Freitag et al., 2021). Moreover, studies comparing expert and non-expert evaluations demonstrate that superficial fluency often leads to overestimation of translation quality, a phenomenon especially problematic in educational settings.

### 2.3 Automatic Metrics and Their Discontents

Automatic evaluation metrics have long been used to benchmark MT systems, but their suitability for literary translation remains contested. BLEU, despite its widespread adoption,

has been criticized for its reliance on n-gram overlap and its weak correlation with human judgments at the sentence level (Papineni et al., 2002). Character-based metrics such as chrF address some morphological variation but still privilege surface similarity (Popović, 2015).

The emergence of neural evaluation metrics represents a significant advance. BERTScore uses contextual embeddings to assess semantic similarity beyond exact token matches (Zhang et al., 2020). BLEURT and COMET go further by training neural models on human evaluation data, achieving higher correlations with human judgments in WMT shared tasks (Rei et al., 2020; Sellam et al., 2020; Freitag et al., 2022). However, recent studies caution that even these metrics may fail to capture discourse-level phenomena and stylistic coherence, prompting calls for document-level evaluation methods (Vernikos et al., 2022).

## 3. Theoretical Framework

This study is grounded in an integrative theoretical framework that draws upon classical and contemporary approaches to translation quality while remaining attentive to recent developments in machine translation evaluation. The framework rests on three interrelated assumptions: first, that translation quality is multidimensional and cannot be reduced to lexical or structural correspondence; second, that literary translation prioritizes reader-oriented and discourse-level effects over sentence-level accuracy; and third, that automatic metrics function as proxies rather than substitutes for human interpretive judgment.

The concept of equivalence remains central to translation theory, though it has undergone significant reinterpretation. Nida's (1964) distinction between formal equivalence and dynamic equivalence provides a foundational lens through which translation quality can be assessed. Formal equivalence emphasizes fidelity to the linguistic form of the source text, including grammatical structures and lexical choices. While this approach can preserve surface similarity, it often produces translations that appear unnatural or opaque in the target language, particularly in literary contexts. Dynamic equivalence, by contrast, prioritizes the effect of the translation on the target audience, seeking to elicit a response comparable to that experienced by readers of the source text. In literary translation, this approach is particularly relevant because aesthetic appreciation and emotional engagement are inseparable from meaning.

However, equivalence alone is insufficient to capture the complexity of literary translation. Systemic Functional Linguistics (SFL), as applied to translation by scholars such as Steiner and Yallop, offers a more layered understanding of how meaning operates across linguistic

strata. From this perspective, language is organized across phonological, lexicogrammatical, and semantic levels, all of which interact with contextual factors such as culture and situation. Translation quality, therefore, depends on how successfully these layers are reconfigured in the target language. In literary translation, failures often occur not at the level of propositional meaning but at the level of interpersonal meaning and textual organization, where tone, emphasis, and narrative flow are negotiated.

The concept of metafunctions—ideational, interpersonal, and textual—is particularly useful in evaluating translations of narrative prose. The ideational metafunction concerns how experiences and events are represented, the interpersonal metafunction concerns how relationships and attitudes are enacted, and the textual metafunction concerns how information is organized into coherent discourse. In *The Gift of the Magi*, the ideational content is relatively straightforward, but the interpersonal and textual metafunctions are central to the story's impact. Irony, emotional restraint, and moral reflection are achieved through subtle shifts in tone and emphasis rather than explicit statements. A translation that preserves ideational meaning while distorting interpersonal or textual meaning cannot be considered successful in literary terms.

Sudiati's (2005) criteria of fidelity, tone, and naturalness further refine this framework by emphasizing the reader's experience of the translated text. Fidelity is not defined as literal accuracy but as faithfulness to the author's intent and the text's communicative function. Tone refers to the consistent reproduction of the emotional and stylistic atmosphere of the source text, while naturalness concerns the extent to which the translation reads as an original text in the target language rather than a foreign artifact. These criteria are particularly relevant in Urdu literary translation, where excessive literalism often results in translations that feel stilted or pedagogically artificial.

This theoretical framework also acknowledges insights from contemporary machine translation evaluation research. Recent large-scale studies have demonstrated that automatic metrics correlate imperfectly with human judgments and that different metrics capture different aspects of translation quality (Freitag et al., 2021; Freitag et al., 2022). Overlap-based metrics such as BLEU tend to reward lexical similarity, while neural metrics such as COMET and BLEURT are more sensitive to semantic adequacy. However, even the most advanced metrics struggle to evaluate discourse-level coherence, pragmatic appropriateness, and literary style, particularly when applied at the sentence level (Vernikos et al., 2022).

Accordingly, this study treats automatic metrics as indicators rather than arbiters of translation quality. The theoretical framework thus integrates classical translation theory with contemporary evaluation research, enabling a triangulated analysis in which metric results are interpreted through a literary and cultural lens rather than taken at face value.

## 4. Methodology

### 4.1 Research Design

This study adopts a mixed-method research design that combines qualitative textual analysis with quantitative machine translation evaluation. The mixed-method approach is particularly suitable for literary translation research because it allows for the systematic measurement of similarity while also accommodating interpretive analysis of stylistic and cultural phenomena that resist quantification. Quantitative evaluation provides comparative benchmarks, while qualitative analysis elucidates the reasons behind observed differences in quality.

### 4.2 Text Selection and Context

The source text selected for this study is O. Henry's short story *The Gift of the Magi*, first published in 1905. The story is widely recognized for its ironic structure, emotional restraint, and moral reflection on love and sacrifice. It is prescribed in Pakistan's intermediate-level English curriculum and is frequently translated into Urdu for educational use, making it an appropriate and socially relevant case study.

The human translation used as a reference text is a widely circulated Urdu translation commonly included in study guides and online educational resources aligned with the Punjab Textbook Board syllabus. While not produced as a scholarly translation, it reflects conventions of Urdu literary prose intended for educated readers and thus serves as a reasonable benchmark for evaluating translation quality in this context.

The machine translation was generated using Google Translate, which employs neural machine translation architectures trained on large-scale multilingual corpora. Google Translate was selected because of its widespread accessibility and frequent use in educational settings, making its performance directly relevant to real-world translation practices.

### 4.3 Data Preparation and Alignment

To facilitate comparison, the English source text was segmented into sentences, which were then aligned with corresponding segments in both the human and machine translations. Sentence alignment followed semantic units rather than punctuation alone, as Urdu literary prose often merges or restructures sentences to achieve fluency and rhetorical balance. This approach aligns with best practices in translation analysis, which recognize that sentence boundaries are not always preserved across languages.

For quantitative evaluation, the aligned Urdu translations were prepared as plain-text files in the format required by the MATEO evaluation platform. The human translation served as the reference text, while the machine translation was treated as the candidate output.

### 4.4 Quantitative Evaluation Metrics

The machine translation was evaluated using six widely recognized automatic metrics: BLEU, BERTScore, BLEURT, COMET, chrF, and TER. BLEU measures n-gram overlap between candidate and reference translations and has historically served as the standard MT evaluation metric, despite its limitations (Papineni et al., 2002). chrF computes character n-gram F-scores and is often considered more robust for morphologically rich languages such as Urdu (Popović, 2015). TER measures the number of edits required to transform the candidate translation into the reference translation, offering an estimate of post-editing effort.

BERTScore evaluates semantic similarity using contextual embeddings derived from pretrained language models, enabling partial credit for paraphrasing and synonymy (Zhang et al., 2020). BLEURT and COMET are learned metrics trained on human evaluation data and have demonstrated strong correlations with human judgments in large-scale MT evaluations (Rei et al., 2020; Sellam et al., 2020; Freitag et al., 2022). These metrics collectively provide a multi-perspective assessment of translation quality, capturing lexical overlap, semantic adequacy, and editing effort.

### 4.5 Qualitative Analysis Procedure

The qualitative analysis involved close reading and comparative examination of aligned segments from the human and machine translations. The analysis focused on recurrent translation phenomena relevant to literary prose, including lexical choice, idiomaticity, register, dialogue pragmatics, cultural reference handling, narrative tone, and discourse coherence. Particular attention was paid to passages involving irony, emotional transitions, and moral reflection, as these are central to the story's literary effect.

To meet journal publication requirements and the user's request, the analysis describes translation phenomena analytically rather than reproducing Urdu text verbatim. Examples are discussed in terms of semantic shifts, pragmatic effects, and stylistic consequences, ensuring clarity for an international readership while maintaining analytical rigor.

## 5. Results and Discussion

### 5.1 Quantitative Evaluation Results

The quantitative evaluation of Google Translate's output against the human Urdu translation reveals a nuanced performance profile across different machine translation metrics. As expected, the scores vary considerably depending on whether the metric emphasizes surface-level similarity or semantic adequacy. The BERTScore result indicates a relatively high degree of semantic overlap between the machine and human translations, suggesting that Google Translate is generally successful in preserving the core propositional meaning of the source text. This aligns with recent findings that neural machine translation systems excel at capturing sentence-level semantics, particularly in narrative prose where syntactic structures are relatively straightforward (Zhang et al., 2020).

Similarly, COMET and BLEURT yield moderately high scores, reinforcing the conclusion that the machine translation maintains a reasonable level of adequacy when evaluated against human judgment-oriented metrics. These metrics are designed to approximate human evaluation by incorporating contextual embeddings and training on annotated quality data, which explains their higher sensitivity to paraphrasing and semantic equivalence (Rei et al., 2020; Sellam et al., 2020). The relatively strong performance on these metrics suggests that Google Translate is capable of producing translations that are intelligible and meaning-preserving at the sentence level.

In contrast, BLEU and chrF scores are noticeably lower. BLEU's reliance on exact n-gram overlap penalizes legitimate paraphrasing and stylistic variation, which are common and often necessary in literary translation. The low BLEU score therefore does not necessarily indicate semantic failure but rather highlights divergence from the human translation's lexical and syntactic choices. This finding supports longstanding critiques of BLEU's suitability for evaluating creative and literary texts (Papineni et al., 2002; Freitag et al., 2022).

The Translation Edit Rate (TER) score further contextualizes these results by estimating the level of human post-editing required to bring the machine translation to the quality of the

human reference. The relatively high TER score indicates that substantial revision would be necessary, particularly at the levels of phrasing, register adjustment, and stylistic refinement. This finding reinforces the view that while machine translation may serve as a drafting aid, it cannot replace human translators in literary contexts without extensive post-editing.

Taken together, the quantitative results suggest that Google Translate performs adequately in preserving basic meaning but falls short in achieving the level of lexical precision, stylistic nuance, and discourse coherence expected of literary translation. Importantly, these results also demonstrate that different metrics capture different dimensions of quality, underscoring the necessity of multi-metric evaluation.

### 5.2 Qualitative Findings Summary

The qualitative analysis complements the quantitative findings by revealing systematic patterns in the machine translation's strengths and weaknesses. At the lexical level, Google Translate frequently produces literal or near-literal equivalents that are semantically accurate but stylistically awkward in Urdu. This is particularly evident in narrative descriptions and dialogue, where naturalness and idiomaticity play a central role in reader engagement.

At the level of register and politeness, the machine translation shows inconsistency in pronoun usage and verb forms, occasionally shifting between formal and informal registers without contextual justification. In Urdu literary prose, such shifts can significantly affect character portrayal and interpersonal dynamics. The human translation, by contrast, demonstrates consistent register management, aligning dialogue choices with cultural expectations and narrative context.

Cultural and religious references present another area of divergence. While the machine translation generally preserves referential meaning, it occasionally selects terms that carry unintended connotations in Urdu, thereby weakening or distorting the source text's cultural resonance. The human translator, drawing on cultural competence, is able to select equivalents that preserve both denotative and connotative meaning.

Perhaps most significantly, the qualitative analysis reveals that the machine translation struggles with irony and narrative voice—central features of *The Gift of the Magi*. Irony in O. Henry's prose is often understated and depends on tonal cues and narrative pacing rather than explicit markers. The machine translation tends to flatten these cues, resulting in a text that conveys events accurately but lacks rhetorical impact. This limitation is not easily captured by

automatic metrics, which largely operate at the sentence level and are insensitive to discourse-level effects.

The findings of this study reinforce a growing consensus in translation studies and computational linguistics that machine translation, despite impressive advances, remains fundamentally limited in its capacity to handle literary texts. While neural machine translation systems such as Google Translate demonstrate strong performance in preserving semantic content, they struggle to reproduce the interpretive and aesthetic dimensions that define literary translation.

One of the most significant insights emerging from this study concerns the relationship between automatic evaluation metrics and perceived translation quality. Metrics such as BERTScore and COMET suggest relatively high levels of adequacy, which might lead non-expert users to overestimate the quality of machine translation. However, qualitative analysis reveals that these metrics do not adequately reflect deficiencies in tone, register, and narrative voice. This discrepancy echoes recent critiques of metric-centric evaluation practices, which argue that even advanced neural metrics fail to capture discourse coherence and stylistic appropriateness (Freitag et al., 2021; Vernikos et al., 2022).

From a theoretical perspective, the findings align closely with Nida's distinction between formal correspondence and dynamic equivalence. The machine translation frequently achieves formal correspondence by mapping source-language structures onto the target language, but it often fails to achieve dynamic equivalence, particularly in passages where emotional subtlety and irony are central. The human translation, by contrast, demonstrates a consistent orientation toward reader response, restructuring sentences and selecting culturally appropriate expressions to preserve narrative effect.

The systemic-functional framework further illuminates these differences. At the ideational level, both translations generally succeed in representing events and actions. However, at the interpersonal level, the machine translation often fails to encode appropriate attitudes and relationships, resulting in flattened emotional expression. At the textual level, weaknesses in thematic progression and emphasis reduce narrative coherence, particularly in climactic and reflective passages. These findings support Steiner and Yallop's contention that effective translation must operate across multiple linguistic strata rather than focusing solely on propositional meaning.

The study also highlights important implications for the use of machine translation in educational contexts. In settings where students rely on MT for literary texts, there is a risk that surface fluency may be mistaken for quality, leading to misinterpretation of tone, theme, and character motivation. This is particularly concerning in literature education, where interpretive depth is a central learning outcome. The findings suggest that educators should treat MT output as a preliminary aid rather than a reliable substitute for human translation or critical reading.

At a broader level, the results underscore the ethical and cultural dimensions of translation technology. Literary texts are carriers of cultural values, historical memory, and aesthetic tradition. When such texts are translated mechanically without interpretive sensitivity, there is a risk of cultural flattening and loss of nuance. Human translators, through their interpretive labor, mediate between cultures in ways that machines currently cannot replicate.

The findings of this study contribute to an increasingly nuanced understanding of machine translation quality, particularly in relation to literary texts and low-resource language pairs such as English–Urdu. While recent advancements in neural machine translation have significantly improved grammatical fluency and sentence-level semantic adequacy, the present analysis demonstrates that these improvements do not extend uniformly to higher-order textual and cultural dimensions. Literary translation, by its very nature, operates at the intersection of language, culture, and interpretation, domains in which human translators retain a decisive advantage.

One of the most salient observations concerns the disconnect between automatic metric performance and perceived literary quality. Metrics such as BERTScore, BLEURT, and COMET indicate relatively strong semantic similarity between the machine-generated translation and the human reference. However, qualitative analysis reveals that semantic similarity alone is insufficient to guarantee interpretive fidelity. This finding reinforces recent critiques suggesting that learned metrics, while superior to surface-level overlap measures, remain limited in their ability to evaluate discourse coherence, pragmatic meaning, and stylistic intent (Freitag et al., 2021; Vernikos et al., 2022).

In literary prose, meaning is often distributed across sentences and realized through narrative progression rather than isolated lexical choices. Irony, understatement, and emotional restraint—hallmarks of O. Henry's narrative style—emerge through cumulative textual effects rather than explicit markers. Machine translation systems, which predominantly operate at the sentence level, are ill-equipped to capture such effects. Even when semantic content is

preserved, the loss of narrative rhythm and tonal consistency undermines the literary experience for the target reader.

The results also underscore the continuing relevance of classical translation theory in evaluating modern translation technologies. Nida's concept of dynamic equivalence provides a particularly useful lens for understanding the observed differences between human and machine translations. While the machine translation frequently achieves formal correspondence, it often fails to reproduce the communicative effect of the source text. The human translator, by contrast, demonstrates sensitivity to audience expectations, cultural norms, and narrative intent, adapting linguistic structures to preserve emotional and rhetorical impact. This confirms that dynamic equivalence remains a central criterion for evaluating literary translation quality, even in an era dominated by neural models.

Similarly, the systemic-functional framework proposed by Steiner and Yallop illuminates the limitations of machine translation across linguistic strata. At the ideational level, the machine translation performs reasonably well, accurately representing events and actions. However, at the interpersonal level, it frequently fails to encode appropriate attitudes, politeness strategies, and relational meanings, leading to flattened character interactions. At the textual level, weaknesses in thematic organization and emphasis disrupt narrative coherence, particularly in reflective passages that frame the story's moral conclusion. These deficiencies are largely invisible to automatic metrics, which further underscores the necessity of human-centered evaluation.

The implications of these findings extend beyond literary translation to broader debates about the role of artificial intelligence in language mediation. While machine translation has undeniably expanded access to information and facilitated cross-linguistic communication, its limitations raise important ethical and pedagogical questions. In educational contexts, reliance on MT for literary texts may inadvertently discourage critical engagement and interpretive analysis. Students may accept fluent machine output as authoritative, overlooking subtle distortions in tone, theme, and cultural meaning. This risk is particularly acute in multilingual societies where MT tools are increasingly integrated into everyday academic practice.

From a cultural perspective, the findings highlight the potential consequences of applying MT indiscriminately to literary and culturally significant texts. Literature functions as a repository of cultural memory and aesthetic tradition. When such texts are translated without interpretive sensitivity, there is a risk of cultural homogenization and loss of nuance. Human translators,

through their interpretive labor, act as cultural mediators who negotiate meaning across linguistic and ideological boundaries. This role cannot yet be replicated by machine systems, regardless of their computational sophistication.

## Conclusion

This study set out to compare human and machine translations of O. Henry's *The Gift of the Magi* from English into Urdu using a mixed-method approach that integrates qualitative literary analysis with quantitative machine translation evaluation metrics. The findings demonstrate that while Google Translate performs reasonably well in preserving sentence-level semantic content, it consistently underperforms human translation in reproducing literary tone, cultural nuance, discourse coherence, and interpretive depth.

The quantitative results reveal that semantic similarity-based metrics such as BERTScore and COMET yield relatively high scores, suggesting adequate meaning preservation. However, overlap-based metrics such as BLEU and TER highlight substantial divergence from the human reference translation, indicating significant post-editing requirements. Crucially, qualitative analysis shows that even high-performing metrics fail to capture essential literary qualities, including irony, emotional pacing, and narrative voice.

These findings reaffirm the enduring relevance of human translation in literary contexts and challenge claims that neural machine translation can function as an autonomous substitute for human expertise. Instead, the evidence supports a complementary model in which machine translation serves as a preliminary aid or drafting tool, while human translators ensure interpretive accuracy, stylistic integrity, and cultural appropriateness.

## References

Bar-Hillel, Y. (1960). The present status of automatic translation of languages. *Advances in Computers*, *1*, 91–163. https://doi.org/10.1016/S0065-2458(08)60516-0

Bassnett, S. (2014). *Translation studies* (4th ed.). Routledge.

Berman, A. (1985). Translation and the trials of the foreign. In L. Venuti (Ed.), *The translation studies reader* (2nd ed., pp. 284–297). Routledge.

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, *19*(2), 263–311.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734. https://doi.org/10.3115/v1/D14-1179

Freitag, M., Foster, G., Grangier, D., Ratnakar, V., & Tan, Q. (2021). Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, *9*, 1460–1474. https://doi.org/10.1162/tacl_a_00419

Freitag, M., Rei, R., Mathur, P., Lo, C., Stewart, C., Avramidis, E., ... Bojar, O. (2022). Results of the WMT22 metrics shared task. *Proceedings of the Seventh Conference on Machine Translation (WMT)*, 46–68. https://doi.org/10.18653/v1/2022.wmt-1.3

Hutchins, J. (2004). *Machine translation: A concise history*. University of Cambridge, Computer Laboratory.

Jakobson, R. (1959). On linguistic aspects of translation. In R. A. Brower (Ed.), *On translation* (pp. 232–239). Harvard University Press.

Koehn, P. (2009). *Statistical machine translation*. Cambridge University Press. https://doi.org/10.1017/CBO9780511815829

Munday, J. (2016). *Introducing translation studies: Theories and applications* (4th ed.). Routledge.

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

Papineni, K., Roukos, S., Ward, T., & Zhu, W. (2002). BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. https://doi.org/10.3115/1073083.1073135

Popović, M. (2015). chrF: Character n-gram F-score for automatic MT evaluation. *Proceedings of the Tenth Workshop on Statistical Machine Translation*, 392–395. https://doi.org/10.18653/v1/W15-3049

Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET: A neural framework for machine translation evaluation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2685–2702. https://doi.org/10.18653/v1/2020.emnlp-main.213

Reiss, K., & Vermeer, H. J. (1984). *Grundlegung einer allgemeinen Translationstheorie*. Niemeyer.

Sellam, T., Das, D., & Parikh, A. P. (2020). BLEURT: Learning robust metrics for text generation. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7881–7892. https://doi.org/10.18653/v1/2020.acl-main.704

Steiner, E., & Yallop, C. (2001). Exploring translation and multilingual text production. *Language Sciences*, *23*(4–5), 413–438. https://doi.org/10.1016/S0388-0001(00)00031-1

Sudiati, I. (2005). Translation as a process of meaning transfer. *TEFLIN Journal*, *16*(2), 125–139.

Toury, G. (1995). *Descriptive translation studies and beyond*. John Benjamins. https://doi.org/10.1075/btl.4

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 5998–6008.

Venuti, L. (1995). *The translator's invisibility: A history of translation*. Routledge.

Vernikos, G., Thompson, B., Mathur, P., & Federico, M. (2022). Document-level machine translation metrics: A critical review. *Proceedings of the Seventh Conference on Machine Translation (WMT)*, 1–15. https://doi.org/10.18653/v1/2022.wmt-1.1

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. *International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=SkeHuCVFDr