Project: Translation of Idioms in Large Language Models (LLMs) and Machine Translation (MT) Tools from English to Slovene

Authors: Pia Polutnik and Lea Vodopivec

Mentor: Assoc. prof. Petra Bago, PhD

Faculty of Humanities and Social Sciences, University of Zagreb

Digital Linguistics Project, Academic Year 2025/26

1. Introduction and Hypothesis

Idioms represent one of the most challenging aspects of translation, as their meanings cannot be derived from the literal interpretation of individual words. They are culturally and linguistically bound expressions that require translators, and increasingly, translation systems, to navigate complex semantic and pragmatic layers. Idioms depend on shared cultural knowledge, which makes direct, word for word translations often nonsensical or misleading in the target language.

A number of typologies have been proposed to account for how idioms are translated across languages. Božena Horváthová (2014) outlines four principal strategies: paraphrasing, equivalence, literal translation, and omission. Paraphrasing involves expressing the idiomatic meaning through non-idiomatic expressions that preserve the semantic content of the source phrase. Equivalence refers to the use of a target-language idiom that conveys the same or a similar meaning as the original, while differing in form but corresponding in pragmatic effect. Literal translation transfers the idiom word for word and frequently results in semantically inaccurate output. Omission occurs when the translator removes the idiom entirely, typically when no suitable equivalent or paraphrase exists.

With the rapid development of machine translation (MT) systems and large language models (LLMs), idiom translation has become a useful test of how well these technologies capture linguistic nuance and cultural specificity. While both MT and LLM systems can generate grammatically correct translations, idioms expose the limitations of their semantic and contextual interpretation.

This project examines how LLMs and traditional MT tools translate idioms from English to Slovene, focusing on how often each strategy occurs and how accurate the translations are. We hypothesize that when translating idioms from English to Slovene, LLMs will predominantly produce literal translations, followed by semantically accurate and fewer culturally adapted translations. However, compared to traditional machine translation tools, we expect LLMs to yield a higher proportion of semantically accurate and culturally adapted translations.

2. Literature Review

In the paper *Improving LLM Abilities in Idiomatic Translation* (Donthi et al, 2025), the authors address the challenges LLMs face in translating idiomatic expressions. They note that LLMs often produce literal translations that fail to preserve idiomatic style and cultural nuance. Existing resources, such as IdiomKB, are limited in language coverage and often inadequate for maintaining idiomatic integrity. The study introduces two methods aimed at improving idiomatic translation. The first, Semantic Idiom Alignment (SIA), employs pre-trained sentence embeddings and cosine similarity to locate idiomatic equivalents in the target language, leveraging datasets like IdiomKB. The second, LLM-based Idiom Alignment (LIA), uses prompt-based generation, instructing LLMs to produce idiomatic counterparts directly in the target language. Human evaluations showed that SIA outperformed both LIA and direct translation in preserving idiomatic style and cultural authenticity. GPT-4o evaluations were also more closely aligned with human judgments than those of GPT-3.5. The results demonstrated that SIA was particularly effective for maintaining idiomatic integrity, especially in English–Chinese translations. LIA performed well but inconsistently, being more sensitive to prompt design, while direct translation frequently missed idiomatic nuances though it succeeded in simpler cases. The study acknowledges limitations, including the scarcity of idiom datasets for low-resource languages and the reduced reliability of GPT-based evaluations for languages like Urdu due to limited human annotators. Overall, the research contributes to improving cross-lingual communication by preserving idiomatic and cultural subtleties, with significant implications for literary and educational translation and for fostering cross-cultural understanding.

Castaldo and Monti (2024) investigate how prompt design affects the ability of LLMs to translate idiomatic expressions accurately. The study evaluates four prompt templates—A, B, C, and D—and finds that Prompt D, which employs a five-shot learning setup, consistently outperformed the others according to BLEURT and COMET scores across both GPT-3.5 API and Mistral-7B models. Among the zero-shot configurations, Prompt C achieved the best results. Overview of the prompt templates used in this study:

Prompt ID	Prompt Template
A	[src]: [input] ♦ [tgt]:
B	Please provide the [tgt] translation for this sentence: [input] ♦ Translation:
C	This is a [src] to [tgt] translation, please provide the [tgt] translation for this sentence: [input] ♦ Translation:
D	[src]: [$source_1$] ♦ [tgt]: [$target_1$] ♦ ... [src]: [$source_k$] ♦ [tgt]: [$target_k$] ♦ [src]: [input] ♦ [tgt]:

In terms of model performance, GPT-3.5 generally surpassed Mistral-7B in idiomatic translation quality. Nonetheless, Mistral-7B demonstrated competitive results when appropriately prompted, particularly under the Prompt D configuration, suggesting that careful prompt design can partially compensate for model scale differences. The authors critique traditional evaluation metrics such as BLEU for their limited ability to assess idiomatic translation, as these metrics emphasize n-gram overlap and sentence length rather than semantic or pragmatic fidelity. In contrast, neural-based metrics like COMET and BLEURT proved more effective in capturing the semantic and contextual nuances of idiomatic meaning. The dataset used in the study consisted of 254 Italian–English sentence pairs containing idiomatic expressions presented in both literal and figurative contexts. These pairs were drawn from the Italian Dodiom corpus and the Reverso Context database. The authors acknowledge several limitations, including the relatively small dataset size, dependence on automated evaluation methods, and the need for more comprehensive testing of prompt strategies across additional language pairs.

In her thesis, Razum (2024) examines the quality of idiom translation produced by two machine translation systems: Google Translate, representing neural machine translation (NMT), and ChatGPT-3.5, representing LLMs. The study analyzes 126

English phraseological units translated into Croatian and evaluates them through human assessment based on the MQM error typology. Both systems achieved comparable accuracy: Google Translate correctly translated 60 out of 126 idioms, while ChatGPT-3.5 achieved 61 correct translations. The most frequent and severe error type in both systems was literal translation, which often failed to capture idiomatic meaning. Contrary to the initial hypothesis that paraphrasing would dominate, both systems more frequently used idiomatic equivalents than paraphrases. However, ChatGPT-3.5 occasionally produced hallucinated or semantically divergent translations that had no basis in the source text, a problem not observed in Google Translate. The author attributes the overall lower performance of both systems partly to the low-resource status of Croatian, which limits training data and model accuracy. When idiomatic equivalents were unavailable, both systems resorted to paraphrasing strategies, similar to human translation behavior. The findings indicate that current MT systems, including advanced LLMs, still lack reliability in handling idiomatic and figurative language. Human evaluation and intervention remain indispensable, and machine translation should be regarded primarily as an assistive tool rather than a replacement for professional human translators.

3. Objectives and Goals

The research analyzes and evaluates the accuracy of translating idiomatic expressions from English to Slovene using LLMs and MT tools. The methodology employs a two-level annotation scheme:

- First-level annotation: Assess whether the meaning of the idiom is conveyed in the translation (yes or no).
- Second-level annotation: classifies the translation as one of the following:
    - Paraphrasing: Semantically accurate translation.
    - Equivalence: Culturally adapted translation.
    - Literal translation: Word for word rendering.
    - Other: Any translation not fitting the above categories.

The research compares translation methods and error rates across different LLMs and MT tools. A small sub-corpus of sentences containing idiomatic expressions is developed, sourced from existing datasets that can be used for further studies. The empirical analysis demonstrates how LLMs and MT tools handle idiomatic expressions and culturally specific units of meaning.

4. Project Timeline

I. Review of literature on idiom translation; defining research questions

II. Creation of the experimental dataset

III. Collection of results from different models (LLMs and MT systems)

IV. Annotation

V. Comparison of results across models; creation of tables and visualizations

VI. Writing and structuring of the final report

VII. Final checks and preparation of the presentation

5. Methodology

5.1. Dataset Construction and Sentence Selection

Our subcorpus of 40 idioms was drawn through random sampling from the EPIE Corpus dataset (https://github.com/prateeksaxena2809/EPIE_Corpus) to ensure that the selection was not guided by prior assumptions about frequency or semantic transparency. Each idiom was then submitted as a search query in SketchEngine within the The English Web 2021 Corpus (enTenTen21). The Good Dictionary Example (GDEX) function was applied to automatically rank and filter concordance lines so that only syntactically clear, contextually interpretable, and semantically representative uses of the idiom were retained. For every idiom, 10 sentences with high GDEX scores were extracted, yielding a balanced collection of 400 naturally occurring idiom instances suitable for the research.

5.2. Translation Systems and Input Preparation

Two MT systems and two LLMs were used to produce the translations. The LLMs included ChatGPT and Google Gemini, with a new chat created for each input to ensure each group of sentences was treated individually. The MT tools used were Google Translate and DeepL. All systems received the same set of 10 unrelated sentences at once, each sentence containing a different idiom. Sentence grouping and organization for analysis were facilitated using a short Python script (corpus_grouping.py), which produced the final text file (ReGrouped_Corpus_of_idioms.txt) containing randomly mixed and grouped sentences, ready for use.

5.3. LLM Prompting

In the use of LLMs, one of the key factors is prompting, which guides how the model interprets input and generates output. Common prompting strategies include: zero-shot prompting, where the model responds without any examples; one-shot prompting, where a single example is provided to guide the response; few-shot prompting, which uses multiple examples to illustrate the desired pattern; and chain-of-thought prompting, which encourages the model to reason step by step to improve accuracy. For this research, a combination of two prompting techniques was employed: persona-based prompting and one-shot prompting. The model was instructed to assume the role of an expert English–Slovene translator and was provided with explicit translation guidelines alongside a single example of an English idiom translated into Slovene. This example served to condition the translation of subsequent idiomatic expressions. The prompt required the translation of each English sentence into Slovene in a manner that preserves semantic accuracy and achieves naturalness in the target language. The prompt was applied at the beginning of each 10 sentence cluster.

Our prompt was:

> You are an expert translator specializing in English–Slovene. Your task is to translate each English sentence into Slovene so that the translated sentence conveys the meaning of the original as accurately and naturally as possible. Follow the style and logic of the example below.

The sentences should be organized like this and are not connected to each other:

ORGANIZATION: [ID] [SENTENCE]

Example:
Input (EN): 1. We wanted to go for a walk, but it suddenly started raining cats and dogs.
Output (SL): 1. Hoteli sva iti na sprehod, a je nenadoma začelo liti kot iz škafa.

Now translate the following sentences:

Input (EN): [SENTENCES]

The sentences are independent, with no connection to each other to prevent influence from surrounding context. Each sentence was assigned a unique number ID to maintain systematic organization and allow precise tracking. The instruction was phrased as “Your task is to translate” rather than a simple command, emphasizing the translator’s professional role and clarifying the objective of the task.

For Gemini, an additional constraint was introduced to control output variability. In initial runs, Gemini systematically generated multiple alternative translations for a single source sentence, separated by a slash, which conflicted with the project requirement of one target sentence per input. To ensure methodological consistency and comparability with other LLMs, the prompt was therefore extended with an explicit restriction instructing the model to generate only a single translation and to avoid offering alternative renderings. Our prompt was modified by adding “Do not give different options (do not use a /) in the sentence translations.” at the end of the first part of the prompt.

6. Annotation of the Translations

After the data was collected, annotation was carried out manually by two annotators. The dataset consisted of 40 idioms, each represented by 10 sentences, yielding 400 source examples. Each example was translated by all four of the systems, resulting in a total of 1600 translated sentences. The annotation workload was divided evenly, so that each annotator annotated 800 translations. Each annotator worked on the output of one MT tool and one LLM.

The annotation was done in Excel, following the two-layer annotation scheme described above. The guidelines were defined before the annotation began and were applied consistently throughout the dataset. In the first layer, the annotation captured whether the meaning of the idiom was conveyed in the translation (yes/no). In the second layer, translations were classified according to the translation strategy used (literal translation, paraphrasing, cultural equivalence, or other).

For clarity, one representative example is provided for each translation strategy. Paraphrasing refers to translations that accurately convey idiomatic meaning without preserving idiomatic form (e.g. *EN: Erratic supply of water can really be a pain in the neck → SL: Neredna oskrba z vodo je lahko res nadležna.*). Equivalence involves the use of a target-language idiom with comparable meaning and pragmatic effect (e.g. *EN: Installing SP2 is a piece of cake and won't take you long. → SL: Namestitev SP2 je mačji kašelj in vam ne bo vzela veliko časa.*). Literal translation corresponds to word-for-word renderings (e.g. *EN: He makes barely enough money to keep body and soul together. → SL: Zasluži komaj dovolj denarja, da ohrani telo in dušo skupaj.*). The category *Other* includes translations that do not fit any of the above strategies, such as incomplete or semantically divergent outputs.

When annotating, the focus was on the idiom itself rather than the overall quality of the sentence. If the idiom was translated correctly and its meaning was conveyed, the translation was marked as correct in the first layer even if the rest of the sentence contained errors or did not fully match the meaning of the source sentence. There were 40 such cases, accounting for 2.5% of the data.

Sentences that were unclear or difficult to classify were marked during annotation and later discussed together. In these cases, a joint decision was reached through discussion. A comparative analysis was conducted across different models and MT tools to assess differences in accuracy and treatment of idiomatic expressions.

7. Data Processing and Analytical Framework

Two Python scripts were used to implement the Translation Idiom Evaluation Pipeline, providing an end-to-end evaluation of idiom translation quality across four systems. The pipeline covers data inspection, structural validation, parsing, statistical analysis, visualization, and report generation from an ODS spreadsheet.

The script *explore_data.py* was used to establish an explicit understanding of the spreadsheet's internal logic before formal parsing and analysis. The script is diagnostic, not analytical. It processes Translations.xlsx.ods loaded via pandas.read_excel(..., engine='odf'). Its preliminary role is to prevent silent parsing errors, ensuring that the spreadsheet follows a hierarchical structure (idiom → sentence → annotation) and that annotation values are consistent enough to be algorithmically interpreted.

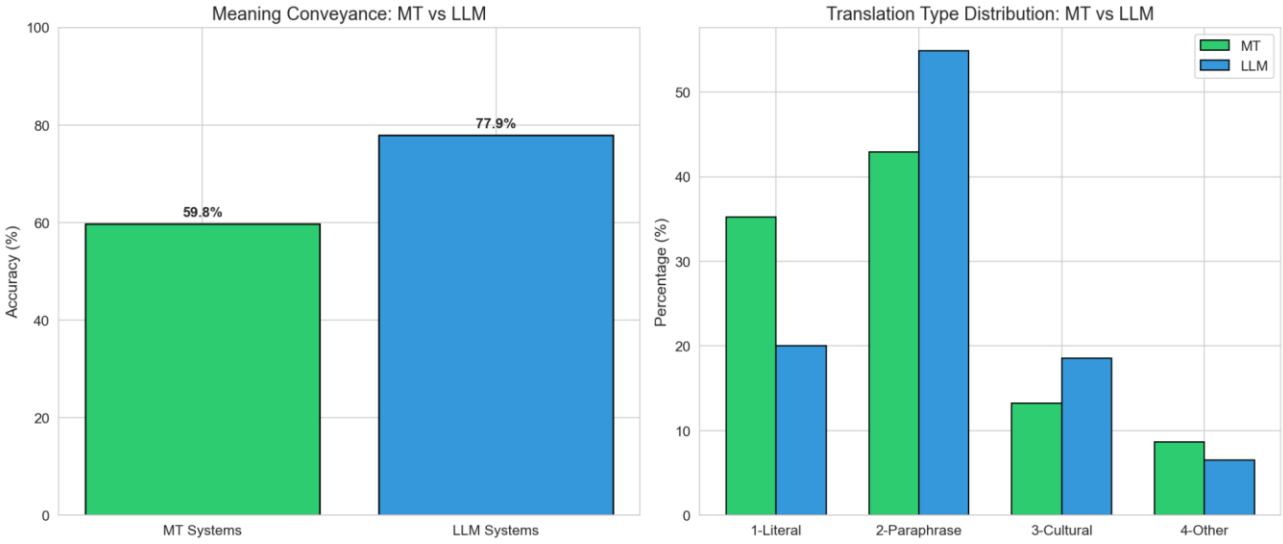
The script *data_analysis.py* was used to transform the semi-structured spreadsheet into a normalized dataset and compute layered evaluation statistics with reproducible outputs. Each observation in the Conceptual Model corresponds to: one idiom, one sentence instance, one system, one translation and two evaluation layers. The result is a tidy DataFrame with one row per system × sentence × idiom.

8. Results

Figure 1 shows the overall accuracy of idiom translation across the four systems. Gemini achieved the highest accuracy, correctly conveying idiomatic meaning in 87.2% of cases. ChatGPT follows with an accuracy rate of 68.5%, while DeepL achieved a comparable result with 67.1%. Google Translate performed the weakest, with only 52.5% of translations successfully conveying the intended idiomatic meaning. The results demonstrate a clear performance gap between the highest and lowest scoring systems, with Gemini outperforming Google Translate by 34.7%.

Figure 1: Meaning Conveyance Accuracy by System (Yes/No)

Figures 2 and 3 depict the aggregated results comparing LLMs and MT tools. Figure 2 shows that LLMs achieved a higher average accuracy (77.9%) than the MT tools, which reached an average accuracy of 59.8%. Figure 3 demonstrates the distribution of translation strategies used by MT systems and LLMs. Both rely most frequently on paraphrasing, with LLMs using this strategy in over 50% of cases and MT tools in over 40%. Literal translation is more common in MT systems than in LLMs, indicating a stronger tendency toward word for word rendering among MT tools. Cultural equivalence is used more often by LLMs than by MT systems, although both rely on this strategy less frequently than on paraphrasing and literal translation.



Figures 2 and 3: Meaning Conveyance; Translation Type Distribution – LLM/MT Comparison

Figure 4 illustrates the distribution of translation strategies used by each system. All systems rely most frequently on paraphrasing (type 2), with Gemini employing this strategy most often (59.9% of the time). Literal translation (type 1) is the second most common strategy for all systems except Gemini, where paraphrasing takes the second spot with 23.3%. Although literal translation also ranks second for ChatGPT, MT systems use it more frequently overall.

Figure 4: Translation Type Distribution by System

The results shown in Figure 5 indicate the success rate of meaning conveyance by translation type. Cultural equivalence (type 3) achieves the highest proportion of correct meaning transfer across all systems, tying with paraphrasing in the case of DeepL. Paraphrasing (type 2) also shows relatively high success rates across all systems. Literal translation (type 1) demonstrates consistently low success rates, remaining below 50% across all systems. Gemini exhibits the highest overall accuracy, with a 100% success rate when opting for cultural equivalence.

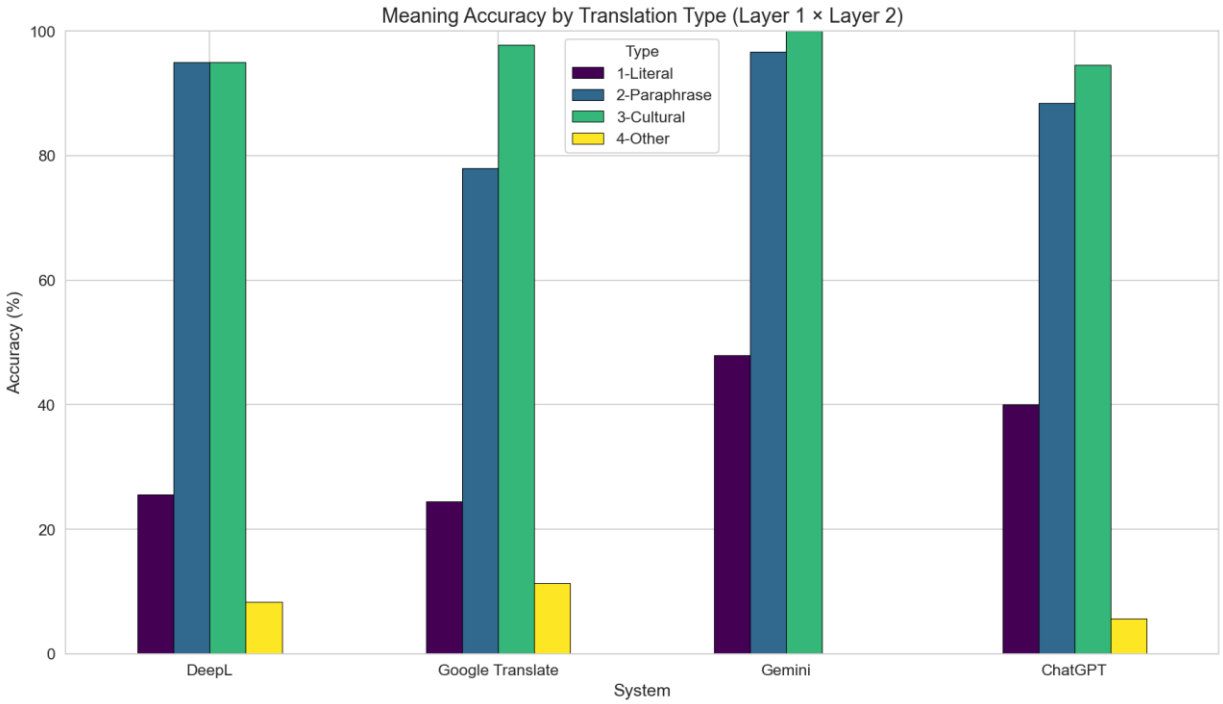


Figure 5: Accuracy by Type per System (%)

Figure 6 shows the accuracy of meaning conveyance for each idiom across the four systems. The heatmap illustrates substantial variation in system performance depending on the specific idiom. The only idiom that reached 100% accuracy rate across all systems was *so far so good*, even though no Slovene literal equivalent exists.

Gemini consistently achieved high accuracy across most idioms, reaching a 100% accuracy for 22 out of 40 idioms. The idioms that stand out are *once in a blue moon, play safe*, and *keep [pron] head above water,* for which its accuracy was under 30%. No idiom achieved 0% accuracy.

DeepL demonstrates mixed performance. While it achieved perfect accuracy for 8 idioms, it performed poorly on several idioms, with the accuracy dropping to 0% for 3 idioms (*out of this world, keep [pron] head above water, drink like a fish*).

Google Translate showed the lowest and most unstable performance across idioms. 12 expressions received 0% accuracy, including *dark horse*, *draw in [pron] horns*, *draw the shortest straw*, *drink like a fish*, *keep body and soul together*, *keep the wolf from the door*, *never-never land*, *not bat an eyelid*, *on cloud nine*, *the last straw*, *walk on air*, and *water under the bridge*. However, the system achieved 100% accuracy in 9 cases.

ChatGPT performed more consistently than the MT systems but showed greater variability than Gemini. It achieved 100% accuracy for only 6 idioms. Its performance dropped significantly for idioms such as *dark horse*, *draw fire*, *not bat an eyelid, on cloud nine, play safe*, and *walk on air*, where accuracy ranges between 20% and 30%. As with Gemini, no idiom achieved 0% accuracy.

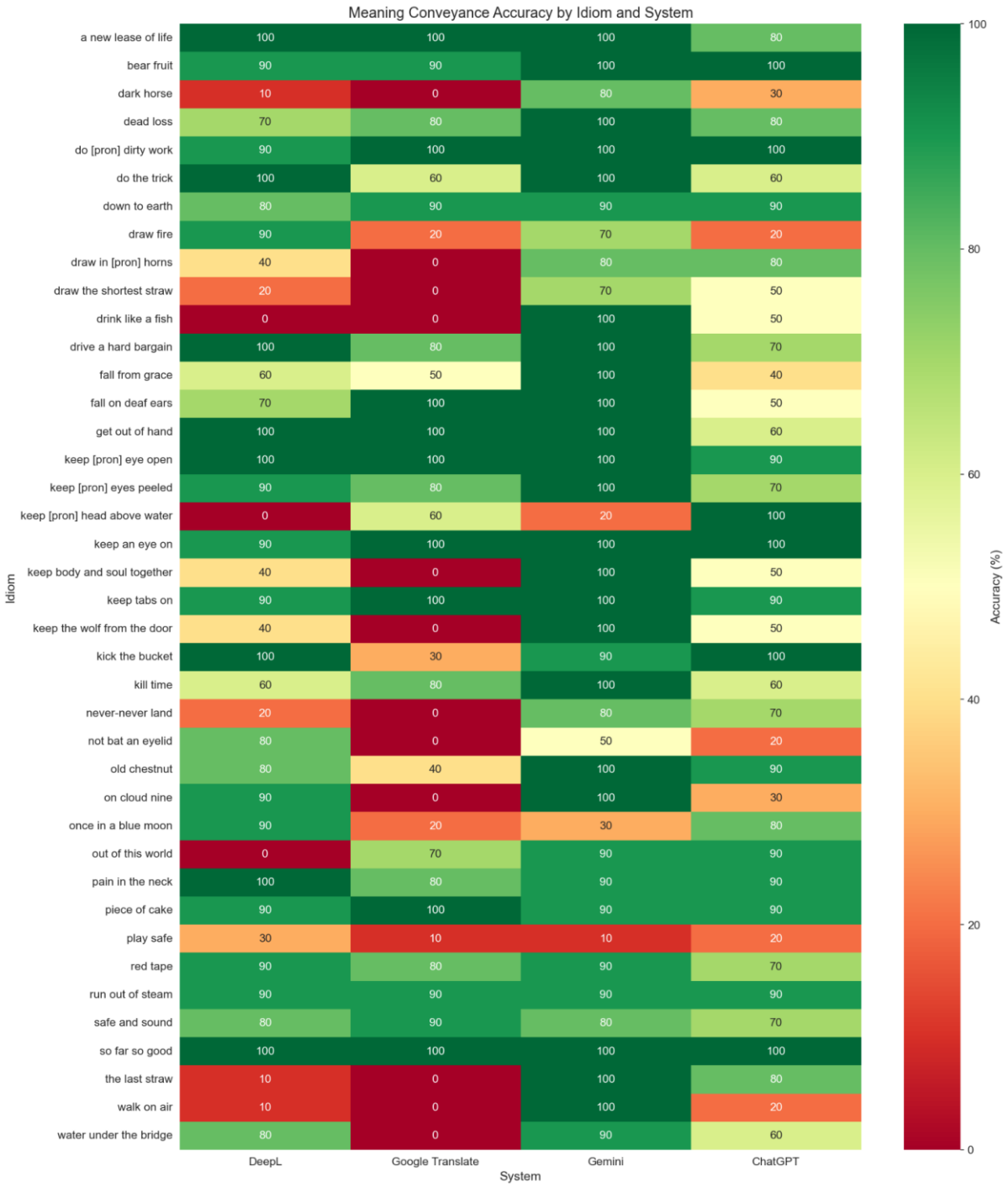


Figure 6: Meaning Accuracy by Idiom and System (%)

Table 1 reports the number of special cases per system. These are examples where the idiom was translated correctly, yet the sentence as a whole is grammatically incorrect (*e.g. EN: A development which seems likely to bear fruit was the creation of a working party from within the panel. → SL: Razvoj, ki se zdi verjetno obrodil sadove, je bil ustanovitev delovne skupine znotraj panela*). ChatGPT produced the highest number of such cases (18), followed by Google Translate and Gemini with 9 cases each. DeepL produced 4 instances of this type. These cases represent translations where the idiom meaning was conveyed correctly, but the sentence-level realization was pragmatically or stylistically unusual. There were a total of 40 such cases, which represent 2.5% of all translations. example

System	Count
DeepL	4
Google Translate	9
Gemini	9
ChatGPT	18

Table 1: Special Cases Count by System

9. Discussion

The findings of this study provide new insights into the performance of LLMs and MT systems in translating idiomatic expressions from English to Slovene. Contrary to the initial hypothesis that LLMs would predominantly produce literal translations, the data indicate that both LLMs and MT systems favor paraphrasing. For DeepL, Google Translate, and ChatGPT, literal translation was the second most common strategy, whereas Gemini's was cultural equivalence. The second part of the hypothesis, namely that LLMs tend to produce a higher proportion of semantically accurate and culturally adapted translations compared to MT systems, was largely confirmed. However, this advantage was not uniform across all models, as ChatGPT and DeepL exhibited comparable performance in several cases.

Gemini achieved the highest overall accuracy in meaning conveyance, followed by ChatGPT, DeepL, and Google Translate. These results may suggest that the LLMs

used are generally more successful than traditional MT systems in conveying idiomatic meaning, although ChatGPT and DeepL achieved comparable performance. Paraphrasing and cultural equivalence accounted for the majority of successful translations, whereas literal translations rarely conveyed the correct idiomatic meaning.

Notably, Gemini consistently outperformed ChatGPT across most idioms, achieving higher overall accuracy and more reliable meaning conveyance. This suggests that not all LLMs are equally effective for idiomatic translation tasks and underscores the importance of evaluating multiple models when selecting tools for cross-lingual translation.

Some limitations should also be acknowledged. Human error during annotation may have influenced the accuracy of the classifications, and no inter-annotator agreement metrics were calculated, limiting the reliability of subjective judgments. The study relied on publicly available versions of LLMs and MT systems; API versions or future model updates could produce different results. Finally, the relatively small size of the sub-corpus constrains the generalizability of these findings, particularly for low-frequency or highly culture-specific idiomatic expressions. Further studies could explore prompt engineering, larger datasets, and cross-lingual comparisons to better understand how LLMs generalize idiomatic knowledge.

10. References

Castaldo, Antonio and Johanna Monti. 2024. Prompting Large Language Models for Idiomatic Translation. In *Proceedings of the 1st Workshop on Creative-text Translation and Technology*, pp. 32–39. European Association for Machine Translation. https://aclanthology.org/2024.ctt-1.4/.

Donthi, Sundesh, Maximilian Spencer, Om B. Patel, Joon Young Doh, Eid Rodan, Kevin Zhu, and Sean O'Brien. 2025. Improving LLM Abilities in Idiomatic Translation. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pp. 175–181. Association for Computational Linguistics. https://aclanthology.org/2025.loreslm-1.13/.

Razum, Sandra. 2024. *Machine Translation of Phraseological Units from English to Croatian*. Master's thesis, University of Zagreb, Faculty of Humanities and Social Sciences. https://urn.nsk.hr/urn:nbn:hr:131:950907.