Impact Factor (SJIF): 3.484 E-ISSN: 2583-7206

International Journal of All Subject Research

Received: 17/April/2023 **IJASR:** 2023; 2(3):15-19 **Accepted:** 02/June/2023

A Linguistic and Comparative Evaluation of English-to-Urdu Machine Translation Paradigms: A Corpus-Based Study

[*1]Fatima Qurratulain Zufa, [2]Mohammad Khalid Mubashiruz Zafar and [3]Dr. Syed Majid Ali

[*1]Research Scholar, Department of Translation Studies, Maulana Azad National Urdu University, Hyderabad, India.

[2]Professor, Department of Translation Studies, Maulana Azad National Urdu University, Hyderabad, India.

[3]Former Research Associate, Directorate of Translation, Translation Studies, Lexicography, and Publications (DTTLP), Maulana Azad National Urdu University, Hyderabad, India.

Abstract

Machine Translation (MT) has emerged as one of the most transformative developments in computational linguistics and translation studies. Although substantial progress has been achieved in high-resource language pairs, the linguistic evaluation of MT systems for morphologically rich and structurally divergent languages such as English and Urdu remains underexplored. This study presents a comprehensive linguistic assessment of four major machine translation paradigms—Rule-Based Machine Translation (RBMT), Statistical Machine Translation (SMT), Neural Machine Translation (NMT), and Hybrid Machine Translation (HMT)—within the context of English-to-Urdu translation.

The research is based on a structured corpus of five thousand English sentences encompassing diverse syntactic, semantic, morphological, and stylistic patterns. Each sentence was translated using the four MT approaches and systematically evaluated across morphological, syntactic, semantic, and stylistic parameters. The analysis identifies recurring patterns of error, performance trends, and areas of strength, offering a detailed multi-layered linguistic perspective on the outputs generated by each system.

The findings indicate that Neural Machine Translation demonstrates superior fluency and contextual sensitivity, while Rule-Based systems perform relatively better in preserving grammatical structure in controlled contexts. Statistical systems exhibit moderate performance but encounter difficulties with long-distance dependencies and morphological agreement, whereas Hybrid systems display a promising balance that requires further linguistic refinement. Overall, the study concludes that for a morphologically complex and culturally nuanced language like Urdu, machine translation functions best as an advanced assistive tool that necessitates human linguistic oversight to ensure semantic accuracy and cultural appropriateness.

Keywords: Machine Translation, Urdu Linguistics, Morphological Complexity, Semantic Fidelity, Corpus-Based Analysis, English–Urdu Translation, Linguistic Evaluation, Rule-Based MT, Statistical MT, Neural MT, Hybrid MT, Syntax, Translation Studies.

Introduction

Translation has historically functioned as a bridge between civilizations, enabling the transmission of knowledge, culture, and intellectual traditions across linguistic boundaries. From the translation of Greek philosophical texts into Arabic during the Abbasid era to the subsequent transmission of Arabic scientific works into Latin, translation has served as a catalyst for global intellectual development.

In the modern digital era, the unprecedented expansion of information has rendered manual translation insufficient to meet global communicative demands. This context facilitated the emergence of Machine Translation (MT) as an automated solution to multilingual communication. MT seeks to computationally model linguistic structures and generate translations without direct human intervention.

English and Urdu represent two typologically distinct languages. English, belonging to the Germanic family, is predominantly analytic with rigid Subject–Verb–Object (SVO) structure. Urdu, an Indo-Aryan language, is morphologically rich and primarily follows a Subject–Object–Verb (SOV) order. These structural differences pose significant challenges for automated translation systems.

This study investigates the linguistic performance of four major MT approaches in English-to-Urdu translation, focusing on their ability to preserve grammatical accuracy, semantic equivalence, and stylistic appropriateness.

Literature Review

The theoretical foundations of translation studies were significantly shaped by scholars such as Eugene Nida (1964) [11], who introduced the concept of dynamic equivalence, emphasizing receptor-oriented translation. Peter Newmark (1988) [10] distinguished between semantic and communicative translation, while Mona Baker (1992) highlighted the role of

***Corresponding Author:** Fatima Qurratulain Zufa

textual and pragmatic factors in translation.

Machine translation research began in the mid-20th century following Warren Weaver's 1949 memorandum proposing automated translation. Early systems relied on rule-based models. However, the ALPAC report (1966) criticized MT's limited success, temporarily slowing research progress.

The 1990s witnessed the rise of Statistical Machine Translation (Brown *et al.*, 1993) [5], which shifted from rule-driven to corpus-driven models. Neural Machine Translation emerged in the 2010s, revolutionizing MT through deep learning architectures such as encoder-decoder models and attention mechanisms (Bahdanau *et al.*, 2015; Vaswani *et al.*, 2017) [3, 13].

Despite technological progress, linguistic evaluations of English–Urdu MT remain limited. Most studies focus on computational metrics (BLEU scores, etc.), while qualitative linguistic analysis—especially concerning morphology and stylistics—remains underexplored. This study addresses that research gap.

A Brief History and Comparative Analysis of English and Urdu Languages

Before proceeding further, we will briefly examine the historical background of both English and Urdu in order to understand their origins, roots, language families, as well as their similarities and differences.

English and Urdu are two major world languages with rich historical backgrounds. English is a global language of communication, science, technology, and education, while Urdu is a significant cultural, literary, and religious language of South Asia. Although both languages belong to the Indo-European language family, they developed in different geographical, cultural, and historical contexts.

A brief Introduction of English

English originated in England and developed mainly from Germanic languages. Its history can be divided into three major stages:

a) **Old English (450–1100 AD):** Old English developed after the arrival of Angles, Saxons, and Jutes from Germany and Denmark. It was heavily influenced by Germanic vocabulary and grammar.

b) **Middle English (1100–1500 AD):** After the Norman Conquest (1066), French and Latin strongly influenced English. Thousands of French words entered the language, especially in law, government, culture, and education.

c) **Modern English (1500–present):** With the Renaissance, printing press, colonial expansion, and industrial revolution, English absorbed vocabulary from Latin, Greek, and many world languages. Today, English is the most widely used international language.

A Brief Introduction of Urdu

Urdu originated in the Indian subcontinent, especially in the military camps and urban centers of Delhi and surrounding regions during the Delhi Sultanate and Mughal period (12th–18th century).

The word *Urdu* comes from the Turkish word "Ordu", meaning *army or camp*, indicating its origin as a contact language among soldiers and civilians.

Urdu developed through contact between local Indo-Aryan languages, especially Khari Boli, and Persian, Arabic, and Turkish languages.

Over time, Urdu became a refined literary language with rich poetry, prose, and religious literature. Today, it is the national language of Pakistan and one of the official languages of India.

Language Family and Classification

Both English and Urdu belong to the Indo-European language family, but they belong to different branches:

Feature	English	Urdu
Language Family	Indo-European	Indo-European
Branch	Germanic	Indo-Aryan
Writing Script	Roman (Latin)	Perso-Arabic
Major Influences	Latin, French, Greek	Persian, Arabic, Turkish

Similarities between English and Urdu

Despite their different origins, English and Urdu share several similarities:

a) **Common Indo-European Roots:** Both languages ultimately trace back to Proto-Indo-European, which explains some basic conceptual similarities in grammar and sentence structure.

b) **Rich Vocabulary from Other Languages**

Both languages borrowed extensively:

- **English:** Latin, Greek, French
- **Urdu:** Persian, Arabic, Turkish

c) **Flexible Word Formation:** Both languages easily adopt new words, which helps them grow continuously.

d) **Literary Traditions:** Both languages possess strong literary traditions, including poetry, fiction, drama, and philosophical writing.

Major Differences between English and Urdu

a) **Script**

- English uses Roman script, written left to right.
- Urdu uses Perso-Arabic script, written right to left.

b) **Grammar Structure**

- English follows Subject–Verb–Object (SVO) order.
  - **Example:** *I eat food.*
- Urdu follows Subject–Object–Verb (SOV) order.
  - **Example:** میں کھانا کھاتا ہوں(Main khana khata hoon).

c) **Use of Gender**

- English has natural gender (he, she).
- Urdu has grammatical gender, where nouns and verbs change according to masculine and feminine forms.
  - Example:
    - **Masculine:** وہ کتاب پڑھ رہا ہے۔ *Woh kitaab parh raha hai.*(He is reading a book.)
    - **Feminine:** وہ کتاب پڑھ رہی ہے۔ *Woh kitaab parh rahi* ہے۔(She is reading a book.)

d) **Verb System**

- English relies heavily on auxiliary verbs and tense forms.
- Urdu uses postpositions, participles, and auxiliary verbs in more complex ways.

e) **Cultural Influence**

- English reflects Western culture and philosophy.
- Urdu reflects South Asian, Islamic, and Persian cultural

traditions.

So the conclusion is English and Urdu are two historically rich languages with deep roots in the Indo-European family. While English developed mainly through Germanic and European influences, Urdu emerged through Indo-Aryan, Persian, Arabic, and Turkish interactions. Their similarities reflect their ancient common origin, while their differences highlight distinct cultural, historical, and social developments. A comparative understanding of these languages helps in translation, linguistic research, and cross-cultural communication.

The Evolution of Machine Translation and the Urdu Context

Translation has historically served as the cornerstone of intellectual continuity and cultural dialogue between civilizations. From the transmission of Greek philosophy into Arabic to the dissemination of modern scientific knowledge, translation is more than a linguistic substitution; it is a vital instrument of cross-cultural communication. In the contemporary era, the sheer volume of global information has necessitated the automation of this process through Machine Translation (MT).

The primary objective of MT is to facilitate fast, scalable, and cost-effective communication across languages. However, the transition from human-centric translation to automated systems introduces profound linguistic challenges, particularly for language pairs that belong to different genealogical families. English, an analytical language, and Urdu, a synthetic and morphologically dense language, present a unique set of structural asymmetries. This research addresses a critical gap in existing literature, where Urdu has often been marginalized or evaluated through purely technical metrics rather than qualitative linguistic standards.

Theoretical Foundations: Paradigms of Automated Translation

The development of machine translation has been marked by several paradigm shifts, each reflecting the dominant linguistic and computational theories of its period.

The Rule-Based Paradigm (RBMT): Rule-Based Machine Translation represents the earliest systematic effort in the field, treating language as a structured system of predefined rules. These systems rely on comprehensive bilingual dictionaries, morphological analyzers, and syntactic parsers. While RBMT provides a high degree of grammatical consistency, it is notoriously rigid. It struggles with the creative and idiomatic nature of natural language, often resulting in literal translations that fail to capture the spirit of the source text.

The Statistical Shift (SMT): Statistical Machine Translation emerged as a data-driven alternative, utilizing large parallel corpora to establish probabilistic correspondences between words and phrases. By shifting away from handcrafted rules, SMT offered better adaptability to high-resource languages. However, for low-resource languages like Urdu, SMT often produces fragmented output due to "data sparsity"—the lack of sufficient bilingual data to train reliable models.

The Neural Revolution (NMT): The introduction of Neural Machine Translation in the early twenty-first century marked a revolutionary shift. By employing artificial neural networks and "transformer" architectures, NMT processes language as an integrated, context-sensitive system. Unlike earlier models that translated word-by-word or phrase-by-phrase, NMT captures long-range dependencies and produces significantly more fluent output.

The Hybrid Approach (HMT): Hybrid systems attempt a pragmatic synthesis, combining the linguistic precision of rule-based models with the data-driven fluency of neural or statistical models. This approach is particularly relevant for languages like Urdu, where explicit morphological rules can help guide a neural model toward more grammatically correct inflections.

Research Methodology: The Five Thousand Sentence Corpus

The strength of this research lies in its empirical foundation, a systematically designed corpus of five thousand (5,000) English sentences. This corpus was not a random collection but a carefully curated dataset representing diverse thematic domains, including academic discourse, everyday conversation, and texts with religious or cultural references.

The research methodology followed a descriptive-analytical approach. Each sentence was processed through the selected MT systems, and the outputs were evaluated across four linguistic parameters: morphology, syntax, semantics, and style. Human evaluation remained the "gold standard," as technical scores often overlook the "hallucinations" or semantic drifts that are common in automated systems.

Corpus

A corpus of 5,000 English sentences was constructed. The dataset includes:

- Simple sentences
- Complex and compound sentences
- Idiomatic expressions
- Technical and academic sentences
- Conversational structures

MT Approaches Evaluated

- Rule-Based Machine Translation (RBMT)
- Statistical Machine Translation (SMT)
- Neural Machine Translation (NMT)
- Hybrid Machine Translation (HMT)

Evaluation Criteria

Each translation was evaluated for:

- Gender and number agreement
- Verb inflection accuracy
- Word order conformity
- Idiomatic translation
- Cultural and stylistic appropriateness

1. Morphological Analysis

Urdu possesses a rich inflectional system where verbs, adjectives, and pronouns must agree with nouns in gender, number, and case. The evaluation revealed that RBMT systems, while structurally consistent, often failed to capture the nuances of Urdu's politeness markers and gender agreements. Conversely, NMT systems demonstrated a higher level of morphological alignment through contextual learning but were prone to "over-generalization," where they might apply a common masculine inflection to a feminine noun incorrectly.

Urdu's morphological richness presents challenges such as:

- Gender agreement (وہ گیا / وہ گئی)
- Pluralization patterns
- Honorific forms

Findings:

- RBMT maintained grammatical gender in structured sentences.
- SMT frequently produced agreement mismatches.
- NMT improved morphological consistency but still showed occasional verb-form errors.
- Hybrid systems performed moderately well but lacked consistency in complex sentences.

2. Syntactic Analysis

One of the most significant challenges in English-to-Urdu MT is the radical difference in word order. English follows a Subject-Verb-Object (SVO) structure, whereas Urdu is a Subject-Object-Verb (SOV) language.
English SVO → Urdu SOV transformation is crucial.

Example:
English: She reads a book.
Correct Urdu: وہ کتاب پڑھتی ہے۔

RBMT: Often rigid and literal. RBMT systems often struggle with "verb-final" positioning, producing sentences that sound unnatural or fragmented to a native speaker.
SMT: Fragmented phrase ordering.
NMT: Most natural SOV reordering. NMT models showed a superior ability to reposition verbs at the end of the sentence and handle complex subordinate clauses.
Hybrid: Balanced but occasionally overcorrected.

3. Semantic Analysis

Semantic fidelity—the accurate transfer of the original message—is the most sensitive criterion for evaluating translation quality. The research findings indicate that while NMT produces the most fluent sentences, it is not always the most accurate.
The Problem of Literalism in RBMT: In cases of idioms or metaphorical language, RBMT systems tend to produce word-for-word translations. For instance, common English expressions like "break the ice" are rendered literally, losing their intended meaning entirely in Urdu.
Semantic Drift in NMT: A critical finding of this study is the phenomenon of "semantic drift" in Neural systems. Because NMT prioritizes fluency, it occasionally produces a sentence that sounds like perfect Urdu but has fundamentally altered the meaning of the source text. This underscores the danger of "blind reliance" on MT for legal or religious documents where every word carries immense weight.

Idioms Posed Serious Challenges
English: "Break the ice."
RBMT: برف توڑنا (literal error)
SMT: Awkward phrasing
NMT: Improved contextual rendering
Hybrid: Partial success

NMT demonstrated superior semantic handling due to contextual modeling.

4. Stylistic, Cultural and Pragmatic Evaluation

Urdu requires sensitivity to formality levels, honorific markers, and cultural expressions, as it is a language deeply embedded in social hierarchy and cultural norms. The appropriate use of honorifics such as *Hazrat, Janāb,* and *Sahib* is essential for achieving communicative competence and ensuring respectful and culturally appropriate interaction.

- **Registers and Formality:** The study found that most MT systems fail to distinguish between formal and informal registers. A sentence intended for a superior might be translated using an informal verb form, which would be considered culturally offensive.
- **Contextual Nuance:** NMT systems performed better in modern, professional contexts (e.g., digital culture) but struggled significantly with the deep spiritual or social nuances of classical Urdu.

Major Findings and Implications

The comparative evaluation leads to several profound conclusions:

- **Fluency vs. Accuracy:** The study challenges the assumption that a "smooth" translation is a "correct" one. NMT often disguises semantic errors with grammatical fluency.
- **Language-Specific Models:** There is a dire need for MT models designed specifically for the morphological and cultural traits of Urdu, rather than force-fitting it into universal frameworks.
- **Human in the Loop:** For sensitive domains such as academia, law, and religion, human post-editing is not just an option but a requirement to ensure intellectual responsibility.

Comparative Findings

Parameter	RBMT	SMT	NMT	HMT
Morphology	Moderate	Weak	Strong	Moderate
Syntax	Strong (rigid)	Weak	Strong	Strong
Semantics	Weak	Moderate	Strong	Moderate
Stylistics	Weak	Weak	Moderate	Moderate

NMT outperformed other approaches overall but requires integration with linguistic rule-based mechanisms for Urdu.

Conclusion

This research demonstrates that machine translation is fundamentally a linguistic and cultural challenge rather than a purely technical one. Although the technological evolution from Rule-Based to Neural approaches marks significant progress, the intrinsic complexities of the Urdu language—its morphological richness, SOV syntactic structure, flexible word order, and deep socio-cultural embedding—continue to pose serious challenges for even the most advanced AI systems. The study confirms that no existing MT system can fully replicate human translation, as effective English–Urdu translation requires sensitivity to cultural nuance, pragmatic variation, and contextual appropriateness.
The findings reveal that Neural Machine Translation currently offers the highest overall performance, particularly in fluency and contextual handling, but remains heavily dependent on the quality and size of training data. For relatively low-resource languages such as Urdu, data scarcity significantly limits system performance. Rule-Based systems, while structurally reliable, lack natural fluency, whereas Statistical models struggle with morphological agreement and syntactic complexity. Hybrid approaches demonstrate promising potential by balancing linguistic rules and data-driven learning, though they require further refinement and linguistic optimization.
The study recommends a shift toward linguistically informed, morphology-aware neural-hybrid frameworks that integrate

human linguistic expertise into the automated translation pipeline. Such interdisciplinary collaboration among linguists, computational scientists, and translation scholars is essential to bridge the gap between machine efficiency and human precision. While machine translation cannot replace human translators, it remains an indispensable tool in the digital era, and its advancement for morphologically rich and culturally nuanced languages like Urdu depends on the deep integration of linguistic modeling and artificial intelligence.

References

1. Baker M. *In Other Words: A Coursebook on Translation*. London: Routledge; 1992.
2. Baker M. *Corpora in Translation Studies: An Overview*. Target. 1995;7(2):223–243.
3. Bahdanau D, Cho K, Bengio Y. *Neural Machine Translation by Jointly Learning to Align and Translate*. In: Proceedings of the International Conference on Learning Representations (ICLR); 2015.
4. Biber D, Conrad S, Reppen R. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press; 1998.
5. Brown PF, Della Pietra SA, Della Pietra VJ, Mercer RL. *The Mathematics of Statistical Machine Translation*. Computational Linguistics. 1993;19(2):263–311.
6. Catford JC. *A Linguistic Theory of Translation*. London: Oxford University Press; 1965.
7. Hatim B, Mason I. *The Translator as Communicator*. London: Routledge; 1997.
8. Koehn P, Knowles R. *Six Challenges for Neural Machine Translation*. In: Proceedings of the First Workshop on Neural Machine Translation, ACL; 2017. p. 28–39.
9. McEnery T, Hardie A. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press; 2012.
10. Newmark P. *A Textbook of Translation*. London: Prentice Hall; 1988.
11. Nida EA. *Toward a Science of Translating*. Leiden: Brill; 1964.
12. Popović M. *Error Classification and Analysis for Machine Translation*. Natural Language Engineering. 2015;21(4):1–26.
13. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. *Attention Is All You Need*. In: Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS); 2017. p. 5998–6008.
14. Weaver W. *Translation*. Memorandum, Rockefeller Foundation; 1949.
15. ALPAC. *Language and Machines: Computers in Translation and Linguistics*. Washington, DC: National Academy of Sciences; 1966.