Registered in Germany Europa Revue Journal
ISSN: 1862-6564 Vol. 6 No. 11 (2026) Published November 13, 2026

Translating Metaphor from English into Uzbek: Human, Neural Machine, and LLM Performance Compared

Laylo Matyakubova

Trainee - Teacher, Uzbekistan State World Languages University

Abstract

Metaphor is a demanding test for any translator because it ties a linguistic form to a culture-specific conceptual mapping. This essay examines how a human translator, a neural machine translation system and a large language model can be expected to handle English metaphors when translating into Uzbek, a language with comparatively limited digital and parallel data. Drawing on conceptual metaphor theory, established translation procedures and recent work on machine and LLM translation, it distinguishes conventional from novel metaphors and argues that the main risk for machine output is not a lack of fluency but the loss of the underlying mapping. It proposes an error-based evaluation, with categories such as literalisation, wrong mapping, omission, unnatural phrasing and cultural mismatch, and sets out the outcomes the theory would predict. The essay concludes that human post - editing remains necessary for figurative language in English - Uzbek translation, and that prompting and context can help but cannot replace cultural and linguistic judgment.

Keywords

metaphor translation, neural machine translation, large language models, English – Uzbek, error analysis, conceptual mapping

Metaphor is easy to mistake for decoration. We tend to think of it as something added to plain language to make it livelier, a flourish that a careful writer could remove without losing the meaning. Cognitive linguistics has argued almost the opposite. Lakoff and Johnson (1980) showed that much of ordinary thought is organised through metaphor, so that we talk about arguments as if they were wars, about time as if it were money, and about ideas as if they were objects, we can grasp or drop. On this view a metaphorical expression is the visible surface of a deeper mapping between two areas of experience. That is exactly why metaphor is so troublesome for translators. A translator who meets a metaphor has to recognise it, work out which mapping lies behind it, and

then decide whether the target language can carry that mapping, needs a different one, or requires an explanation instead. Doing this well takes cultural knowledge and judgment as well as vocabulary, and this is the reason the translation of figurative language remains one of the hardest tests of any translator, human or machine.

Translation scholars have not always agreed on how far this test can be passed. Dagut (1976) famously asked whether metaphor can be translated at all, arguing that a metaphor in the source language is a semantic novelty and so can have no ready-made equivalent in the target language. Later work took a more practical line. Newmark (1988) listed a set of procedures that translators actually use, from reproducing the same image in the target language, to replacing it with a different but conventional image, to turning it into a simile, to paraphrasing its sense, and finally to leaving it out. Mandelblit (1996) added a cognitive argument: when the source and target languages share the same conceptual mapping, translation is quick and accurate, but when the mappings differ, the translator is slowed down and errors become more likely. Schäffner (2004), working with English and German political texts, showed that the way a translator handles a metaphor affects the text itself and how its readers receive it. Taken together, these studies give us a vocabulary for describing what a translator has done with a metaphor, even if they do not tell us what should be done in every case.

One more distinction is needed before the comparison with machines can begin. Not all metaphors are equally alive. Phrases such as "the price of a mistake" or "a bright idea" are so familiar that most speakers no longer notice the image, and such conventional metaphors usually have established equivalents in other languages, or at least can be rendered in plain sense without much loss. Novel metaphors are different. A writer invents them for a particular text, and their effect depends on the reader noticing the unexpected comparison. A translator cannot look these up in a dictionary and must rebuild them. Researchers who want to study this difference need a repeatable way of deciding what counts as a metaphor in the first place, and the most widely used method is the Metaphor Identification Procedure and its extension, MIPVU, which asks whether a word in context has a more basic meaning that contrasts with its contextual meaning (Steen et al., 2010). The method is slow, but it protects the analysis from the analyst's

own intuitions. Over the last decade the question of who translates has become more complicated. Neural machine translation, which replaced older statistical systems after the introduction of attention-based and Transformer architectures (Vaswani et al., 2017), now produces fluent text for many language pairs and is used daily by millions of people. More recently, large language models trained on enormous quantities of text have been shown to translate at a level close to dedicated systems for languages with plenty of data, although their performance is more limited for low-resource languages (Hendy et al., 2023). They can also follow instructions, which a conventional machine translation engine cannot. A user can tell a language model that a passage is a poem, that the reader is a child, or that a particular phrase is figurative and should not be translated word for word. In principle, that gives the model a way to handle metaphor that older systems lacked. Whether it works in practice is a separate question, and the answer is likely to depend on how much relevant text the model has seen in the target language.

This is where Uzbek becomes an interesting case. It is an agglutinative Turkic language with a rich morphology, it has been written in a Latin alphabet since the language reforms of the 1990s, and the amount of digital text available in it, especially parallel text aligned with English, is small compared with the languages that dominate machine translation research. Systems trained on limited data tend to be strongest where the pattern is frequent and weakest where it is rare, and a novel metaphor is by definition rare. There is also a cultural dimension. Many of the images that Uzbek speakers use for generosity, patience, honour or hardship come from domestic life, hospitality, the body and the natural world in ways that do not map neatly onto English images. A familiar example is the saying that time is gold, which in Uzbek takes a form close to "Vaqt – oltin", while English speakers more often say that time is money. The two images overlap, because both treat time as a valuable resource, but they are not identical, and a translator choosing between a literal and an adapted version is making a decision about which mapping the reader should receive. An idiom such as "qo'li ochiq", literally "open-handed", works for generosity in much the same way as its English counterpart, and a word-for-word rendering happens to succeed. Such coincidences are helpful, but they are exactly the cases in which a system can appear skilful without understanding

anything. A human translator brings to this problem the ability to read the whole text, to know who the readers are, and to accept that a good translation will sometimes look quite different from its source. A trained translator will notice that a metaphor is carrying an argument and will choose a strategy accordingly, sometimes keeping the image, sometimes replacing it, sometimes dropping it and compensating elsewhere in the sentence. Human translators are not infallible, though. They get tired, they miss images that have become invisible to them, and they sometimes keep a source image out of loyalty to the original even when it confuses readers. A realistic comparison therefore should not assume that the human is the gold standard in every case. It should instead ask what kinds of mistakes each kind of translator tends to make.

Making that comparison rigorous is harder than it looks. Automatic scores such as BLEU (Papineni et al., 2002) measure how many words or word sequences a translation share with a reference, and a good metaphor translation may share almost none with the reference, so the score punishes creativity and rewards literalness, which is the opposite of what a figurative text need. Human evaluation with defined error categories is a better approach. The Multidimensional Quality Metrics framework (Lommel et al., 2014) asks trained raters to mark each error by type and severity rather than to give a single impression of quality. For metaphor, the categories can be adapted. One can distinguish literalisation, where the image is carried over word for word and the meaning is lost; wrong mapping, where a different image or sense is substituted; omission, where the metaphor disappears without compensation; unnatural phrasing, where the sense survives but the Uzbek sounds foreign; cultural mismatch, where the image is inappropriate or puzzling for Uzbek readers; and ordinary grammatical errors in the metaphorical phrase. A study could then collect a corpus of English metaphors from literary, journalistic and popular-science texts, label each as conventional or novel, obtain translations from a human translator, a neural system and a language model, and have bilingual raters judge them without knowing their source. Agreement between raters could be reported with Cohen's kappa.

What would such a study be likely to find? Prior work does not answer the question for English and Uzbek, and it would be a mistake to state results before the data exist,

but the theory suggests some reasonable expectations. Conventional metaphors should be handled well by every kind of translator, since they are frequent and often have settled equivalents. The differences should appear with novel metaphors, where a system without real understanding is most likely to translate literally and produce Uzbek that is grammatical but strange, or to choose a plausible but wrong sense. Culture-bound images should cause trouble for machine systems for the same reason, because the expression that a human would choose to preserve the effect is rarely the one that appears most often in training data. Context should matter as well: a metaphor translated inside a full paragraph gives a system more clues about what is being compared than the same sentence on its own. And a language model that is told a metaphor is present may do better than one that is not told, because the instruction directs it away from the most literal reading. If these expectations were confirmed, the main result would be that fluency and fidelity to metaphor come apart: a system can produce smooth Uzbek and still lose the mapping that made the original interesting. If they were not confirmed, that would be equally valuable, because it would show that the newest systems are better at figurative language than the theory predicts. For translators and teachers, the practical implications follow from this. Machine output for English and Uzbek, even when it reads smoothly, should be post-edited by someone who can recognise a metaphor and judge whether its effect has survived. Training programmes could use the error categories as a teaching tool, asking students to identify what went wrong in a machine translation and to explain it in terms of conceptual mapping rather than simply marking it as a mistake. Users who rely on language models for translation can improve their results by giving context and by flagging figurative language, although the instruction is no substitute for a reader who knows Uzbek. For developers, the cases that matter most are the rare and culture-bound ones, and a benchmark made of such cases would be more informative than another test set of ordinary sentences.

Any such work has limits that should be stated plainly. A corpus of a hundred or so expressions can only suggest patterns, not establish them. One human translator cannot stand for all human translation, and two raters cannot speak for every Uzbek reader. Machine systems change from month to month, so results are true only for the versions

tested, and the output of a language model depends on how it is prompted. These are reasons for caution, not reasons for avoiding the question. The question is worth asking because Uzbek readers will meet machine-translated text more and more often, and it matters whether the figurative side of the original survives the trip.

The central argument of this essay is that metaphor is a useful test of translation quality precisely because it cannot be passed by fluency alone. A translator, whether a person, a neural network or a language model, has to recover the mapping behind the words and find a way to give the target reader something that works in the same way. Human translators do this through knowledge of both cultures and judgment about readers, and machines do it, when they do it at all, through patterns learned from data that for Uzbek remain thin. Comparing the three directly, with careful identification of metaphors, clear error categories and transparent evaluation, would show where each succeeds and where it fails, and would give translators, teachers and developers evidence to work with instead of impressions.

References

1. Dagut, M. B. (1976). Can "metaphor" be translated? *Babel*, *22*(1), 21–33.

2. Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M., & Awadalla, H. H. (2023). *How good are GPT models at machine translation? A comprehensive evaluation*. arXiv. https://doi.org/10.48550/arXiv.2302.09210

3. Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press.

4. Lommel, A., Uszkoreit, H., & Burchardt, A. (2014). Multidimensional Quality Metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica: Tecnologies de la Traducció*, *12*, 455–463. https://doi.org/10.5565/rev/tradumatica.77

5. Mandelblit, N. (1996). The cognitive view of metaphor and its implications for translation theory. In M. Thelen & B. Lewandowska-Tomaszczyk (Eds.), *Translation and meaning: Part 3* (pp. 483–495). Hogeschool Maastricht, School of Translation and Interpreting.

6. Newmark, P. (1988). *A textbook of translation*. Prentice Hall.

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

8. Schäffner, C. (2004). Metaphor and translation: Some implications of a cognitive approach. *Journal of Pragmatics*, *36*(7), 1253–1269. https://doi.org/10.1016/j.pragma.2003.10.012

9. Steen, G. J., Dorst, A. G., Herrmann, J. B., Kaal, A. A., Krennmayr, T., & Pasma, T. (2010). *A method for linguistic metaphor identification: From MIP to MIPVU*. John Benjamins. https://doi.org/10.1075/celcr.14

10. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30, pp. 5998–6008). Curran Associates.