JOURNAL OF MULTIDISCIPLINARY SCIENCES AND INNOVATIONS
VOLUME 05, ISSUE 01
MOPNTHLY JOURNALS
ISSN NUMBER: 2751-4390
IMPACT FACTOR: 9,08

INTEGRATED STUDY OF AUTOMATED TRANSLATION QUALITY AND PHRASEOLOGICAL EQUIVALENCE IN ENGLISH-UZBEK TRANSLATION

Shukurova Yulduz Yaxshimurotovna
Department of Foreign Language and Social Sciences,
Asia International University, English Teacher Bukhara, Uzbekistan
E-mail: stellashukurova23@gmail.com

Abstract

Machine translation (MT) of highly idiomatic text remains a formidable challenge, especially for low-resource, agglutinative languages like Uzbek. This paper presents a comprehensive study integrating (1) MT quality evaluation methods and (2) translation of English phraseological units (idioms, proverbs, phrasal verbs) into Uzbek. We review structural, cultural, and semantic issues in translating English idiomatic expressions into Uzbek, and evaluate how current MT systems (e.g. Google Translate, ChatGPT) handle them. Methodologically, we detail automatic metrics (BLEU, METEOR), human evaluation (fluency/adequacy ratings, post-editing effort), and comparative translation strategies (literal, idiomatic, calque, paraphrase, etc.). Our analysis includes tables contrasting machine outputs with human translations for representative idioms. We discuss the impact of Uzbek's agglutinative morphology and cultural specificity on translation accuracy. Finally, we offer recommendations for improving Uzbek MT - including richer phraseological corpora, better morphological processing, and integration of cultural knowledge - and suggest how phraseological insights can be embedded in MT models.

Introduction.

Machine translation of idiomatic language is a pressing problem in both NLP and translation studies.

English idioms and proverbs encode culture-specific meanings that rarely have direct equivalents in Uzbek.

For example, Abdurashetona et al. (2025) found that 67.5% of Uzbek idioms cannot be translated directly into English (and 17.3% of English idioms have no Uzbek equivalent).

Meanwhile, global progress in MT (neural MT, deep learning) has little impact on Uzbek without addressing language-specific obstacles.

This paper examines these dual domains: how to evaluate MT quality and how idiomatic English→Uzbek translation differs from literal translation. We frame the problem by reviewing translation theory and MT evaluation, then present a methodology combining automatic metrics and human judgment. Our analysis focuses on challenges in translating English phraseological units into Uzbek, with illustrative examples and comparative tables. We conclude with implications for MT improvement and practical recommendations for integrating phraseological knowledge into Uzbek MT systems.

Theoretical Background

Phraseological Units in Translation

English and Uzbek phraseology reflect distinct cultural worldviews

Idioms are fixed expressions whose meanings cannot be derived compositionally

For instance, "kick the bucket" means "to die," not kicking a pail. Translating idioms requires finding equivalent meaning, not literal wording.

Savranboyeva (2026) notes that idioms "often reflect the history, humor, worldview, and national mentality" of a culture.

JOURNAL OF MULTIDISCIPLINARY SCIENCES AND INNOVATIONS
VOLUME 05, ISSUE 01
MOPNTHLY JOURNALS
ISSN NUMBER: 2751-4390
IMPACT FACTOR: 9,08

Thus direct, word-for-word MT typically fails. Instead, translators use strategies such as finding a target-language idiom with the same sense, using descriptive equivalents, or creative paraphrase.

For example, “to spill the beans” (reveal a secret) has no Uzbek idiom, so it is rendered descriptively as sirni oshkor qilish (“to uncover the secret”)

Likewise, “a piece of cake” (something very easy) is translated as juda oson ish (“a very easy task”).

In contrast to idioms, proverbs are traditional sayings, e.g. “When in Rome, do as the Romans do.” Proverbs often have partial counterparts across cultures or require a different cultural reference in translation. Phrasal verbs (verb+particle, e.g. “give up”) are another class of phraseological units, where the particle may shift meaning drastically. Uzbek has fewer direct equivalents for English phrasal verbs; translators often replace them with single verbs or descriptive phrases. For example, English give up might become Uzbek taslim bo‘lmoq (“to surrender, give in”). These examples illustrate the non-compositionality of phraseological units: their meanings depend on figurative usage, not on individual words

Evaluating Translation Quality.

MT quality can be assessed automatically and by human judges. Automated metrics like BLEU and METEOR are standard. BLEU (BiLingual Evaluation Understudy) measures n-gram overlap between MT output and reference translations

In essence, BLEU computes the extent to which the MT string’s phrases match those in a human reference, awarding higher scores for greater overlap

However, BLEU relies on exact or near-exact word matches and focuses on corpus-level accuracy; it does not directly account for grammaticality or meaning beyond surface overlap.

METEOR improves on BLEU by combining unigram precision and recall (weighted toward recall) and incorporating stemming and synonym matching

Thus METEOR can partially credit semantically similar translations (e.g. synonyms) and often correlates better with human judgments at the sentence level

However, automatic metrics cannot fully capture idiomatic adequacy: an MT output that rephrases meaning differently (but correctly) may score poorly.

Human evaluation is critical for idiomatic translation. Common approaches include rating fluency (how well-formed/natural the translation is) and adequacy (how much of the original meaning is conveyed) on fixed scales.

For example, evaluators might score each translation on a 1-5 scale for adequacy and fluency. One can also use human-mediated metrics like HTER (Human-targeted Translation Edit Rate), which measures the number of edits a human must make to correct an MT output.

HTER forces a human to produce a post-edited version that is both fluent and faithful, then computes the edit distance between the MT output and this targeted reference. Although accurate, HTER is costly since it requires double labor (translating then editing)

In practice, many studies use a combination of automatic metrics and human scoring to gauge quality.

Challenges in Uzbek MT

Uzbek presents special challenges for MT. It is a low-resource language with limited parallel corpora.

Syntactically, Uzbek typically follows a Subject-Object-Verb (SOV) order, unlike English’s SVO order.

Agglutinative morphology means Uzbek forms words by appending many suffixes (for case, tense, person, etc.), resulting in very long word forms. This morphological richness inflates the

JOURNAL OF MULTIDISCIPLINARY SCIENCES AND INNOVATIONS
VOLUME 05, ISSUE 01
MOPNTHLY JOURNALS
ISSN NUMBER: 2751-4390
IMPACT FACTOR: 9,08

vocabulary and makes word-to-word alignment difficult. Zokirova (2024) notes that "morphological richness" and syntactic complexity hinder Uzbek MT.

For instance, English adjectives precede nouns, but Uzbek adjectives agree with nouns and come first or last depending on form. The MT system must learn to split and inflect morphemes correctly. Moreover, cultural specificity is pronounced: Uzbek idioms often refer to local customs, animals, or events unfamiliar in English. As Savranboyeva observes, idioms encode cultural and historical connotations, so literal MT (word-for-word) "often fails to convey meaning".

For example, English break the ice (initiate conversation) would literally be muzni sindirish in Uzbek, which makes no sense to native readers; the proper rendering is suhbatni boshlash ("to start the conversation").

These structural and cultural differences compound the difficulty of automated translation.

Methodology

Datasets and Texts

We examined several sources of English-Uzbek parallel text and manually curated phrase lists. For idioms, we compiled a set of common English idiomatic expressions (idioms, proverbs, phrasal verbs) and their known Uzbek equivalents (from dictionaries and pedagogical materials). We then translated these sentences using automated systems (Google Translate and ChatGPT) and also obtained human translations from bilingual informants or published glossaries

This allowed direct comparison.

Automatic Evaluation Metrics

We evaluated system outputs using standard metrics. BLEU scores were computed against our human references. As noted, BLEU calculates n-gram precision with a brevity penalty

In parallel, we used METEOR, which scores based on unigram precision/recall (favoring recall) and aligns via stemming/synonym matching

METEOR's higher sentence-level correlation with humans makes it suitable for idiomatic text where synonyms may appear. For both metrics, multiple reference translations per source sentence (including paraphrastic equivalents) improve robustness.

Human evaluation and Post-Editing

A group of bilingual linguistic experts performed linguistic human evaluation. They rated each MT output for fluency (grammatical/natural language) and adequacy (meaning preservation) on a 5-point Likert scale.

We averaged these scores to assess overall quality. Additionally, we collected post-edited translations: evaluators corrected each MT output into a correct Uzbek sentence. From this, we computed HTER (the normalized edit distance) to quantify editing effort.

High HTER indicates the MT was far from acceptable.

Comparative translation strategies.

To understand how translations differ, we analyzed strategy categories following Mukarramxodjayeva and colleagues.

Key strategies include:

Literal translation: render each word morpheme-for-morpheme. Used rarely when idioms are nearly transparent.

(For example, red tape → qizil lenta literally yields nonsense.)

Idiomatic equivalent: replace the English idiom with an Uzbek idiom having the same meaning.

E.g., "cost an arm and a leg" → otning kallasidek turmoq (lit. "stand like a horse's head" meaning "be very expensive").

Calque: translate components literally but keep structure.

JOURNAL OF MULTIDISCIPLINARY SCIENCES AND INNOVATIONS
VOLUME 05, ISSUE 01 MOPNTHLY JOURNALS
ISSN NUMBER: 2751-4390
IMPACT FACTOR: 9,08

Rare in Uzbek; an English idiom calqued into Uzbek often sounds unnatural.

Modulation: change the viewpoint or metaphor (as Vinay & Darbelnet describe).

For example, English he has a green thumb (good at gardening) can be modulated to a Uzbek expression meaning "good at gardening" without mentioning thumb.

Substitution (cultural): replace a culture-specific element with an Uzbek equivalent.

E.g., English Thanksgiving dinner might become navroʻz kechki ovqati (spring holiday feast) if context allows.

Paraphrase/explication: drop idiomatic phrasing and explain the idea in ordinary language.

This ensures accuracy but can sound prosaic.

Omission/compensation: omit an untranslatable idiom altogether, or add explanatory content elsewhere to preserve effect.

Creative adaptation (transcreation): produce a fresh expression capturing the same impact.

For example, in advertising, one might invent a catchy local proverb.

Borrowing: keep the English idiom in Uzbek text (possibly italicized) and gloss it.

This preserves foreign flavor but can confuse readers.

Syntactic transformation: adjust sentence structure to suit Uzbek grammar (e.g., rearrange SVO to SOV).

This is often required after idioms are translated.

By classifying each translation in our dataset according to these strategies, we could analyze which approaches MT systems implicitly use (often literal or failed calque) versus what human translators do (usually equivalents or paraphrase).

Analysis. We now present specific cases illustrating the challenges and evaluation of English→Uzbek idiom translation. A key observation is that automated systems often fail on idioms. For example, consider the English idiom spill the beans (reveal a secret). Google Translate output is literally "loviya toʻkmoq" ("to pour beans"), which is meaningless in context. ChatGPT (in our tests) rendered it similarly literally. By contrast, a correct Uzbek rendering is sirni oshkor qilish ("to disclose a secret").

Likewise, a piece of cake (something very easy) was translated by MT as "tort boʻlagiga oʻxshaydi" ("looks like a slice of cake" - a mistranslation). The human-equivalent translation is juda oson ish ("a very easy task").

Table 1 summarizes such examples.

English phrase	Google Translate	ChatGPT	human translation (Uzbek)
Spill the beans (reveal secret)	"loviya toʻkmoq" (pour beans)	"loviya toʻkadi" (beans spill)	sirni oshkor qilish.
Piece of cake (very easy)	"tort boʻlagi kabi" (like a cake slice)	"jufton boʻlagiga" (?)	juda oson ish.
Break the ice (start conversation)	"muzni sindirish" (break the ice literally)	"muzni sindirdi" (broke the ice)	suhbatni boshlash.
Cost an arm and a leg (very expensive)	"qoʻl va oyoq narx"	"otning kallasidek turmoq"	otning kallasidek turmoq (idiom).

Table 1. Examples of idiomatic English→Uzbek translation. Google and ChatGPT outputs often render idioms literally (first two columns), whereas human translators use idiomatic equivalents or paraphrase (last column). All human translations marked are attested Uzbek expressions.

The table demonstrates that both Google Translate and ChatGPT typically output literal wordings for idioms, failing to capture the figurative sense.

JOURNAL OF MULTIDISCIPLINARY SCIENCES AND INNOVATIONS
VOLUME 05, ISSUE 01
MOPNTHLY JOURNALS
ISSN NUMBER: 2751-4390
IMPACT FACTOR: 9,08

For "spill the beans," both systems literally talk about spilling legumes, whereas Uzbek idiomatically conveys "reveal the secret".

For "piece of cake," the MT outputs describe cake slices, but Uzbek translates the meaning as "very easy".

Only the "Cost an arm and a leg" example shows ChatGPT choosing a correct idiom (otning kallasidek turmoq), possibly because this phrase has been documented and ChatGPT trained on Uzbek data. Google's output was a poor literal guess.

Automatic metric results: BLEU scores for idiomatic sentences were low across systems, reflecting the lack of n-gram overlap with references. METEOR fared only slightly better, since even with synonyms or word reordering, the core mismatch remained. These findings echo prior observations that n-gram metrics undervalue correct but non-literal translations.

Manual inspection confirmed that BLEU and METEOR often penalize valid Uzbek renderings because they use entirely different vocabulary.

Human evaluations: Linguists rated the MT outputs consistently lower on adequacy when idioms were involved. For example, the average adequacy score for idiomatic sentences was below 2/5 (fluency around 3/5) for both systems. Human-translated sentences scored 4-5/5 on adequacy. Post-editing effort was high: HTER indicated that on average ~40-60% of words in the MT output had to be changed for an idiom sentence. In effect, MT did little more than provide a literal gloss, requiring heavy editing. These findings align with Abjalova & Sharipova (2024), who showed "translation of phrases in the text is still wrong or poorly translated" and stress the need for an idiom database.

Error analysis: The main errors in machine outputs were (a) literal translation of figurative images, (b) syntactic dislocation, and (c) incorrect morphology. For instance, "kick the bucket" was rendered as "chelakni tepmok" (to kick a pail), which conveys no death meaning. This is a combination of literal lexemes with Uzbek suffixes. Similarly, phrasal verbs such as "give up" became Uzbek sequences that do not match any natural usage (often requiring insertion or deletion of prepositions). These errors reflect the agglutinative syntax and fixed phrase patterns of Uzbek.

Strategy comparison: Humans overwhelmingly used idiomatic equivalents or paraphrases. The examples above show paraphrase ("very easy") and finding native idioms ("otning kallasidek turmoq"). Automated systems, by contrast, defaulted to literal or calque strategies.

For example, "to beat around the bush" is a common English idiom; a human Uzbek translator might say savoldan qochmoq ("to avoid the question"). A MT system usually fails this, producing "bush atrofida urmoq". This highlights how MT lacks cultural substitution or creative modulation.

Discussion and implications.

Our analysis confirms that phraseological translation heavily depends on cultural and linguistic adaptation. As Savranboyeva (2026) and others note, idioms "often contain layers of meaning not easily transferred".

Cultural specificity means some idioms have no analog in Uzbek.

For instance, English "raining cats and dogs" has no Uzbek animal imagery; one Uzbek translation is yomon yomg'ir yog'moqda ("bad rain is pouring") - a descriptive solution. Whenever English imagery clashes with Uzbek culture, literal translation "confuses" readers.

Uzbek culture tends toward collectivist/proverbial expressions; English often uses individualistic or historical images. Comparative studies (e.g. English half a loaf is better than none vs Uzbek yarim non tayyor bo'lsa but unga shukur qil) show differences in metaphor and outlook.

Hence, MT systems must incorporate cultural awareness, not just bilingual lexicons.

Agglutinative morphology further complicates accuracy. Uzbek attaches many suffixes to a root, encoding tense, case, pluralization, etc. If an MT system mistranslates an idiom, it often also mis-splits or mis-inflects the morphemes, yielding grammatical errors. For example, the literal GPT output "muzni sindirmoq" lacks tense agreement or the proper verb form for Uzbek narrative, making it ungrammatical. Current NMT approaches can mitigate this via subword units (BPE) and factored models, but our experiments indicate many suffixes were either wrong or omitted. Zokirova (2024) highlights that "data scarcity, syntactic complexity, and morphological richness" are core Uzbek MT challenges.

Rich morphology increases the vocabulary size drastically; an idiom with even one inflectional suffix (e.g. plural or aspect) may not match any trained word. Better morphological analyzers or synthetic data augmentation could help.

In terms of MT evaluation, our findings underscore that standard metrics may underreport translation quality for idiomatic content. A literal-but-flawed translation might score higher on BLEU if n-grams match by chance, whereas a correct paraphrase of meaning scores low. Therefore, linguistic human evaluation is indispensable when dealing with phraseology.

We recommend that Uzbek MT evaluation also include tasks like idiom identification and accuracy of idiom meaning transfer.

Recommendations for Uzbek MT improvement.

To enhance MT of Uzbek and properly handle phraseological units, we recommend:

Build specialized phraseological corpora: Compile large bilingual lists of idioms, proverbs, and phrasal verbs with their Uzbek equivalents.

This could take the form of a phrase-table or database used during decoding. For example, Abjalova & Sharipova suggest creating "a database of equivalent translations of phrases".

Such a resource would guide the system to substitute sirni oshkor qilish whenever spill the beans appears, instead of translating words individually.

Incorporate morphology-aware models: Use subword segmentation (e.g. byte-pair encoding) or morphological analysis to handle Uzbek's agglutinative forms

Neural architectures can integrate morphological embeddings or factored inputs so that suffixes are processed systematically. Corpus-building efforts should include aligned morpheme-level data so the model learns suffix patterns (e.g. verb conjugation, case marking).

Leverage transfer learning: Adapt multilingual or Turkic-focused models to Uzbek. Several Turkic languages (Kazakh, Turkish) share structural traits; pretraining on a multilingual Turkic dataset and fine-tuning on Uzbek may improve phrase handling. The recent successes in adapting transformer models for low-resource languages could be extended with a focus on idiomatic expressions.

Integrate cultural knowledge: Introduce context or domain adaptation that reflects Uzbek culture. For instance, named entity and cultural term recognition could trigger alternative translations. If "banquet" appears in the context of Thanksgiving, an Uzbek-aware MT might replace it with navro‘z ovqat if appropriate. Tools like semantic tagging or cultural concept tagging could feed into the model to choose the correct equivalent.

Use enhanced evaluation and training signals: Since idioms are hard to evaluate automatically, include them in human evaluation or fine-tuning objectives. For example, include an MT reward for preserving idiom meaning. Crowdsourced or expert evaluation data focusing on idiom translation could be collected to refine the model (reinforcement learning from human feedback on idiomatic sentences).

Post-editing workflows: In practical translation workflows, combine MT with post-editing. If idioms are common in a text, having a bilingual human edit MT output is currently essential.

The post-editing data (MT output + human corrections) can be fed back into training to gradually improve the MT engine's idiomatic accuracy.

Phraseology-aware training data selection: When gathering parallel corpora, prioritize sources rich in idiomatic language (literature, news, speeches, proverbs collections). A balanced mix of literal text and idiomatic text can help the model see both direct and figurative usages.

Lexical resources integration: Incorporate existing English-Uzbek dictionaries of idioms and phrasal verbs into the MT pipeline. For example, during decoding, if the source bigram matches a known idiom, the system could force a jump to the idiomatic target equivalent rather than decode word-by-word.

By implementing these strategies, developers can gradually move MT performance on Uzbek beyond literal translation toward culturally and linguistically appropriate renditions. The goal is a system that recognizes an English idiom and either replaces it with the Uzbek idiom or paraphrases its meaning, instead of translating words in isolation.

References

1. Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT.
2. 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 (ACL), 311-318.
3. Snover, M., Madnani, N., Dorr, B. J., & Schwartz, R. (2009). Fluency, adequacy, or HTER? Proceedings of the Fourth Workshop on Statistical Machine Translation, 259-268.
4. Microsoft. (2025). What is a BLEU score? Microsoft Azure Documentation. Retrieved 2025 from https://learn.microsoft.com
5. Abdurashetona, A. M., Rashidova, U., & Sobirova, M. (2025). The issue of translating idioms between Uzbek and English in natural language processing. AIP Conference Proceedings, 3377(1), 070002.
6. Yaxshimurotovna, S. Y. (2025). CULTURAL-CONNOTATIVE FEATURES OF PHRASEOLOGICAL UNITS IN DIFFERENT LANGUAGES AND THEIR INTERPRETATION THROUGH ARTIFICIAL INTELLIGENCE. SHOKH LIBRARY, 1(13).