Project: Translation of idioms in large language models (LLMs) and machine translation (MT) tools from English to Slovene

1. Introduction and Hypothesis

Idioms represent one of the most challenging aspects of translation, as their meanings often extend beyond 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. The difficulty lies in the fact that idioms depend on shared cultural knowledge, making direct, word-for-word translations often nonsensical or misleading in the target language.

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

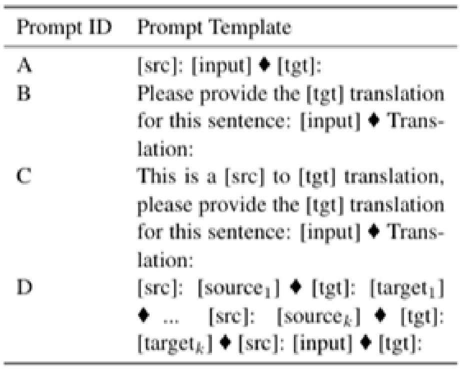
With the rapid development of machine translation (MT) systems and large language models (LLMs), idiom translation has become a revealing test of how well these technologies capture linguistic nuance and cultural specificity. While both MT and LLM systems can generate grammatically correct translations,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

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. Our hypothesis is 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, LLMs are expected to yield a higher proportion of semantically accurate and culturally adapted translations.

2. Literature Review and Previous Researches

In the paper *Improving LLM Abilities in Idiomatic Translation* (Sundesh Donthi, Maximilian Spencer, Om Patel; https://aclanthology.org/2025.loreslm-1.13/), the authors address the challenges large language models (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.

In the paper *Prompting Large Language Models for Idiomatic Translation* (Antonio Castaldo, Johanna Monti; https://aclanthology.org/2024.ctt-1.4/), the authors investigate how prompt design affects the ability of large language models 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 *Machine Translation of Phraseological Units from English to Croatian* (Sandra Razum; https://repozitorij.ffzg.unizg.hr/islandora/object/ffzg:12162), the author examines the quality of idiom translation produced by two machine translation systems: Google Translate, representing neural machine translation (NMT), and ChatGPT-3.5, representing large language models (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 (word-for-word) 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 involves analyzing and evaluating the accuracy of translating idiomatic expressions from English to Slovene using large language models (LLMs) and machine translation (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: classify 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 also involves comparing error rates across different LLMs and MT tools. A small sub-corpus of sentences containing idiomatic expressions will be developed, sourced from existing datasets that can be used for further studies. The empirical analysis will demonstrate 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 translators)
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. Used Datasets

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 our research.

6. Methods and Tools

The analytical approach consists of qualitative techniques aimed at evaluating idiom translations. Translations are classified according to previously mentioned typologies of idiom translation, followed by qualitative data analysis to identify patterns, errors, and translation strategies. A comparative analysis is conducted across different models and MT tools to assess differences in accuracy and treatment of idiomatic expressions.

The tools employed include LLMs and MT tools. Among the LLMs were used ChatGPT, with a new chat for each input to ensure each group sentences are treated individually, and Google Gemini. Among MT tools were used Google Translate and DeepL. The same set of sentences (with different idioms) is provided to all systems. Sentence grouping and organization for analysis is facilitated using a short Python script (corpus_grouping.py). This script's result is the final text file (ReGrouped_Corpus_of_idioms.txt) with randomly mixed and grouped sentences, ready for use.

7. LLMs 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, one-shot prompting is employed, providing the model with instruction and a single example of translation of an idiom from English to Slovene to guide the translation of idiomatic expressions. In addition, the setup adopts a fixed ratio of one prompt per ten sentences, each sentence containing a different idiom from the previously described mixed groups. Each prompt is executed in a new chat instance to prevent context carryover and isolate model behavior.

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 prompt directs an expert English–Slovene translator to translate each English sentence into Slovene, ensuring that the meaning is conveyed accurately and naturally. The sentences are independent, with no connection to each other, and each is assigned a unique ID to maintain systematic organization and allow precise tracking. The instruction is 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. These design choices serve specific purposes: independent sentences prevent influence from surrounding context, ID numbers enable clear structure and reference, the phrasing frames the task as a professional responsibility, and the focus on meaning and naturalness guides translators to produce semantically faithful and fluent translations. Overall, the prompt ensures clarity, consistency, and reliability in translation research.

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. This adjustment targets Gemini's tendency toward option generation and forces a deterministic output format aligned with the research design. By eliminating parallel translations, the revised prompt ensures that each source sentence corresponds to exactly one Slovene output, preserving the integrity of quantitative and qualitative analysis. Our prompt was modified by adding "Do not give different options (don't use a /) in the sentence translations." at the end of the first part of the prompt.

8. Annotation of the translations

The annotation was carried out manually by two annotators. The dataset consisted of 400 translated sentences, which were divided evenly so that each annotator annotated 200 sentences. The division was done in a way that each annotator worked on the output of one machine translation tool and one large language model.

The annotation was done in Excel, following the two-layer annotation scheme described above. The annotation guidelines were defined before the annotation began and were applied consistently throughout the dataset. In the first layer, we annotated 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).

When annotating, we focused 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 10% of the data.

Sentences that were unclear or difficult to classify were marked during annotation and later discussed together. In these cases, we reached a joint decision through discussion.

9. Data interpretation and analysis

For the Translation Idiom Evaluation Pipeline were used two Python scripts that together implement an end-to-end evaluation pipeline for idiom translation quality across four systems: two MT engines (DeepL, Google Translate) and two LLMs (Gemini, ChatGPT). The pipeline covers data inspection, structural validation, parsing, statistical analysis, visualization, and report generation from an ODS spreadsheet.

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

The script *data_analysis.py* transforms the semi-structured spreadsheet into a normalized dataset and computes 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 (layer 1: meaning conveyance (yes/no) and layer 2: translation strategy (1–4)). The result is a tidy DataFrame with one row per system × sentence × idiom.
Analytical Layers:

- Layer 1: Meaning Conveyance
    - Accuracy per system (% yes).
    - Aggregated MT vs LLM comparison.
    - Per-idiom accuracy matrices.
    - Raw yes/no counts retained for transparency.
- Layer 2: Translation Strategy
    - Distribution and counts of types 1–4 per system.
    - MT vs LLM preference profiles.
    - Explicit labeling:
        - Type 1: Literal
        - Type 2: Paraphrase
        - Type 3: Cultural equivalent
        - Type 4: Other
- Combined Analysis (Layer 1 × Layer 2)
    - Success rates by translation type.
    - System-specific effectiveness of each strategy.

- Layer 3: Special Cases (“Red Cells”)
    - Operationalized as meaning == yes combined with type == 4.
    - Flagged as potential cases where idiom translation succeeds but sentence-level adequacy fails.
    - Counts and full case listings produced per system.

10. Results

Data Visualization: Overall meaning accuracy by system, MT vs LLM accuracy and strategy comparison, Stacked type distributions per system, Idiom × system accuracy heatmap, Accuracy by translation type per system, Meaning conveyance pie charts per system, Strategy distribution pie charts per system.