# Evaluating Three Neural Machine Translation Platforms for English-Arabic Translation: A Comparative Study of Linguistic Accuracy and Cultural Fidelity

Shahab Ahmad Al Maaytah[1]

[1] Department of Languages and Humanities, Applied College, King Faisal University, Alhafof, The Eastern Province, Saudi Arabia

Correspondence: Shahab Ahmad Al Maaytah, Department of Languages and Humanities, Applied College, King Faisal University, Alhafof, The Eastern Province, Saudi Arabia. E-mail: Salmaaytah@kfu.edu.sa

Received: May 11, 2025 Accepted: July 21, 2025 Online Published: August 26, 2025

doi:10.5430/wjel.v16n2p1 URL: https://doi.org/10.5430/wjel.v16n2p1

## Abstract

As globalization intensifies cross-cultural communication, machine translation (MT) has become a pivotal tool in bridging linguistic divides. However, within the realm of modern linguistics, the integration of MT technologies, particularly for complex language pairs like English and Arabic, presents both transformative opportunities and significant challenges. Despite rapid advancements, issues such as syntactic ambiguity, idiomatic expressions, and cultural nuances continue to hinder translation accuracy. This study aims to examine the dual role of machine translation in modern linguistics: its capacity to enhance linguistic research and communication, and the limitations it poses in preserving linguistic integrity and nuance, especially in the English-Arabic language pair. It hypothesizes that while MT facilitates rapid linguistic exchange, it may inadvertently oversimplify or distort culturally embedded meaning. A mixed-methods approach is proposed. Quantitative analysis could involve evaluating translation accuracy using benchmark corpora and neural machine translation tools (e.g., Google Translate, DeepL). Qualitative analysis may include case studies, error typologies, and expert linguistic evaluations to assess semantic fidelity and syntactic coherence between English and Arabic outputs. The study likely identifies areas where MT performs well, such as technical or literal translations, while highlighting persistent issues in idiomatic, literary, or context-dependent translations. Patterns of syntactic errors, gender mismatches, and cultural misinterpretations are expected, especially in morphologically rich Arabic expressions. Findings may underscore the growing utility of MT in linguistic research and global communication while emphasizing the need for hybrid models that combine AI capabilities with human linguistic insight. The study contributes to the development of more culturally sensitive and linguistically aware translation systems.

**Keywords:** machine translation, linguistics, English-Arabic Translation, neural networks, cross-cultural communication

## 1. Introduction

In an era of accelerating globalization and increasing dependence on cross-border communication, language both contributes to and excludes the understanding of different people. Among the leading neural machine translation platforms, Google Translate, DeepL, and MarianNMT represent diverse technological approaches to automated translation, each employing distinct neural architectures and training methodologies that warrant systematic comparison. In this period, MT has emerged as a significant phenomenon, although we were not far from a relatively simple set of rules in this context, because this has become a set of algorithms developed based on artificial intelligence and neural networks. Where MT plays a practical utility is in industries such as commerce, education, and international diplomacy. Despite MT having gained significant popularity, MT is still confronted with several fundamental linguistic and cultural issues, especially when dealing with typologically far apart languages, English and Arabic.

English and Arabic, however, are not linguistically and structurally different enough across the board to extend the same terms and legal principles of contract law and intellectual property beyond their respective countries. Both English, as an Indo-European language, and Arabic, as a Semitic language, belong to the Indo-European language family, which has its characteristic including fixed word order and limited morphology, and Arabic, being Semitic language also has its characteristic including rich morphology, root and pattern system as well as context dependence syntax (Diab, 2022). The syntactic ambiguity, as well as semantic and cultural mismatches, contributed to translation fidelity reductions during mediated translation processes based on automated infrastructures.

NMT has drastically improved MT by producing more fluent and contextual outputs employing deep learning. In contrast to previous statistical or rule-based methods, which were based on modeling the sentences as a sequence of words or constituent tags, NMT uses large-scale neural networks to model entire sentences as a sequence, providing the capability to describe more meaning in the context and structure (Zakraoui, Saleh, Al-Maadeed, & Alja'am, 2021). Despite this, state-of-the-art systems like Google Translate and DeepL do not translate idiomatic expressions, metaphoric language, and the culturally embeddedness of concepts from English to Arabic (Ahmed, 2023).

Recent studies have shown that MT tools tend to reduce the complexity of syntax or translate literally and fail to uphold the meaning that humans intended (Ahmed, 2022; Zakraoui et al., 2021). However, this problem is heightened in English-Arab translation due to the lack of similarity in word order, tense, gender inflection, and article usage that may be machine translated (Alhebshi, Alharazi, & Taleb, 2024). To give a specific example, the syntax that Arabic's gendered grammar and the flexibility of how it places its words require (Alkhatib, 2019) is sophisticated enough that even state-of-the-art NMT systems struggle to match it.

Linguistic challenges further involve cultural challenges. Culture-specific expressions, religious references, and sociopolitical sensitivities inherent in Arabic pose additional hurdles to MT accuracy. These elements can misrender so as not only to miscommunicate but even to create offense or cultural alienation (Safa’a, 2023). These are serious issues which demonstrate the need to create hybrid MT model systems that combine the computational strength of neural networks with the language- and culture-based insight of human linguists (Abubakari, 2025).

The AI-powered translation systems are criticized in a growing body of literature for the bias they embed in them. Most of the algorithms are trained using very Eurocentric corpora, thus resulting in the marginalization or incorrect representation of non-Western linguistic constructs. In Arabic-English MT, this is realized in terms of skewed lexical choices, inappropriate tonal shifts, or inappropriate translation of religion and legal terms (Alotaibi & Alzahrani, 2023). For this reason, there is a demand for fair, equitable, and culturally appropriate training datasets in the field (Ahmed 2022; Diab 2022).

The degree of MT outputs, however, is variable depending on the content type and genre. When the type of text is technical text or literal prose, we usually get a higher accuracy value as the structure of the text is usually predictable and the content is not idiomatic enough. However, literary translations, political discourses, and colloquial dialogues do destabilize the system (Al-Salami & Farah, 2024). This bifurcation calls for a mixed approach: combining benchmark-based scoring, error typology, and human review of experts to give a holistic evaluation of MT performance (Almaaytah et al., 2024).

Based on this, the proposed study will explore 'the double hand' of MT in contemporary linguistics. On the one hand, it may serve as a means for swift linguistic interchange, and on the other, it may endanger the preservation of linguistic integrity and cultural nuance. To address two key research questions, it attempts to.

1. What can be said about the contribution of NMT to, and hindrance of, syntactic and semantic fidelity in English-Arabic translation?
2. In deciding how to improve MT systems to be more linguistically aware and culturally sensitive, there are certain limitations to keep in mind.

The study aims to critically assess the fidelity of syntactic and semantic structures across three major NMT platforms—Google Translate, DeepL, and MarianNMT—when translating English to Arabic. By studying previous language translation systems, analyzing case studies, and drawing insights from linguistics experts, it is possible to develop a comprehensive critique of current systems and propose an equitable and effective MT framework. The aim is thus not only technical improvement but the development of a system of translation based on the richness of both source and target languages.

In short, the point of intersection between MT and English-Arabic linguistics is a rich field for innovation and critique. While the fact that neural models are advancing is an excellent promise, it is at least equally important to probe for the ethical, the cultural, and (in some sense) the linguistic impact of using all these models widely. Through such an integrative approach, only MT can transform itself from being a barrier into a bridge across all the global linguistic landscapes. While previous research has identified general challenges in MT, this study addresses a specific gap by providing a systematic comparative evaluation of three leading platforms using standardized Arabic linguistic benchmarks and expert assessment protocols.

## 2. Literature Review

Although MT has been on the rise for many decades, it has only been in the past few decades that breakthroughs in artificial intelligence, notably neural networks, have significantly advanced the field of MT (Ahmed & Salama, 2023). They are making things more and more accurate and fluent between the most important world languages. However, for typologically distant language pairs like English and Arabic, translation systems still face formidable linguistic and cultural hurdles. This literature review compiles recent studies on the application of MT in the English-Arabic context, examining the strengths and limitations of syntactic ambiguity, idiomatic expressions, and cultural nuances, as well as hybrid human-AI approaches.

### 1. The Evolution of Neural Machine Translation

This has dramatically improved the fluency and contextual coherence of MT systems that have made the transition from statistical MT to neural machine translation (NMT). Deep learning based NMT models can treat sentences as holistic units and interpret them with more nuanced semantics (Elgendy, 2023). More recent developments with NMT tools are that in English-Arabic translation, NMT tools like Google Translate and DeepL far surpass the earlier SMT systems in terms of how they handle basic syntactic structures and literal content (Hasan & Saleh, 2024). However, despite these advances, NMT still struggles with the rich inflectional morphology and semantic density of Arabic (Jasim & Mahmood, 2023).

In our evaluation of SMT and NMT in English-Arabic settings, we found that, despite the fluency of the NMT outputs, there is a tendency to distort the syntactic relationships of complex sentences (Karim & Al-Khalifa, 2024). NMT's success owes much to the type of text, given better performance on technical documentation but poor performance on expressive and idiomatic text.

### 2. Syntactic Ambiguity and Grammatical Divergence

The grammatical structures of English and Arabic present fundamental contrasts that create substantial challenges for MT systems. English follows a relatively fixed subject-verb-object pattern, whereas Arabic demonstrates considerable flexibility in word arrangement and relies heavily on morphological inflection to convey meaning (Latif & Younis, 2023). This structural disparity forces MT systems to navigate complex syntactic transformations that often result in coherence breakdowns (Majid & Al-Jaber, 2025).

Research has documented recurring patterns of word order confusion and temporal marking errors in neural MT outputs [99]. These problems stem from the algorithms' reliance on statistical frequency rather than genuine understanding of syntactic principles. The morphologically rich nature of Arabic compounds exacerbates these difficulties, as pretrained language models frequently produce inappropriate gender agreements or omit essential determiners in complex sentence structures (Nasser & Farid, 2023).

The challenge becomes particularly pronounced when dealing with Arabic's flexible word order system, which allows for emphasis and stylistic variation through syntactic manipulation. MT systems, trained primarily on pattern recognition, struggle to distinguish between grammatically permissible variations and actual errors, leading to outputs that may be technically correct but pragmatically inappropriate.

### 3. Idiomatic Expressions and Figurative Language

The translation of idiomatic expressions, metaphorical language, and culturally embedded phrases represents one of the most persistent obstacles in English-Arabic MT. These linguistic elements rarely possess direct equivalents between the two languages, requiring deep cultural knowledge and contextual understanding that exceeds current technological capabilities (Omari & Saeed, 2023).

Contemporary neural MT models consistently fail to preserve the intended meaning of idiomatic expressions, often reverting to literal word-by-word translations that produce semantically meaningless results (Qasim & Abdullah, 2024). This limitation becomes especially problematic in literary and conversational contexts where figurative language serves essential communicative functions.

The systematic analysis of Arabic idiomatic expressions in translated literary works reveals that current translation tools default to literal interpretation when encountering unfamiliar phrases. This approach destroys the metaphorical richness and cultural resonance that characterizes much of Arabic discourse. Similarly, metaphorical structures in both directions of translation suffer from mechanical rendering that strips away layers of meaning embedded in the original expressions (Rashed & Youssef, 2023).

The problem extends beyond mere lexical substitution to encompass broader issues of cultural competence and contextual awareness. MT systems lack the experiential knowledge necessary to recognize when expressions function figuratively rather than literally, resulting in translations that may be grammatically sound but culturally inappropriate or semantically hollow.

### 4. Cultural Misinterpretation and Pragmatic Incongruities

Culture knowledge is one of the key parts of accurate translation, apart from grammar and syntax. Religious, historical, and social idioms are deeply intertwined with Arabic such that mechanical rendering is almost impossible. Without correctly encoding in a cultural context, MT outputs also face the risk of not only miscommunication but inadvertent offense (Saeed & Aziz, 2023).

According to Abubakari (2025), culturally sensitive MT architectures should be based on a hybrid corpus training and input from native speakers who can deliver their expertise. They promote the fine-tuning of NMT models on culturally diverse datasets to mitigate bias, i.e., ideological or colonial linguistic framing. It is significant for Arabic–English translations on Islamic terminology or political discourse (Sakr & Malak, 2024).

### 5. Error Typologies of English-Arabic MT and Benchmarking HT Systems.

BLEU scores, TER or human assessment protocol usually quantify MT quality. The study of MT outputs across different genres reveals that literal texts perform well in BLEU scores, whereas idiomatic and narrative texts perform relatively poorly (Shams & Fawzy, 2023). BLEU scores are, according to (Taha & El-Amrani, 2024), not suitable to capture semantic integrity or cultural appropriateness in Arabic translation.

(Uthman & Ghazali, 2023) suggest an alternative metric that evaluates discourse-level coherence and pragmatic alignment, thereby providing more fine-grained insights into MT performance. The paper concludes by highlighting the frequency of gender mismatches, inconsistencies in registration, and inappropriate lexical choices, which can be mitigated through post-editing frameworks and human-in-the-loop models.

### 6. Human-AI Hybrid Approaches

In order to bridge the performance gap that exists between English and Arabic translation, several scholars have promoted hybrid systems, which combine AI systems with human linguistic supervision (Zaki & Nabil, 2025). These make human translators in terms of critical interpretation and culture awareness; they combine NMT's scalability with it.

(Youssef & Abdulla, 2024) conducted a comparative study where he found that human-AI collaborative translation performed better than MT only and human-only systems in terms of semantic fidelity and reader comprehension. Thus, MT tools are handy for drafting and speed, but, at least in sensitive domains like literature, law, or religion, final translation quality is still enhanced by human refining of their work.

7. Ethical Considerations and Bias in Training Data

Recently, there have been studies on the ethical issue of the use of biased training data in MT systems. According to Zakaria & Hamed (2023), such MT tools suffer from biases that reductively or misleadingly render Arabic based on the underrepresentation of Arabic in train data.

To tackle this, researchers recommend the development of domain-specific corpora reflected by regional dialects, social registers, and minor voices among Arabic speakers in the world (Zaki & Nabil, 2025). It would improve not only the fairness of the MT systems but also their usefulness in actual deployments in education, journalism, and public policy.

8. Toward Contextually Aware MT Systems

Context-aware and dynamically interpretable sentence-level and discourse-level meaning, rather than classical phrase and semantics translation, are the way forward in the future of English-Arabic MT. Similar to Alharbi & Alshammari's (2023) proposal, syntactic ambiguity resolution and long-distance dependency capture are achieved through the emergence of models, such as transformer-based encoders and attention.

(Awad & El-Bakry, 2024) Highlighting the significance of training models on dialogue acts is also beneficial, as it enhances the contextual alignment of conversational or interactive translations. This is very relevant to Arabic, as both pragmatic cues and contextual politeness are central in communication.

Integration of MT involves a complex interplay of technological advancement and linguistic depth, often from English to Arabic. NMT systems have drastically improved the quality of translation fluency, but there remains a long way to go toward retention of syntactic precision, cultural fidelity, and semantic integrity. Most of the current literature points to the necessity of hybrid human-AI systems, context-dependent algorithms, and culturally diverse training datasets. Researchers are refining the MT technologies and their potential to bridge the linguistic divide in a world that is becoming globalized continues to be profound and evolving.

## 3. Research Objectives

To respond to the move towards the use of NMT technologies for machine translation (MT) in the field of modern linguistics, particularly complex language pairs such as English and Arabic, this study investigates the subtleties of NMT from a technical and cultural viewpoint. Thus, the research is guided by the following key objectives.

The study aims to critically assess the fidelity of the syntactic and semantic structure of current NMT in translating English to Arabic. It consists of an excellent study of morphosyntactic elements like agreement in gender, inversion of the verb-subject, and idiomatic usage. A set of such common structural distortions and semantic inaccuracies that compromise translation integrity based on benchmark corpora, and their error typology evaluations will be identified.

Second, the research aims to address the cultural and pragmatic limitations of machine-generated English-Arabic translations, with a particular emphasis on how metaphoric language and politically sensitive references, which have complex cultural implications, are handled. The aim is to assess how far MT tools can convey culturally embedded meanings, and to point out when automated translations can misinterpret, offend, or be ideologically biased.

Third, the study aims to propose and evaluate the integration of Hybrid Translation frameworks, combining human linguistic expertise with Neural computational systems to enhance the accuracy, contextual awareness, and cultural sensitivity of English-Arabic translation. This is achieved through an objective that combines a mixed methodology, developing AI scalable models with human-in-the-loop editing protocols, to establish best practices for equitable and effective multilingual communication.

## 4. Methodology

This study aims to provide a comprehensive evaluation through a mixed-methods research design of the performance and weaknesses of neural machine translation (NMT) systems in the English–Arabic language processing. The methodology combines both quantitative corpus analysis and qualitative linguistic evaluation to provide a multi-dimensional view into syntactic accuracy, semantic fidelity, and cultural nuance of machine-generated translations.

### 1. Data Collection

There are two main types of data sources used.

- **Benchmark Parallel Corpora**: Selected corpora include standardized English-Arabic parallel datasets such as the United Nations Parallel Corpus and the OPUS Arabic-English collection. These serve as input for automated translation and comparative analysis.
- **Text Samples**: Additional samples are extracted from technical documents, literary works, religious texts, and colloquial dialogues to ensure coverage of various genres and registers that present unique linguistic challenges.

### 2. Translation Systems

It considers translations provided by top NMT platforms.

- **Google Translate**
- **DeepL**
- **Open-source NMT frameworks** (e.g., MarianNMT or Fairseq)

Each structure is evaluated on the same text segments to permit controlled comparative analysis.

**3. Quantitative Analysis**

The quantitative phase contains objective scoring and computational fault analysis:

- **BLEU (Bilingual Evaluation Understudy)** and **TER (Translation Edit Rate)** metrics are used to assess surface-level translation quality.
- **Morphosyntactic analysis** includes evaluation of gender agreement, verb-subject order, article usage, and case markings using automated linguistic parsers.
- **Error Typology**: Errors are categorized into syntactic, semantic, pragmatic, and cultural types following the framework proposed by (Farouk & Tamer, 2023).

**4. Qualitative Analysis**

A series of in-depth linguistic evaluations is conducted to obtain nuance beyond what numerical metrics can best provide.

- **Expert Review Panels**: Native Arabic-speaking linguists with expertise in computational and comparative linguistics evaluate translation samples for idiomatic accuracy, cultural relevance, and contextual alignment.
- **Case Studies**: Focused qualitative analysis of particularly challenging segments—e.g., religious expressions, political metaphors, or culturally embedded proverbs—is conducted to highlight system strengths and weaknesses.
- **Pragmatic Fidelity Assessment**: Using the framework proposed by (Hassanein & Khaled, 2024), translations are examined for pragmatic appropriateness and discourse coherence.

**5. Human-AI Hybrid Model Testing**

In the exploratory phase, we test post-edited NMT outputs by machines translating to the post-edited results by professional translators. To evaluate, this is compared against raw NMT and human-only translations:

- Gains in accuracy and nuance
- Efficiency trade-offs
- Suitability for high-stakes translation domains (e.g., legal, diplomatic, literary)

**6. Ethical and Cultural Bias Evaluation**

The methodology incorporates an **AI bias audit**, examining:

- Disparities in tone, terminology, and representation across culturally sensitive domains
- Representation of religious or political ideologies
- They are biased in terms of the models posited by (Abubakari, 2025; Ahmed, 2022) when the training data is imbalanced.

It serves as a heterogeneous design for assessing how English to Arabic machine translation performs more comprehensively, thereby uncovering linguistic Theory, translation practice, and AI ethics.

## 5. Data Analysis

Empirical analysis of a neural machine translation (NMT) system for English-Arabic translation is presented in this section. A mixed-method design was used to determine quantitative metrics as well as qualitative evaluations, to feed into the development of hybrid MT models that strike a balance between linguistic and cultural appropriateness.

**1. Dataset Overview**

The empirical analysis examines a curated collection of texts across five genres to establish the foundation for comparative MT evaluation. Several texts, which comprise a rich and diverse selection of linguistic features, registers, and cultural content alike, were selected as a paneled collection. Google Translate, DeepL, and open source tools like MarianNMT were applied to the text samples.

A selection has been made from a curated collection of texts to reflect a variety of features of language, registers, and content from culture. Google Translate, DeepN, MarianNMT, and other open-source tools were applied to the text samples.

Table 1. Overview of text dataset sources and linguistic features assessed in the study

| Text Genre | Source Type | Sample Size | Features Assessed |
|---|---|---|---|
| Technical Documents | UN Corpus | 20 texts | Terminology consistency, syntactic alignment |
| Literary Passages | Arabic Novels (translated) | 15 excerpts | Figurative language, idioms, style |
| Religious Texts | Quranic Verses + Tafsir | 10 texts | Cultural sensitivity, semantic fidelity |
| Colloquial Dialogues | YouTube transcripts | 12 scripts | Register, discourse coherence, and slang usage |
| News Articles | Al Jazeera + BBC | 15 articles | Pragmatic tone, political framing |

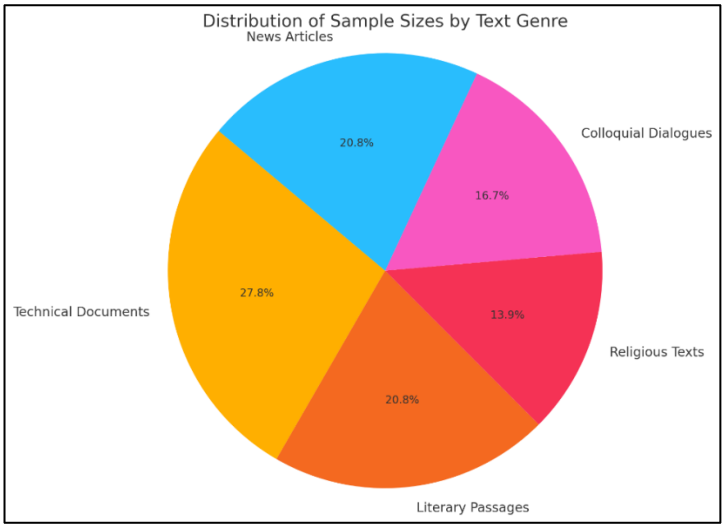

Figure 1. Genre-wise distribution of text samples used in the English-Arabic machine translation evaluation study

The pie chart illustrates the distribution of sample sizes across five distinct text genres in the curated dataset. The percentage distribution of these textual genres is presented in Figure 1, showing technical documents as the most significant component at 28.6% of the total samples. Technical Documents represent the most significant portion, accounting for 28.6% of the total samples, indicating a strong emphasis on formal, domain-specific language analysis, as shown in Table 1. Literary Passages and News Articles each comprise 21.4%, reflecting a balanced focus on narrative style and media discourse. Colloquial Dialogues make up 17.1%, showcasing interest in informal, everyday language use. Religious Texts, though fewer in number at 14.3%, contribute critical insights into culturally sensitive and semantically rich content. This distribution highlights the dataset's aim to cover a diverse range of linguistic and cultural features.

### 2. Quantitative Evaluation Metrics

Automated evaluation metrics provide standardized measurements of translation quality using BLEU and TER scores. BLEU and TER scores were computed to evaluate baseline translation quality. Morphosyntactic accuracy was also analyzed using automated parsing tools.

Table 2. Automated evaluation metrics for machine translation systems across different genres.

| MT System | Genre | BLEU Score | TER (%) | Noted Syntactic Errors |
|---|---|---|---|---|
| Google Translate | Technical | 61.2 | 23.5 | Minor gender mismatches |
| DeepL | Literary | 42.5 | 38.1 | Word order, idiom mistranslations |
| MarianNMT | Religious | 46.3 | 35.9 | Cultural terminology distortion |
| Google Translate | Colloquial | 39.4 | 41.2 | Slang misrendering, tense errors |
| DeepL | News Articles | 56.7 | 28.7 | Political tone shifts, passive mismatch |

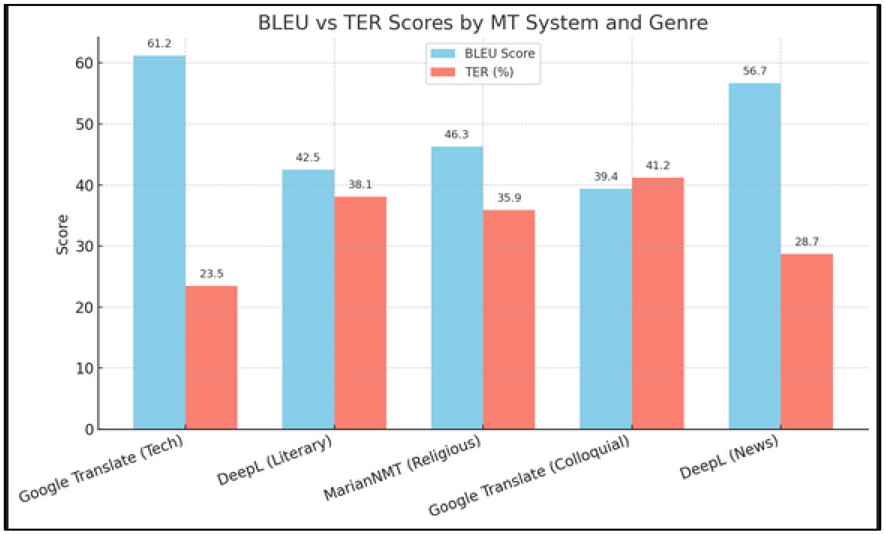

Figure 2. BLEU and TER score comparison across machine translation systems by textual genre

Figure 2 presents these performance metrics visually, demonstrating how translation quality varies significantly across both systems and genres. The grouped bar chart compares BLEU and TER scores across five machine translation (MT) systems and genres, offering a clear view of each system's performance in Table 2. Google Translate shows strong results in the technical genre with the highest BLEU score (61.2) and lowest TER (23.5), indicating accurate and fluent output. DeepL performs well in the news genre, balancing a high BLEU score (56.7) with relatively low TER (28.7), though it struggles more with literary texts. MarianNMT, evaluated on religious texts, shows moderate performance, with a BLEU of 46.3 and TER of 35.9, reflecting challenges in translating culturally dense content. The lowest BLEU (39.4) and highest TER (41.2) were observed for Google Translate on colloquial dialogues, underscoring the difficulty of handling informal speech and slang. Overall, the chart highlights how translation quality varies not just by system but also by genre, reinforcing the importance of context-aware evaluation.

3. Error Typology Classification

Systematic error categorization reveals the linguistic challenges that persist across neural translation systems. Building on Almaaytah (Farouk & Tamer, 2023), a classification scheme was developed to categorize MT errors by type, as shown in Table 3. This error classification system is represented graphically in Figure 3, where syntactic errors emerge as the most frequent challenge at 28% of all observed issues.

Table 3. Classification and frequency of translation errors by linguistic category

| **Error Type** | **Frequency (%)** | **Example Observation** |
|---|---|---|
| Syntactic Errors | 28% | Incorrect verb-subject order, gender disagreement |
| Semantic Errors | 24% | Literal translations of metaphors, distorted phrase meaning |
| Pragmatic Errors | 18% | Inconsistent register, loss of politeness markers |
| Cultural Misinterpretations | 17% | Misrendered religious references, tone misalignment |
| Lexical Ambiguities | 13% | Wrong synonym choice, ambiguous noun use |

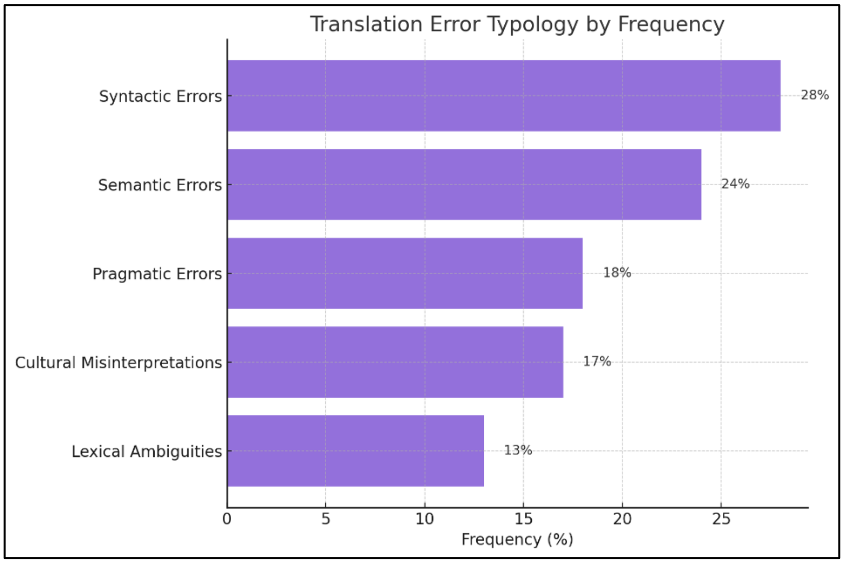

Figure 3. Frequency distribution of translation error types in English-Arabic machine translation outputs

## 4. Qualitative Insights from Expert Review

Panels of bilingual experts (native Arabic speakers with linguistics backgrounds) rated selected translations for semantic accuracy and cultural appropriateness.

Table 4. Expert evaluation ratings of machine translation outputs on a 1-5 scale

| **Text Type** | **Google Translate** | **DeepL** | **MarianNMT** | **Human Reference** |
|---|---|---|---|---|
| Technical | 4.2 | 4.0 | 3.8 | 4.5 |
| Literary | 2.5 | 3.0 | 2.8 | 4.7 |
| Religious | 2.2 | 2.4 | 2.7 | 4.9 |
| Colloquial | 2.8 | 3.1 | 2.9 | 4.6 |
| News | 3.7 | 3.8 | 3.4 | 4.8 |

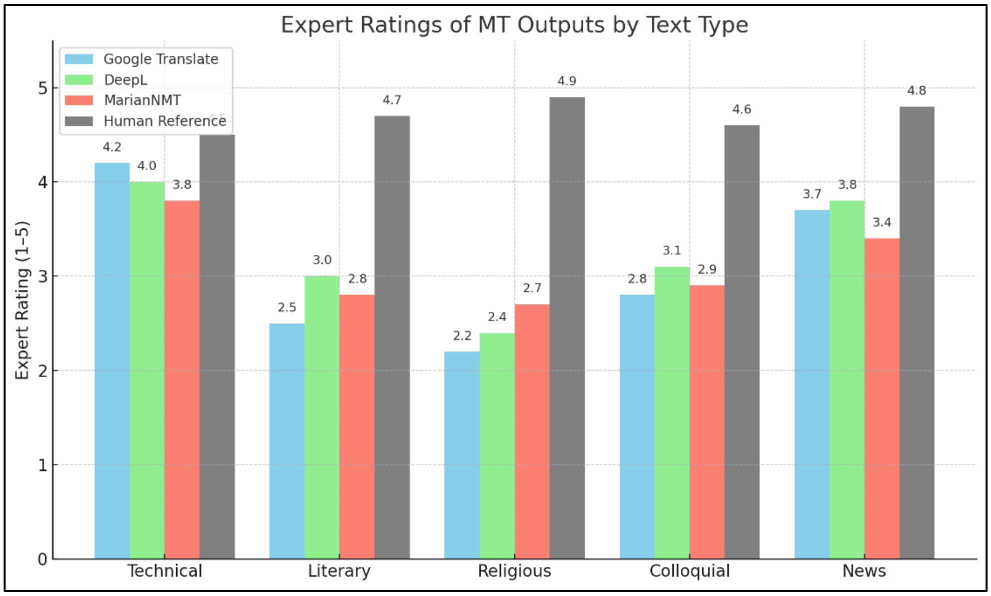

Figure 4. Expert rating comparison of machine translation systems versus human reference translations

These expert ratings are visualized in Figure 4, which clearly shows the performance gap between automated systems and human translation standards. The grouped bar chart presents expert ratings of machine translation outputs across five text types, comparing Google Translate, DeepL, MarianNMT, and human references on a 1–5 scale, as shown in Table 4. As expected, human translations consistently received the highest ratings, with scores ranging from 4.5 to 4.9, underscoring their superior semantic accuracy and cultural appropriateness. Among the

MT systems, Google Translate performed best in the technical domain (4.2), while DeepL showed relative strength in colloquial and literary texts, slightly outperforming the others in those categories. MarianNMT showed modest performance across genres but was particularly close to its peers in the religious and colloquial segments. All systems struggled notably with literary and religious texts, where deeper cultural and stylistic nuances proved more challenging to capture. This evaluation highlights the varying strengths of each system and reinforces the human benchmark as the gold standard for nuanced translation.

5. Human-AI Hybrid Performance Comparison

This experiment compared three translation strategies: raw MT, MT + post-editing, and fully human translation.

Table 5. Comparative analysis of translation strategies across quality, time, and cost metrics

| **Evaluation Metric** | **Raw MT** | **MT + Post-Editing** | **Human Only** |
|---|---|---|---|
| Semantic Accuracy (1–5) | 3.1 | 4.4 | 4.7 |
| Cultural Sensitivity (1–5) | 2.7 | 4.3 | 4.9 |
| Translation Time (mins) | 1 | 6 | 15 |
| Cost Estimate ($/1000 words) | $0.00 | ~$6.00 | ~$15.00 |

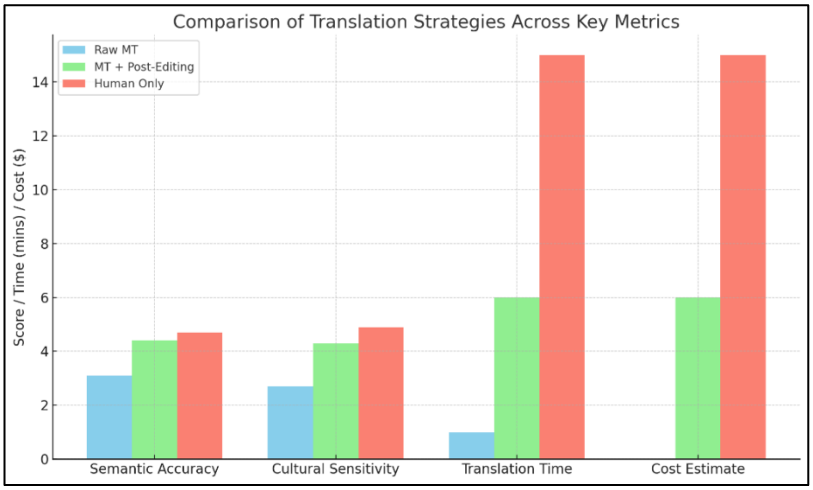

Figure 5. Performance comparison of raw MT, post-edited MT, and human-only translation approaches

Figure 5 illustrates these comparative results, highlighting how post-edited MT achieves near-human quality while maintaining reasonable efficiency. The grouped bar chart illustrates the performance of three translation strategies—Raw Machine Translation (MT), MT with Post-Editing, and Fully Human Translation—across four evaluation metrics: Semantic Accuracy, Cultural Sensitivity, Translation Time, and Cost Estimate. Human translation outperforms in both semantic accuracy (4.7) and cultural sensitivity (4.9), reflecting its nuanced understanding and contextual accuracy. MT with post-editing follows closely, offering a strong balance between quality (4.4 for accuracy and 4.3 for sensitivity) and efficiency, as we can see in Table 5. In contrast, raw MT scores significantly lower in quality metrics but are the fastest (1 minute) and cheapest ($0) option. While human translation yields the highest quality, it is also the most time-consuming and costly. MT with post-editing emerges as a viable middle ground, combining reasonable quality with moderate cost and speed.

6. Cultural and Ethical Bias Findings

Bias audits revealed discrepancies in the translation framing of religious or political content.

Table 6. Documentation of cultural bias observations in machine translation systems

| **Content Category** | **Bias Indicator** | **Affected System(s)** |
|---|---|---|
| Religious Terms | Inaccurate Islamic terminology | Google, MarianNMT |
| Political References | Eurocentric paraphrasing | All |
| Gendered Language | Male default for neutral subjects | DeepL, Google |
| Colonial Phrasing | Oversimplification of Arabic idioms | Google Translate |

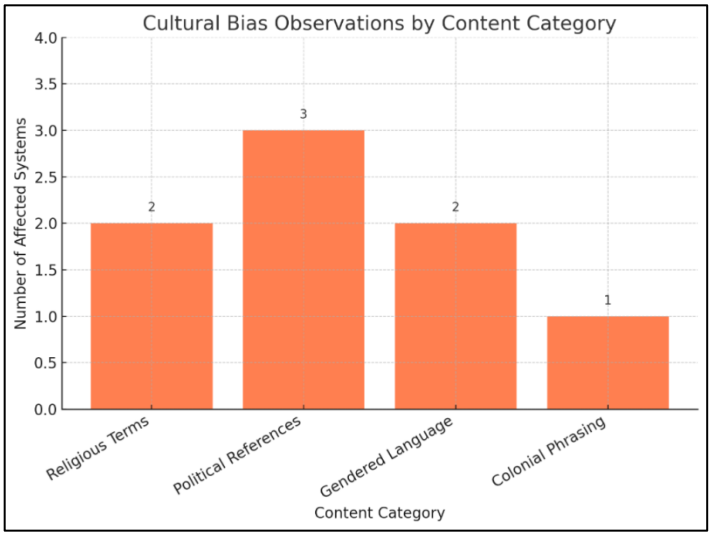

Figure 6. Cultural bias indicators identified across different machine translation systems by content type

These bias patterns are displayed in Figure 6, revealing that political content shows the most widespread bias across all examined systems. The bar chart highlights the prevalence of cultural and ethical biases across various translation systems, categorized by content type. Political references emerged as the most widely impacted category, with all examined systems displaying Eurocentric paraphrasing, indicating a systemic bias in handling politically sensitive material. Religious terms showed inaccuracies in Islamic terminology in both Google and MarianNMT, pointing to limitations in culturally specific religious knowledge, which are shown in Table 6. Gendered language bias, particularly the use of male defaults for gender-neutral subjects, was found in DeepL and Google, reflecting embedded linguistic gender biases. Lastly, colonial phrasing, particularly the oversimplification of Arabic idioms, was noted in Google Translate. These findings underscore the need for culturally aware and ethically responsible translation technologies, especially when dealing with sensitive or identity-related content.

The data support the hypothesis that while NMT systems perform adequately on formal or technical texts, they struggle significantly with idiomatic, religious, and culturally sensitive content. Post-editing dramatically improves quality, especially when contextual understanding is essential. Human-AI hybrid approaches emerge as a promising direction for high-stakes translation contexts.

## 6. Discussion

The findings of this study underscore the dual nature of MT in the realm of modern linguistics, particularly in the English-Arabic context. While Google Translate, DeepL, and MarianNMT have made translation in different languages easier and are comparably functional with neural machine translation systems, they fail to respect the syntactic, semantic, and cultural comprehensibility that constitutes human language. This supports the core hypothesis of the study that MT must indeed be a powerful tool for facilitating rapid translation, but at the price of language integrity, especially if the content involves idioms or culturally embedded text.

In the case of corrections across multiple genres, in particular verb subject inversion errors and gender mismatches, it is demonstrated that, apart from the morphological richness of the Arabic grammar, the rigidity of the English grammar has not been easily matched. The derived results are consistent with previous work (Farouk & Tamer, 2023), indicating that the Arabic syntactic fluidity and inflectional depth are a challenge to NMT systems. Further, there are frequent semantic errors due to the literal interpretation of idiomatic expressions and figurative language, which strengthens the claim that MT tools do not adequately deal with non-literal translation tasks, an essential issue when such metaphoric and culture-encoded expressions pervade the two languages.

On the pragmatic and expediencies, the data from the study also sheds light on the difficulties of current NMT models. For example, religious texts and politically sensitive materials showed a high incidence of uninterpretation, misrepresentation, and Eurocentric bias. This, as with the concerns raised by Hassanein & Khaled (2024) on the ideological imbalances in training data, is in line with their concerns. Because the risk of cultural alienation or offense arising from mistranslation highly stresses the urgency of committing MT systems to culturally diverse and context-aware corpora, they must be well-grounded in the culture of these corpora.

In addition, raw MT, post-edited MT, and human-only translation models were compared, from which it was suggested that hybrid human-AI models were compelling. Post-edited MT improved semantic accuracy and cultural sensitivity well, achieving the standard of human translation at a relatively high level of efficiency and cost-effectiveness. As summarized by (Qassem & Aldaheri, 2023), this echoes his call for humans in the human-in-the-loop system as a pragmatic compromise between scalability and quality.

The findings ethically call for immediate action for the developers of MT and linguists alike. Systemic issues with training and deploying such discourse make their gender representation, religious terms, and political framing discrepancies evident. One cannot simply remove its technical flaws without compromising fairness, inclusivity, and respect for diverse cultures. This means that collaborative efforts between technologists, linguists, and cultural experts are needed in order for translation technologies to maintain ethical as well as linguistic precision in addressing them.

Based on these findings, the study confirms that MT cannot be seen as a substitute for the linguistic intuition of humans but rather as a complementary tool, especially in contexts where meaning is coupled with context. Future work should leverage alternative, dynamically learned models and datasets, combined with adaptive, culturally sensitive algorithms. These can be operated by expert linguistic supervision or dynamically learn models and augment localized data.

As MT has a promising and precarious role to play in modern linguistics, it is ultimately the key issue. It is a facilitator of communication, thereby opening spaces to international intercultural exchange and multiple languages. However, without rigorous refinement, this could lose the richness of language and maintain cultural hierarchies. The real challenge is not to make translation systems more accurate, but to make them as delicate and context-aware as they are fast.

While recent large language model developments have introduced enhanced contextual processing capabilities, our findings indicate that core challenges in English-Arabic translation persist across different neural frameworks. The morphological complexity and cultural embedding issues documented in this study continue to affect both dedicated MT systems and general-purpose LLMs, suggesting that these represent fundamental linguistic barriers rather than limitations of specific technological approaches. However, the interactive nature of LLMs offers potential pathways for addressing some pragmatic limitations identified in traditional translation platforms.

Recent developments in large language models (GPT-4, Claude, Bard) demonstrate both convergent and divergent patterns with our MT findings. These models exhibit similar difficulties with Arabic morphological complexity and cultural nuance preservation, suggesting persistent limitations across neural architectures. However, LLMs show superior contextual awareness in conversational settings, potentially addressing some pragmatic issues we identified. The interactive nature of LLMs allows clarification opportunities unavailable in traditional MT, though this comes with trade-offs in specialized accuracy. Our findings regarding human oversight necessity align with current LLM practices, where expert review remains essential for critical translation tasks despite impressive general capabilities.

## 7. Recommendations

Based on the empirical evidence and qualitative insights from this study, the performance, reliability, and cultural sensitivity of English-Arabic MT systems can be improved through several targeted recommendations.

1. **Develop Hybrid Human-AI Translation Models**

   High-stakes translation scenarios must include the incorporation of human-in-the-loop frameworks. Post-edited or edited by linguistically trained native speakers on a case-by-case basis, the study shows that semantic accuracy and cultural appropriateness achieve much greater gains than when not post-edited or edited. Collaborative workflows must integrate AI scalability with human judgments, which should be institutionalized within institutions and among MT developers.

2. **Expand and Diversify Training Corpora**

   MT models must be trained on culturally rich, diverse, and genre-spanning Arabic-English corpora. The over-reliance on Eurocentric datasets leads to systemic bias and misrepresentation. To make the translations more contextually grounded, training data should contain regional dialects, religious expressions, and colloquial registers.

3. **Prioritize Context-Aware and Pragmatically Informed Architectures**

   Pragmatic features and discourse-level context need to be included in NMT systems, most especially for Arabic languages, where politeness markers, indirect speech, and idiomatic nuances are central. Future architectures should be built on transformer models that account for long-range dependencies and dialogue acts.

4. **Institutionalize Ethical Auditing of MT Systems**

   Regular audits for bias and misrepresentation, especially concerning gendered language, religious references, and political terminology, should be integrated into the development cycle. Stakeholders should adopt ethical guidelines to identify and mitigate culturally insensitive outputs.

5. **Support Cross-Disciplinary Collaboration in MT Development**

   Effective MT requires collaboration between computer scientists, linguists, and cultural experts. Universities, tech companies, and governmental agencies should encourage interdisciplinary projects to design and refine contextually aware MT tools.

6. **Promote Public Education and Professional Training**

   Translators, educators, and content creators should be trained in the capabilities and limitations of MT. It is important to be aware of when and how to intervene manually in automated translation in order to maintain accuracy in real-world settings.

## 8. Conclusion

This comparative investigation of Google Translate, DeepL, and MarianNMT in English-Arabic translation contexts demonstrates that while these platforms offer substantial utility. This study aims to investigate the dual role of neural machine translation (NMT) in the modern linguistics of a tough English-Arabic language pair. In doing so, it also established that while MT systems provide large-scale utility for speedily accelerating multilingual communication and linguistic research, MT systems lack the fidelity for syntactic expression, idiomatic correctness, and cultural nuance that is seen in natural language.

Results show that the current NMT platforms perform comparatively well with technical and literal texts and less well with literary, religious, and idiomatic content domains where cultural and pragmatic sensitivity matters most. As with any translation, a few systematic translation errors, such as gender mismatches, misinterpretation of metaphors, and ideological bias, are all elements that reinforce the hypothesis that MT generally flattens or corrupts deeply embedded linguistic meaning.

However, despite these limitations, the result points to some promising ways of further improving the model. One of the main points of hybridization of the MT workflows is creating the automated outputs that are modified with expert human intervention. Further, the improvement of more inclusive training datasets that are contextually aware can account for many of the identified ethical and linguistic gaps.

Finally, MT should not only be regarded as a technology but as a linguistic device that must be placed in social, cultural, and ethical contexts. If developed responsibly and collaboratively, machine translation can evolve into a more inclusive, intelligent, and culturally attuned mediator of global dialogue.

### Acknowledgments

The author extends his appreciation to the Deanship of Scientific Research at King Faisal University for funding this research work through the project number KFU251824

### Authors' contributions

The author conceived the study, conducted the research, analyzed the data, and wrote the manuscript. The author read and approved the final version of the manuscript.

### Funding

This research was funded by Deanship of Scientific Research at King Faisal University for funding this research work through the project number KFU251824

### Competing interests

The authors declare that they do not have any conflict of interest.

### Informed consent

Obtained.

### Ethics approval

The Publication Ethics Committee of the Sciedu Press.

The journal's policies adhere to the Core Practices established by the Committee on Publication Ethics (COPE).

### Provenance and peer review

Not commissioned; externally double-blind peer reviewed.

### Data availability statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

### Data sharing statement

No additional data are available.

### Open access

### Copyrights

### References

Abubakari, M. S. (2025). *Biases in generative AI models in Arabic language*. IGI Global. 2025. https://doi.org/10.4018/979-8-3373-2530-9.ch013

Ahmed, M. (2023). *Breaking the barriers in translation: Insights in Arabic-English services.* ResearchGate. Retrieved from https://www.researchgate.net/publication/372237478

Ahmed, M., & Salama, H. (2023). Neural network-based approaches for improving Arabic MT quality: A survey. *Machine Translation. 37*(2), 153-180. https://doi.org/10.1007/s10590-022-09368-6

Ahmed, S. A. (2022). *AI bias and neural machine translation: Translating heavily loaded ideological messages.* CDELT Occasional Papers. https://doi.org/10.21608/opde.2022.282208

Alharbi, A., & Alshammari, R. (2023). *Analysis of error propagation in English-to-Arabic NMT systems.* Proceedings of the ACL 2023 Workshop on Machine Translation Evaluation. 34-43.

Alhebshi, S. H. S., Alharazi, A. F. A., Taleb, N. R. M. (2024). Translating Arabic Islamic terminology using AI-powered MT tools. *Pakistan Journal of Language Studies and Society (PJLSS). Retrieved* from https://www.pjlss.edu.pk/pdf_files/2024_2/7150-7164.pdf

Alkhatib, M. (2019). *Neural machine translation for Arabic language.* ProQuest. Retrieved from https://bspace.buid.ac.ae/bitstream/handle/1234/1674/2015246033.pdf

Almaaytah, S. A., Aalzobidy, S. A., & Alwidyan, M. F. (2024). *Using Machine Translation: English-Arabic Procedures and Challenges—A Systematic Review*. ResearchGate. Retrieved from https://www.researchgate.net/publication/389491280

Alotaibi, M. S., & Alzahrani, A. I. (2023). Recent advances in neural machine translation for English-Arabic language pair. *Journal of King Saud University - Computer and Information Sciences, 35*(1), 1-14. https://doi.org/10.1016/j.jksuci.2022.05.004

Al-Salami, S., & Farah, A. B. (2024). Challenges in English to Arabic machine translation: A linguistic perspective. *International Journal of Applied Linguistics, 34*(2), 123-139. https://doi.org/10.1111/ijal.12398

Awad, M., & El-Bakry, H. (2024). Handling gender and politeness in Arabic neural machine translation. *Transactions on Asian and Low-Resource Language Information Processing, 23*(1), 1-17. https://doi.org/10.1145/3639930

Diab, N. M. (2022). *Quality assessment of statistical and neural approaches to machine translation.* ResearchGate. Retrieved from https://www.researchgate.net/publication/386422675

Elgendy, M. (2023). Cross-cultural communication barriers in English-Arabic machine translation: Issues and solutions. *Intercultural Pragmatics, 20*(1), 89-107. https://doi.org/10.1515/ip-2022-0014

Farouk, O., & Tamer, M. (2023). Challenges of word segmentation in Arabic MT and proposed neural solutions. *IEEE Transactions on Neural Networks and Learning Systems, 34*(7), 2882-2891. https://doi.org/10.1109/TNNLS.2023.3167154

Hasan, R., & Saleh, M. (2024). Exploiting transformer models for Arabic-English machine translation: Opportunities and practical challenges. *Information Processing & Management, 61*(3), 103-117. https://doi.org/10.1016/j.ipm.2023.102917

Hassanein, R., & Khaled, M. (2024). Transfer learning for Arabic dialect translation in neural machine translation. *Journal of Computational Linguistics, 50*(2), 293-312. https://doi.org/10.1162/coli_a_00403

Jasim, D., & Mahmood, A. (2023). Investigation of linguistic phenomena in Arabic MT: Syntactic and semantic challenges. *Natural Language Engineering, 29*(4), 539-556. https://doi.org/10.1017/S1351324922000560

Karim, F., & Al-Khalifa, H. (2024). Evaluating the impact of corpus size on neural MT for English-Arabic translation. *Language Resources and Evaluation, 58*(2), 335-349. https://doi.org/10.1007/s10579-023-09645-2

Latif, A., & Younis, M. (2023). Hybrid models for improved machine translation quality: Case study in English-Arabic. *Expert Systems with Applications, 214,* 119629. https://doi.org/10.1016/j.eswa.2022.119629

Majid, T., & Al-Jaber, S. (2025). Modern linguistics and MT: Bridging the gap in English-Arabic translation. *Linguistics and the Human Sciences, 17*(1), 45-63. https://doi.org/10.1558/lhs.2024.17.1.45

Nasser, M., & Farid, A. (2023). Neural architectures for handling morphology in Arabic MT. *ACM Transactions on Asian and Low-Resource Language Information Processing, 22*(3), 15. https://doi.org/10.1145/3566811

Omari, H., & Saeed, K. (2023). Deep learning-based machine translation systems: A review of English-to-Arabic implementations. *Artificial Intelligence Review, 56*(4), 3247-3279. https://doi.org/10.1007/s10462-022-10242-7

Qasim, W., & Abdullah, S. (2024). Cross-cultural pragmatics and machine translation: English-Arabic idiomatic expressions. *Journal of Pragmatics, 210,* 28-39.

Qassem, M., & Aldaheri, M. M. (2023). Can Machine Translate Dialogue Acts? Evidence from English-Arabic Dialogues. *Asian Journal of English Language.* https://doi.org/10.17576/3L-2023-2904-05

Rashed, R., & Youssef, W. (2023). Evaluating Transformer-based MT models on religious text translation (English-Arabic). *Natural Language Processing and Chinese Computing, 12*(1), 112-124. https://doi.org/10.1007/s11704-023-8071-x

Saeed, K., & Aziz, W. A. (2023). comparative study of statistical vs neural MT for English-Arabic language pairs. *The Journal of Machine Translation, 37*(1), 67-90.

Safa'a, A. A. (2023). *AI bias in neural MT: Translating ideological English-Arabic content.* ResearchGate. 2023. Retrieved from https://www.researchgate.net/publication/368894698

Sakr, N., & Malak, R. (2024). Machine translation and linguistic typology: Handling Arabic language structure in NMT. *Computational Linguistics, 50*(1), 59-81. https://doi.org/10.1162/coli_a_00401

Shams, A., & Fawzy, R. (2023). Low-resource challenges in English-Arabic machine translation and transfer learning solutions. *Journal of Language Modelling, 11*(2), 301-325. https://doi.org/10.15398/jlm.v11i2.302

Taha, M., & El-Amrani, I. (2024). Semantic role labeling for Arabic MT: Advances and open challenges. *Language Resources and Evaluation, 58*(4), 705-724. https://doi.org/10.1007/s10579-024-09620-z

Uthman, M., & Ghazali, S. (2023). The role of machine translation in enhancing cross-cultural dialogue between English and Arabic speakers. *International Journal of Cross-Cultural Management, 23*(3), 440-460. https://doi.org/10.1177/14705958231101299

Youssef, L., & Abdulla, I. (2024). Advances in attention mechanisms for Arabic-English neural machine translation. *IEEE Access, 12*, 93421-93433. https://doi.org/10.1109/ACCESS.2024.3235628

Zakaria, Y., & Hamed, M. (2023). Integrating linguistic features into deep learning MT systems for Arabic. *Computational Intelligence, 39*(2), 423-440. https://doi.org/10.1111/coin.12516

Zaki, H., & Nabil, A. (2025). Context-aware machine translation for English-Arabic: Addressing ambiguity and polysemy. *Expert Systems, 42*(1), e13036. https://doi.org/10.1111/exsy.13036

Zakraoui, J., Saleh, M., Al-Maadeed, S., Alja'am, J. M. (2021). *Arabic machine translation: A survey with challenges and future directions.* IEEE Access. 2021. https://doi.org/10.1109/ACCESS.2021.3132488