Transidiomation: Optimizing translation of idioms embedded in text

Dhondup Dolma, Jaeeun Lee, Yongtian Ou, Jiyoon Pyo
Team NLPitch
{dolma012,lee03881,ou000036,pyo00005}@umn.edu

Abstract

Using neural machine translation (NMT) tools, such as Google Translate and DeepL, as a mode of translation has become prevalent in recent years. Though the performance of translation tools has improved significantly over the past decades, translation software and large language models (LLMs) like ChatGPT still suffer when translating sentences that involve idiomatic expressions. In this work, we propose a novel and simple prompting framework, **Transidiomation**, that can improve language models' capability of translating sentences with idioms.

1 Introduction

An idiom is an expression in the usage of a language that is peculiar to itself either in having a meaning that cannot be derived from the conjoined meanings of its elements (e.g., *up in the air* for 'undecided') or in its grammatically atypical use of words (e.g., *give way*). A phrase used as an idiom is interpreted figuratively according to convention and is usually indirectly derived from the individual meanings of its components. Furthermore, often the ability to recognize and grasp the meaning underlying an idiomatic expression demonstrates how proficient one is in a language. Therefore, the task of translating a sentence with an idiom embedded requires extra steps of identifying the potentially idiomatic expression (PIE) and determining whether it should be interpreted literally or figuratively. While LLMs are reshaping the landscape of natural language processing (NLP) by empowering machines to detect, generate, and translate human language, idioms are challenging for the task of NMT (Dankers et al., 2022; Avramidis et al., 2019; Constant et al., 2017). While state-of-the-art language models such as SeamlessM4T (Barrault et al., 2023), Google Neural Machine Translation (Wu et al., 2016), and OpenAI's GPT models (OpenAI et al., 2024) provide adequate results of translating to target language, their translation output of sentences with PIE are not perfect (Table 1). Some of the errors in the example include hallucinating non-existing idioms in the target language and literal translation of figurative phrases.

In this work, we introduce Transidiomation, a prompting framework that improves the translation quality of sentences that are embedded with idioms by adding a process where LLMs specifically identify PIEs before translating. We quantitatively and qualitatively analyze the effectiveness of the Transidiomation framework in translating idiomatic expressions.

The key contributions of this work are as follows:

- **We introduce a simple yet effective prompt that improves the quality of translation on sentences with idioms.** We show that LLMs can produce better translation by adding a step to identify the meaning of the idiomatic expression.

- **We expand the current scope of research in machine translation.** Research in machine translation aims to improve translation quality in formal written English. Prior research in idiom translation deals with identifying idiom-to-idiom equivalence rather than focusing on translating the whole sentence embedded with idioms. Our work focuses on improving translation in sentences embedded with idioms.

2 Related Works

The following sections describe prior works conducted in applying natural language processing (NLP) techniques to figurative language identification (Section 2.1) and idiom translation (Section 2.2). We also introduce previous research on machine translation using few-shot prompting (Sec-

English sentences with PIE	GPT 3.5	Google Translate	SeamlessM4T
It was raining cats and dogs last weekend.	지난 주말에는 개구리가 비 내리는 것 같았어요. It was like frogs were raining last weekend. (hallucination)	지난 주말에는 고양이와 개에게 비가 내렸습니다. Cats and dogs were rained on last weekend. (incorrect)	지난 주말에는 고양이와 개가 비를 내고 있었다. Not translatable
Getting that done was a piece of cake.	그것을 해내는 것은 쉬운 일이었다. It was easy to get that done.	그 일을 끝내는 건 아주 쉬운 일이었어 It was easy to get that done.	그 일을 처리하는 것은 쉬운 일이었습니다. It was easy to get that done.
Teaching is my bread and butter.	가르치는 것이 내 소득원이다. Teaching is my source of income.	가르치는 것이 나의 빵과 버터이다. Teaching is my bread and butter. (literal translation)	가르치는 게 내 과 버터야. Not translatable

Table 1: Examples of various NMT models translating English sentences with PIE to Korean. Text that have red Xs are incorrect translations.

tion 2.3) and chain-of-thought prompting (Section 2.4).

2.1 Figurative Language Identification

The identification of figurative language (i.e., idiom) has been a challenge due to its non-compositionality and ambiguity. However, it is generally assumed that idioms are semantically less compatible compared with a phrase with literal meaning, which provides a clue that can be utilized by idiom identification models; in other words, the vector representations of an idiom and its context are more distant than that of a phrase with a literal meaning. Therefore, as the first step of idiom identification, state-of-the-art (SoTA) models, such as Zeng and Bhat (2021) and Tedeschi et al. (2022), incorporate both contextualized and static word embeddings to evaluate the semantic compatibility of PIEs. While the static embeddings maintain the compositional meaning of the token, the contextualized embeddings encode the semantic content under a certain context, tackling the non-compositionality. Zeng and Bhat (2021) also introduced an attention phase including two attention flow layers: the first one fuses static word embeddings and part-of-speech (POS) information to capture the idiosyncratic syntactic structures of idioms and the other, and the second one evaluates the semantic compatibility between static and contextualized embeddings. Adopting a similar approach, Tedeschi et al. (2022) generalized this structure to the idiom identification of 10 languages. Both studies applied the Bidirectional Long Short-term Memory (BiLSTM) network to enrich their representation and optimize their prediction.

2.2 Idiom Translation

The translation of idiomatic expressions has been a challenging task because the meaning of the expression cannot be deduced from the composition of the expression. Studies have shown the drawbacks of NMT models and statistical machine translation (SMT) models in translating idiomatic expressions. These models tend to translate the idiomatic expressions leading to erroneous and ambiguous translation outputs (Baziotis et al., 2022). To solve this problem Baziotis et al. (2022) propose a novel metric for automatically evaluating the literal translation errors (LitTER), which also complements word-alignment-based metrics into a unified targeted evaluation framework. Baziotis et al. (2022) states that exploitation of monolingual pre-training yields strong targeted gains during evaluation.

Another class of model, transformer-based models, has gained significant prominence due to its architecture. Transformer models generate a contextual embedding representation for any word (i.e., token) in the training corpus, which is based on both the context and the token as well (Santing et al., 2022). Multiple approaches were proposed in Santing et al. (2022) to fine-tune a pre-trained Text-to-Text Transfer Transformer (T5) model on performing idiom translation from English to German. However, only the first approach promised a direct idiom-to-idiom translation from English to German. Four tasks were directly involved with this approach: identification of idiomatic expressions within a sentence using a transformer model, sense disambiguation (determining whether an idiom is used literally or idiomatically in a sentence) using a transformer model, and finally translating the idiom using the T5 model. Various T5 translation models were fine-tuned for sequence-to-sequence learning tasks, which required prepending input sequences with the prefix "translate English to German:" (predefined) or "translate English to German with idiom:" (custom) (Santing et al., 2022). The best-performing model was based on continuing to fine-tune the T5 model for the predefined English-to-German task without the idiomatic expression specified in the input sequence. This resulted in significant p-values of 0.010 and 0.016 for BLEU and COMET, proving its performance improvement compared to the baseline model.

2.3 Few-shot Prompting

The best approach to utilizing LLMs for machine translation is still an unresolved and ongoing area of research and exploration. Difficulties can arise during the implementation of zero-shot translation when there is a lack of direct mappings between the source and target languages in the training data (Zan et al., 2024). However, in the framework of in-context learning (i.e., few-shot prompting), LLMs have demonstrated the ability to produce more relatable translations by being presented with multiple priming examples where they comprise a source sentence paired with its translated target sentence (Sia and Duh, 2022).

Moslem et al. (2023) observed that the translation quality achieved through few-shot in-context learning can exceed that of robust encoder-decoder machine translation systems, particularly for languages with ample linguistic resources. Domain-specific translation especially benefits most from few-shot methods (Agrawal et al., 2022; Moslem et al., 2023; Zhang et al., 2023) as LLMs adapt their output to adhere to the pattern and style used in previously approved outputs.

2.4 Chain-of-Thought Prompting

Chain-of-thought (CoT) prompting (Kojima et al., 2022) is a form of prompt engineering that emphasizes the step-by-step process of solving problems. By inputting logical problem-solving examples in addition to the main question, the LLM can arrive at the final answer by utilizing the logic of the prompt. For example, Raunak et al. (2023) requests GPT-4 (OpenAI et al., 2024) to enhance translations by incorporating the edit suggestions generated by this strategy. Feng et al. (2024) devised a structured self-correcting translation framework TER (Translate, Estimate, and Refine) based on LLMs and demonstrated that diverse estimation strategies result in different effects on AI feedback, consequently influencing the efficacy of the ultimate corrections. Another use of the CoT model was its implementation of cross-lingual transfer for low-resource languages. Translation-Assisted Cross-Linguality (TaCo) approach employs translations in a CoT process to fine-tune instruction-tune LLMs for new languages via a curriculum-learning procedure (Upadhayay and Behzadan, 2023). Researchers managed to double the performance in contrast to instruction tuning alone, demonstrating the effectiveness of CoT prompting.

3 Broader Impact

Machine translation revolutionized global communication by enabling the seamless conversion of text across different languages. However, cultural context and nuanced translation are still lacking even in the SoTA models. Idioms, which are culturally specific phrases whose meanings cannot be deduced from their words, pose a unique challenge in translation. Thus, while automation streamlines the translation process, it is essential to maintain a high level of concentration and expertise to accurately render idioms in the target language. This emphasis on correctly translating idioms would enhance cross-cultural understanding and foster more authentic interactions across linguistic boundaries.

4 Approach

In this work[1], we designed prompts that utilize prompting techniques like CoT and few-shot prompting to assist in translating sentences embedded with idioms. We then compared the quality of translation generated by our prompts and baseline approaches using both automatic and human evaluations.

4.1 Hypothesis

Current NMT and naive translation produced by large language models (LLMs) struggle with figurative language translation and often make mistakes like translating the source sentence into its literal meaning. However, recently developed LLMs are trained with extreme amounts of human-written text, which led to their impressive knowledge of idioms and their figurative meanings. Though different languages have drastically different idioms

[1] All codes are publicly available under `https://github.com/NLPitch/transidiomation`. The framework can be used from any source language to any target language when the correct data are provided.

depending on their respective cultures, the ideas that are being expressed by the idioms are often universal. It is thus possible that LLMs can capture the figurative meaning of sentences embedded with idioms and find idioms with the same meaning in different languages. Prompting an LLM in a specific manner may guide the model to leverage its knowledge of idioms in multiple languages and integrate them into the translation task, thus improving its translation quality. Therefore, we hypothesize that prompting an LLM to use its idiom knowledge in translation will guide the model to outperform the naive zero-shot prompted LLM as well as the widely used NMT software.

4.2 Prompt Design

We designed a prompting pipeline that integrated CoT prompting (Wei et al., 2022) and few-shot prompting (Brown et al., 2020). Inspired by CoT, we designed a step-by-step instruction (head prompt) as follows:

> “Translate the sentence into target language by following these steps:
>
> Step1. Identify the idiom.
>
> Step2. Find an idiom with the same meaning in the target language. If there is no equivalent idiom, give the figurative meaning of it.
>
> Step3. Include response from Step2 to translate the sentence."

In addition to the instruction, we also adopted the few-shot prompting technique (Brown et al., 2020) by providing examples following the three steps delineated in the prompt. An example for English-to-Spanish translation is as follows:

> User: When I learn Spanish at the Spanish Academy exams are a piece of cake.
>
> Assistant:
>
> Step 1. The idiom in the sentence is "a piece of cake".
>
> Step 2. The idiom translates to "pan comido".
>
> Step3. Full sentence translates to "Cuando aprendo espanol en ’The Spanish Academy’ los examenes son pan comido."

After experimenting with different numbers of "shots" (1 to 5 shots), we found that, surprisingly, two-shot instead of five-shot prompting performed the best. It is possible that the examples provided in the five-shot were somewhat biased thus resulting in overfitting and unnatural translations. We therefore used two-shot prompting in our approach for the experiment.

4.3 Experiment Setup

In the experiment, we aimed to compare the translation quality of these three conditions:

> Condition 1: NMT (Google Translate)
>
> Condition 2: Zero-shot naive prompting of LLM (GPT 3.5 Turbo)
>
> Condition 3: Two-shot Transidiomation prompting of LLM (GPT 3.5 Turbo)

By comparing our proposed approach (Condition 3) with the baseline approaches (Condition 1 and Condition 2), we would be able to see if our proposed approach can significantly improve the quality of translation. The following sections will introduce the datasets, models, and evaluation metrics used in the experiment.

4.3.1 Datasets

To test the effectiveness of our prompting pipeline across languages with ground-truth human translations, we used a Korean-English Idiom-in-Sentences dataSet (KISS) (Choi, 2020) and a Spanish-English idiom dataset extracted from a language learning website (SpanishDict.com). Both datasets contain sentences embedded with idioms and the ground-truth human translation of these sentences. The sentences embedded with idioms were translated in the experiment setup of the three conditions. The model outputs were then evaluated with the ground-truth translation of the sentences.

4.3.2 Models

In the experiment condition, we applied our prompting pipeline to GPT 3.5 Turbo. For one of the baseline conditions, we used Google Translate, a sophisticated NMT software, to represent the current performance of machine translation. For the other baseline condition, we applied a naive prompt to GPT 3.5 Turbo: "Translate {sentence}." The output of this condition can reflect the baseline performance of this LLM without the proposed prompting pipeline.

4.4 Evaluation

Through our initial qualitative observation, we found that the output of our proposed approach did improve the quality of translation. 20 randomly selected examples of Spanish translations are shown in Appendix A, and 20 randomly selected examples of Korean translations are shown in Appendix B. We then adopted automatic evaluation and human evaluation to systematically evaluate the improvement.

4.4.1 Automatic Evaluation

In automatic evaluation, we evaluated the fluency and adequateness of the translated sentences with several different metrics. Fluency was evaluated by Word Error Rate (WER) (Morris et al., 2004), which measures how natural and grammatically correct the sentence is. Adequateness was evaluated by BLEU score (Papineni et al., 2002) and ROUGE F1 score (Lin, 2004), as well as the cosine similarity between the embeddings (obtained from OpenAI models) of the generated text and the reference text, all of which reflect the semantic similarity between the generated text and the reference text. Additionally, we used the BLEURT score (Sellam et al., 2020) to evaluate both fluency and adequateness. Although these metrics failed to show a significant difference between our approach and the baseline approaches, we did observe an improvement by directly comparing the translation generated by the proposed approach and the baselines. Given that the automatic evaluation metrics could not capture the improvement, we turned to human evaluation as a gold-standard evaluation metric.

4.4.2 Human Evaluation

We performed a human evaluation for the Korean-English translation, as two of the authors are speakers of both Korean and English. The two raters were asked to rank the translated sentences generated in the three conditions as "best", "middle", and "worst". Before evaluation, we randomly shuffled the order of the three conditions for each sentence so that the raters were blind to the conditions these sentences were in. We then calculated the average ranking of each condition to reflect the human preference for the translation generated in the three conditions.

4.5 Results

Our evaluation results for BLEURT, ROUGE F1, and WER indicate that our proposed method has no evidence of statistically significant difference when compared to the baseline methods we examined [2]. Our proposed method exhibits a broader spread of scores in both datasets, implying variability in performance across all evaluation metrics shown in Figure 1, Figure 2, and Figure 3. Despite our expectations, Google Translation displayed slightly better to equal performance compared to **Transidiomation** across all evaluation metrics and languages. The Korean dataset generated more outliers across all evaluation metrics compared to the Spanish dataset. These outliers suggest that the method's performance can also lead to translations that perform poorly than usual. We can infer that these outliers may cause a skewed distribution, resulting in a high error rate and the potential for inaccurate and misleading results.

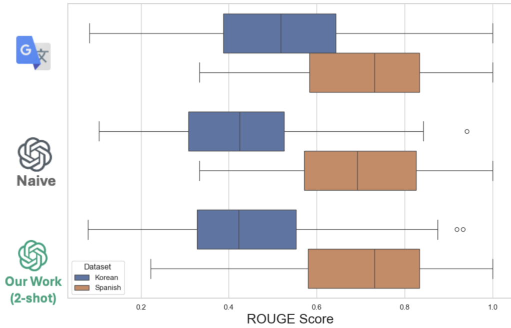

Figure 1: Distribution of ROUGE F1 score by prompting methods

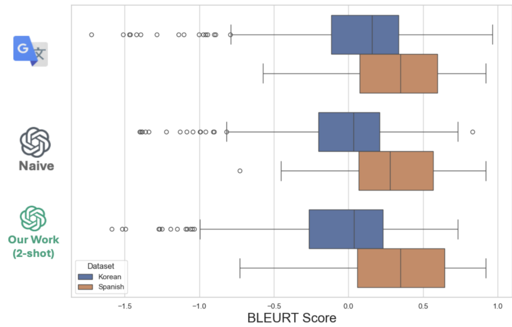

Figure 2: Distribution of BLEURT score by prompting methods

For human evaluation, we have computed the average ranking of the three translation approaches based on two data conditions: (1) All, where the average ranking from all 200 source texts was computed and (2) With Equivalent English Idioms,

[2]For results on other metrics such as BLEU, METEOR, and OpenAI rating, see `https://github.com/NLPitch/transidiomation/tree/main/output/stats`.

Data	Google Translate	Naive Prompting	Transidiomation
All	2.191	1.965	**1.844**
With Equivalent English Idioms	2.200	1.950	**1.850**

Table 2: Average ranking obtained from human evaluation on 200 Korean sentences translated into English. A lower value signifies that the translated sentence was preferred by the evaluators.

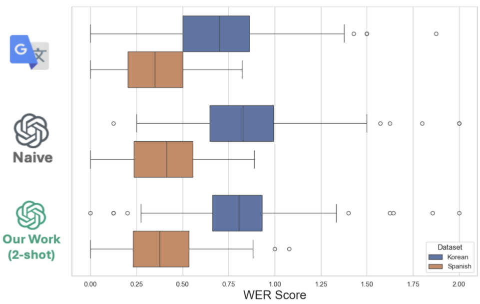

Figure 3: Distribution of WER score by prompting methods

where the average ranking of sentences that have equivalent English idioms was computed. Table 2 displays the result from human evaluation. From the table we can see that Transidiomation has the lowest rank among all three approaches, both in the All condition and in the With Equivalent English Idioms condition. This result suggests that our approach generates translation that are more preferred by human raters, which supports our hypothesis.

4.6 Error Analysis

We identified three main sources of erroneous outputs from GPT 3.5 throughout the experimental process: no output generation, incomplete output generation, and inconsistent generation requiring rigorous data processing (Figure 4). Additionally, we identified that about 6.7% of the BLEURT scores for the Korean dataset fell into the outlier category indicating possible cause of distortion in statistical outcomes. We were able to resolve erroneous outputs from GPT 3.5 through rigorous data processing and reprompting. However, we have not yet implemented any solutions for resolving cases of outliers in the BLEURT evaluation metric for the Korean dataset. We have discussed possible indications of a lack of fluency in the Korean language across all prompting methods. We believe that this issue could be resolved by excluding the outlier data points and including additional datasets to introduce variability.

5 Conclusion

We introduce **Transidiomation**, a novel prompting framework that improves the translation quality of sentences embedded idioms by formulating the translation task into a series of tasks. Evaluation conducted by human evaluators has shown a preference towards translation generated by **Transidiomation** over those generated by naive LLMs and widely used translation tools. Our research has demonstrated the possibility of further improving machine translation by taking culture and language-specific phrases, such as idioms, into consideration.

Limitations

We identified several limitations that apply to our proposed method and evaluation metrics. During LLM prompting, idioms identified in the translated sentence were often not used in the final translation step. We have also found cases of multiple idiom equivalences under different contexts, which were not considered during data collection. To address this issue, we have conducted a study on a few English-to-Korean idioms. Both the Spanish and Korean datasets only had one reference text for each candidate text, limiting the robustness of the evaluation results.

Though the prompting framework is applicable regardless of the language, the study has been conducted with English, Korean, and Spanish texts. Our work can be extended further by evaluating the performance of the prompting framework on text across a diverse language. Furthermore, additional studies can be conducted on whether the proposed method is also effective on smaller-sized LLMs, such as Mistral 7B[3].

Ethics Statement

Idiom translation requires extensive awareness of cultural implications and sensitivity. Translation processes are subjected to various ethical concerns due to a wide range of factors that could deviate

[3]An initial study has been conducted in this direction, and results on a few sentences are available in Appendix C.

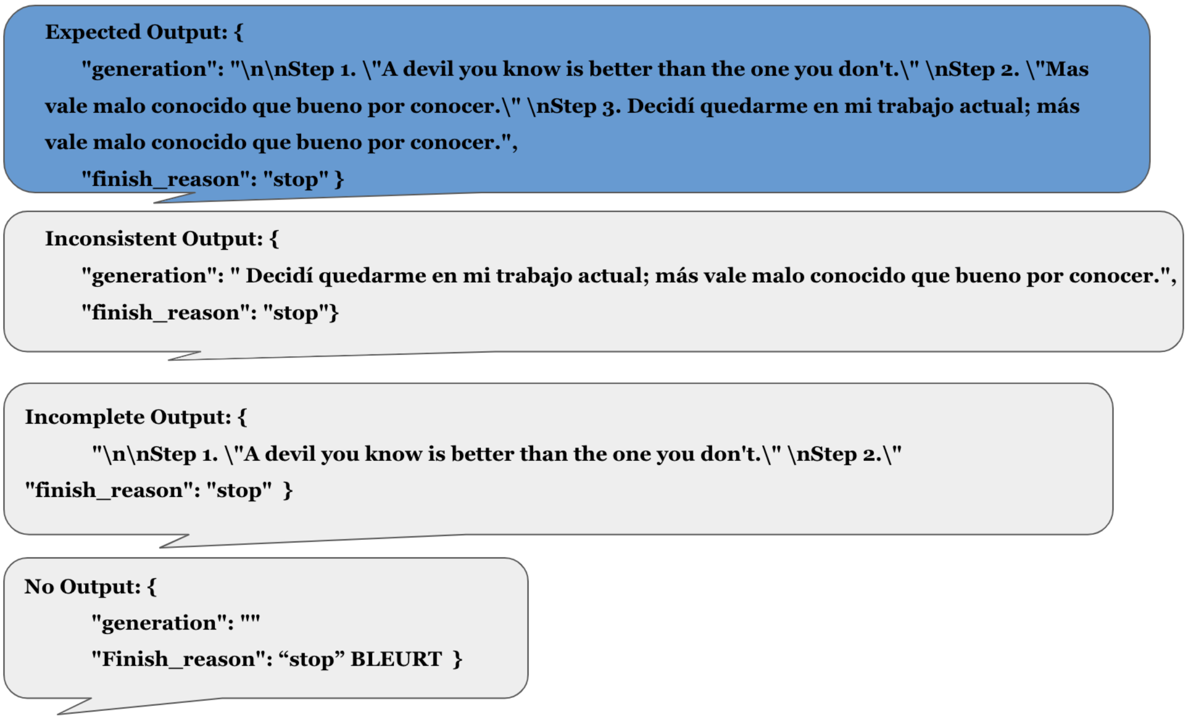

Figure 4: Examples of GPT 3.5 error outputs

from the intended meaning for personal or ideological reasons. Our work relies on machine translation systems trained on data that often contains biases and lacks sensitivity to nuances, which could result in potentially harmful or inappropriate translations. Loss of cultural authenticity expressed through unique semantic properties of a language during translation also plays a key role in ethical translation processes. Overall, careful analysis and attention to detail are necessary to maintain a balance in preserving the intent, accuracy, and cultural sensitivity in machine translation processes.

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2022. In-context examples selection for machine translation. *arXiv preprint arXiv:2212.02437*.

Eleftherios Avramidis, Vivien Macketanz, Ursula Strohriegel, and Hans Uszkoreit. 2019. Linguistic evaluation of german-english machine translation using a test suite. *arXiv preprint arXiv:1910.07457*.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. 2023. Seamlessm4t-massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

Christos Baziotis, Prashant Mathur, and Eva Hasler. 2022. Automatic evaluation and analysis of idioms in neural machine translation.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Min-Joo Choi. 2020. A data set and evaluating method for korean-english idiom machine translation. *Korean Information Science Society Conference*.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke Van Der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

Verna Dankers, Christopher G Lucas, and Ivan Titov. 2022. Can transformer be too compositional? analysing idiom processing in neural machine translation. *arXiv preprint arXiv:2205.15301*.

Zhaopeng Feng, Yan Zhang, Hao Li, Wenqiang Liu, Jun Lang, Yang Feng, Jian Wu, and Zuozhu Liu. 2024. Improving llm-based machine translation with systematic self-correction. *arXiv preprint arXiv:2402.16379*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Andrew Morris, Viktoria Maier, and Phil Green. 2004. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition.

Yasmin Moslem, Rejwanul Haque, John D Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. *arXiv preprint arXiv:2301.13294*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Vikas Raunak, Amr Sharaf, Hany Hassan Awadallah, and Arul Menezes. 2023. Leveraging gpt-4 for automatic translation post-editing. *arXiv preprint arXiv:2305.14878*.

Lukas Santing, Ryan Sijstermans, Giacomo Anerdi, Pedro Jeuris, Marijn ten Thij, and Riza Batista-Navarro. 2022. Food for thought: How can we exploit contextual embeddings in the translation of idiomatic expressions? In *Proceedings of the 3rd Workshop* *on Figurative Language Processing (FLP)*, pages 100–110, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: learning robust metrics for text generation. *CoRR*, abs/2004.04696.

Suzanna Sia and Kevin Duh. 2022. Prefix embeddings for in-context machine translation. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 45–57.

Simone Tedeschi, Federico Martelli, and Roberto Navigli. 2022. Id10m: Idiom identification in 10 languages. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2715–2726.

Bibek Upadhayay and Vahid Behzadan. 2023. Taco: Enhancing cross-lingual transfer for low-resource languages in llms through translation-assisted chain-of-thought processes. *arXiv preprint arXiv:2311.10797*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Changtong Zan, Liang Ding, Li Shen, Yibing Zhen, Weifeng Liu, and Dacheng Tao. 2024. Building accurate translation-tailored llms with language aware instruction tuning. *arXiv preprint arXiv:2403.14399*.

Ziheng Zeng and Suma Bhat. 2021. Idiomatic expression identification using semantic compatibility. *Transactions of the Association for Computational Linguistics*, 9:1546–1562.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning*, pages 41092–41110. PMLR.

A Spanish Translation Examples

The following table (Table 3) displays 10 randomly selected Spanish translation outputs from the three translation approaches (i.e., Google Translate, Naive GPT 3.5, Transidiomation)[4].

	Sentence
Original	It is up to Castro and Roque to show what they are made of and get in the bandwagon.
Reference	Corresponde a Castro y Roque demostrar de qué madera están hechos y subirse al carro.
Google Translate	A Castro y Roque les toca demostrar de qué están hechos y subirse al carro.
Naive GPT	Depende de Castro y Roque mostrar de qué están hechos y subirse al carro.
Transidiomation	Depende de Castro y Roque demostrar de qué están hechos y subirse al carro.
Original	It's fascinating that the mechanic has his car broken down. As it is said, the shoemaker's son always goes barefoot.
Reference	Es curioso que el mecánico tenga su coche averiado. Como se dice, en casa del herrero, cuchillo de palo.
Google Translate	Es fascinante que al mecánico se le haya averiado el coche. Como se dice, el hijo del zapatero siempre anda descalzo.
Naive GPT	Es fascinante que el mecánico tenga su auto averiado. Como se dice, el hijo del zapatero siempre va descalzo.
Transidiomation	Es fascinante que el mecánico tenga su carro averiado. Como se dice, el hijo del zapatero siempre anda descalzo.
Original	I decided to stay in my current job; a devil you know is better than the one you don't.
Reference	Decidí quedarme en mi trabajo actual; más vale lo malo conocido que lo bueno por conocer.
Google Translate	Decidí quedarme en mi trabajo actual; Un diablo que conoces es mejor que el que no conoces.
Naive GPT	Decidí quedarme en mi trabajo actual; un demonio que conoces es mejor que uno que no conoces.
Transidiomation	Decidí quedarme en mi trabajo actual; más vale malo conocido que bueno por conocer.
Original	The restaurant we went to for dinner in the boonies, it took us almost an hour to get there.
Reference	El restaurante al que fuimos a cenar estaba en el quinto pino, nos llevó casi una hora llegar.
Google Translate	El restaurante al que fuimos a cenar en los boonies nos llevó casi una hora llegar allí.
Naive GPT	El restaurante al que fuimos a cenar en el campo nos tomó casi una hora para llegar allí.
Transidiomation	El restaurante al que fuimos a cenar en el quinto infierno, nos tomó casi una hora llegar allí.
Original	Carlos's old car finally kicked the bucket in the middle of the trip.
Reference	El viejo coche de Carlos finalmente estiró la pata en medio del viaje.
Google Translate	El viejo auto de Carlos finalmente murió en medio del viaje.
Naive GPT	El viejo coche de Carlos finalmente se murió a mitad del viaje.
	Continued on next page

[4]Full table is available here: https://github.com/NLPitch/transidiomation/blob/main/output/compiled_Spanish.xlsx

	Sentence
Transidiomation	El viejo coche de Carlos finalmente murió en medio del viaje.
Original	After talking to my friend about my problems, I felt better; it's true that two in distress makes sorry less.
Reference	Después de hablar con mi amiga sobre mis problemas, me sentí mejor; es cierto que las penas compartidas saben a menos.
Google Translate	Después de hablar con mi amiga sobre mis problemas, me sentí mejor; es cierto que dos en apuros hacen que se arrepienta menos.
Naive GPT	Después de hablar con mi amigo sobre mis problemas, me sentí mejor; es cierto que dos en aprietos hacen que la pena sea menor.
Transidiomation	Después de hablar con mi amigo sobre mis problemas, me sentí mejor; es cierto que dos en apuros hacen que lo siento menos.
Original	When she forgot about our anniversary, it was the straw that broke the camel's back and I decided to end the relationship.
Reference	Cuando olvidó nuestro aniversario, fue la gota que colmó el vaso y decidí terminar la relación.
Google Translate	Cuando ella se olvidó de nuestro aniversario, fue la gota que colmó el vaso y decidí terminar la relación.
Naive GPT	Cuando ella se olvidó de nuestro aniversario, fue la gota que colmó el vaso y decidí terminar la relación.
Transidiomation	Cuando ella se olvidó de nuestro aniversario, fue la gota que colmó el vaso y decidí terminar la relación.
Original	I always like to speak plainly, to call a spade a spade, so we all know where we stand.
Reference	Siempre me gusta hablar claro, al pan, pan y al vino, vino; así todos sabemos a qué atenernos.
Google Translate	Siempre me gusta hablar claro, llamar las cosas por su nombre, para que todos sepamos cuál es nuestra posición.
Naive GPT	Siempre me gusta hablar con franqueza, llamar a las cosas por su nombre, para que todos sepamos en qué posición estamos.
Transidiomation	Siempre me gusta hablar con franqueza, llamar a las cosas por su nombre, para que todos sepamos dónde estamos parados.
Original	Even though I got fired, I found a better job, so every cloud has a silver lining.
Reference	Aunque me despidieron, encontré un trabajo mejor, así que no hay mal que por bien no venga.
Google Translate	Aunque me despidieron, encontré un trabajo mejor, así que cada nube tiene un lado positivo.
Naive GPT	esde que perdió su trabajo, parece que no puede tener un respiro; ya sabes, cuando llueve, cae a cántaros.
Transidiomation	Aunque me despidieron, encontré un trabajo mejor, así que todo tiene su lado positivo.
Original	Look at those corrupt politicians together; birds of a feather flock together.
Reference	Mira a esos políticos corruptos juntos; dios los cría y ellos se juntan.
Google Translate	Miren a esos políticos corruptos juntos; dios los cría y ellos se juntan.
Naive GPT	Mira a esos políticos corruptos juntos; pájaros de la misma pluma vuelan juntos.
Transidiomation	Mira a esos tramposos jugando cartas juntos; aves de la misma pluma vuelan juntas.
Original	Don't worry about tomorrow's exam, it's going to be a piece of cake.

Continued on next page

	Sentence
Reference	No te preocupes por el examen de mañana, va a ser pan comido.
Google Translate	No te preocupes por el examen de mañana, será pan comido.
Naive GPT	Estaba tan feliz como una almeja cuando recibió la noticia de que había pasado el examen.
Transidiomation	No te preocupes por el examen de mañana, va a ser pan comido.

Table 3: 10 Spanish Translation Examples

B Korean Translation Examples

The following table (Table 4) displays 20 randomly selected Korean translation outputs from the three translation approaches (i.e., Google Translate, Naive GPT 3.5, Transidiomation)[5]. The idiomatic expression in the source text and target text are underlined, if applicable.

	Sentence
Original	가재는 게 편이라고 그 사람들은 자기네 동네 사람을 도와줄 거야.
Reference	Birds of a feather flock together. They're going to help the people in their town.
Google Translate	Those people said that crabs are more convenient, so they will help their fellow villagers.
Naive GPT	They say crawfish are good and they will help people in their neighborhood.
Transidiomation	The locals will help each other as birds of a feather flock together.
Original	고생 끝에 낙이 온다는 말처럼, 그는 온갖 고생을 했지만, 마침내 대단한 것을 이뤄냈다.
Reference	Just like every cloud has a silver lining, he jumped through hoops and finally achieved something great.
Google Translate	Just like the saying "there is light at the end of the tunnel", he went through all sorts of hardships but in the end, he achieved something great.
Naive GPT	As the saying goes, happiness comes after hard work, and although he went through all kinds of hardships, he finally achieved something great.
Transidiomation	Just like the saying, every cloud has a silver lining, he went through all sorts of hardships, but finally achieved something great.
Original	저는 귀가 얇아서 다른 사람 말을 잘 믿어요.
Reference	I am gullible, so I easily believe what others say.
Google Translate	I have thin ears, so I tend to believe what other people say.
Naive GPT	I have small ears, so I tend to trust what other people say.
Transidiomation	I am gullible and easily believe what others say.
Original	금강산도 식후경이라고, 배고픈 상태에서는 아무것도 집중할 수 없어
Reference	Pudding rather than praise. You can't focus on anything when you're hungry.
Google Translate	Even Mount Geumgang looks like a painting after a meal, I cannot concentrate on anything in a hungry state.
Naive GPT	Even Mt. Geumgang is said to be after a meal, and you can't concentrate on anything when you're hungry.
Transidiomation	Geumgangsan Mountain is also a postprandial view, you can't concentrate on anything when you're hungry.
Original	앤은 언제나 지각하는 주제에 내가 지각하면 뻔뻔스럽게 여러 사람에게 불고 다니는 거야. 정말이지 똥 묻은 개가 겨 묻은 개를 나무란다더니.
	Continued on next page

[5] Full table is available here: https://github.com/NLPitch/transidiomation/blob/main/output/compiled_Korean.xlsx

	Sentence
Reference	Ann is always late, but she was rude enough to tell everyone when I was late. Now, that's the pot calling the kettle black!
Google Translate	Anne is always late, and when I am late, she shamelessly blames and complains to many people. It's like a dog with poop on its body blaming a dog with mud on its body.
Naive GPT	Anne is always late, and when I'm late, she shamelessly lashes out at several people. It's true that a dog with poop scolds a dog with bran on it.
Transidiomation	Ann is always late but has the nerve to criticize others when they are late. It's like the pot calling the kettle black.
Original	나는 그 책을 마음에 두고 있었습니다.
Reference	I had the book in mind.
Google Translate	I had my eye on that book.
Naive GPT	I had that book in mind.
Transidiomation	I was really taking that book to heart.
Original	중국도 결기를 보이고는 있지만 미국에 비해서는 머리가 복잡한 모습이다.
Reference	China is also showing signs of commitment, but compared to the U.S., its head looks more complicated.
Google Translate	China also shows signs of arrogance, but compared to the United States, its stance is more complicated.
Naive GPT	China is also showing signs of resolve, but it appears to be more complex than the United States.
Transidiomation	China does show its teeth sometimes, but compared to the United States, their actions are more complex.
Original	그는 2년 동안 준비한 공무원 시험에서 미역국을 먹었다.
Reference	He failed the civil service exam that he had prepared for for two years.
Google Translate	He ate seaweed soup during the civil service exam that he had been preparing for two years.
Naive GPT	He ate seaweed soup for the civil service exam, which he had prepared for for two years.
Transidiomation	He had seaweed soup in the exam for a civil servant that he prepared for two years.
Original	나는 그녀를 완전히 신뢰했는데, 믿는 도끼에 발등 찍힌 격이 되었어.
Reference	I completely trusted her, but it ended up being a case of being stabbed in the back
Google Translate	I completely trusted her, but it turned out to be a case of the axe hitting the foot of the believer.
Naive GPT	I trusted her completely, but it was like being stabbed in the foot with an axe of trust.
Transidiomation	I completely trusted her, but it turned out to be a case of being stabbed in the back by the trusted axe.
Original	많은 분들이 도와줘서 지금까지 올 수 있었다는 것을 시간이 지날수록 점점 더 피부로 느끼고 있다.
Reference	As time goes by, I can feel in my bones that many people have helped me to come to this point.
Google Translate	As time passes, I increasingly feel that I have been able to come this far because many people have helped me.
Naive GPT	As time goes by, I feel more and more that I was able to come this far thanks to the help of many people.
	Continued on next page

	Sentence
Transidiomation	The more time goes by, the more I feel in my bones that I could get this far thanks to the help of many people.
Original	그 만남이 성사되면 내 손에 장을 지지겠다.
Reference	Pigs will fly when the meeting takes place.
Google Translate	"I will take care of the arrangements if that meeting happens."
Naive GPT	If that meeting takes place, I will hold the cabinet in my hands.
Transidiomation	When pigs fly, I'll hold the reins in my hands if that meeting actually happens.
Original	한라산 등반하는 건 식은 죽 먹기야.
Reference	Climbing Hallasan mountain is a piece of cake.
Google Translate	Climbing Hallasan Mountain is like a piece of cake.
Naive GPT	Climbing Hallasan Mountain is a piece of cake.
Transidiomation	Climbing Hallasan is a piece of cake.
Original	나는 울며 겨자 먹기로 그 요구에 응했다.
Reference	I had to bite the bullet and backed down on the demand.
Google Translate	I reluctantly agreed to his demand.
Naive GPT	I responded to the request by crying and eating mustard.
Transidiomation	I reluctantly complied with the demand.
Original	어제 술을 너무 많이마셔서 필름이 끊겼어.
Reference	I drank too much yesterday and I got blacked out.
Google Translate	I drank too much alcohol yesterday and blacked out.
Naive GPT	I drank too much yesterday so the film cut out.
Transidiomation	I drank too much yesterday so I blacked out.
Original	호랑이도 제말하면 나타난다더니만, 이제 오시는구만.
Reference	Speak of the devil, here he comes now.
Google Translate	They say that even tigers appear when you talk about them, and look who's here now.
Naive GPT	They say a tiger will appear when you tell it to, but it's coming now.
Transidiomation	Speak of the devil, and here he comes now.

Table 4: 20 Korean Translation Examples

C Results from Prompting on Mistral 7B

We have conducted an initial study to measure the effectiveness of our work on a smaller-sized model. We have used Mistral 7B and compared the output from Transidiomation prompting with naive prompting (i.e., Translate *source_sentence* to *target_language*).

As opposed to the case when using GPT 3.5-Turbo, using the Transidiomation framework on Mistral 7B fails to use equivalent idiomatic expressions in the target language. Regardless, Transidiomation prompting generates a better translation output when compared to naive prompting in that it uses figurative meaning instead of a literal translation. This pattern is visible in the examples shown in Table 5.

	Sentence
Original	가재는 게 편이라고 그 사람들은 자기네 동네 사람을 도와줄 거야.
Reference	Birds of a feather flock together. They're going to help the people in their town.
Naive Mistral	The clam is more comfortable being handled by people from their own neighborhood.
Transidiomation	They believe that their fellow townspeople will come to their aid when needed.
Original	고생 끝에 낙이 온다는 말처럼, 그는 온갖 고생을 했지만, 마침내 대단한 것을 이뤄냈다.
	Continued on next page

	Sentence
Reference	Just like every cloud has a silver lining, he jumped through hoops and finally achieved something great.
Naive Mistral	After enduring countless hardships, he finally achieved something remarkable.
Transidiomation	Despite going through numerous hardships, he finally accomplished something great, just as one might say "the rain will come someday."
Original	앤은 언제나 지각하는 주제에 내가 지각하면 뻔뻔스럽게 여러 사람에게 불고 다니는 거야. 정말이지 똥 묻은 개가 겨 묻은 개를 나무란다더니.
Reference	Ann is always late, but she was rude enough to tell everyone when I was late. Now, that's the pot calling the kettle black!
Naive Mistral	Anne always pulls out sensitive topics that make me cringe and then gleefully shares them with multiple people. It's really like a dog with sht on its paws pointing it out to another dog.
Transidiomation	And you, always late, become the center of attention, boasting about being just as bad as others, acting like two scoundrels!
Original	그는 2년 동안 준비한 공무원 시험에서 미역국을 먹었다.
Reference	He failed the civil service exam that he had prepared for for two years.
Google Translate	He ate seaweed soup during the civil service exam that he had been preparing for two years.
Naive Mistral	He had maechoonggi (meal made of seaweed) for two years while preparing for the civil service examination.
Transidiomation	He ate miyeokguk, a traditional Korean dish made of seaweed soup, during the past two years while preparing for the civil service exam.
Original	나는 그녀를 완전히 신뢰했는데, 믿는 도끼에 발등 찍힌 격이 되었어.
Reference	I completely trusted her, but it ended up being a case of being stabbed in the back
Naive Mistral	I trusted her completely, but now I've stepped on the trusty knife's blade.
Transidiomation	I trusted her completely, but I've been let down.
Original	많은 분들이 도와줘서 지금까지 올 수 있었다는 것을 시간이 지날수록 점점 더 피부로 느끼고 있다.
Reference	As time goes by, I can feel in my bones that many people have helped me to come to this point.
Naive Mistral	The more and more people have helped me, I have been able to come this far, and I have been feeling it more and more deeply on my skin as time goes by.
Transidiomation	I've been coming to deeply realize that it's only because of the help of many people.
Original	한라산 등반하는 건 식은 죽 먹기야.
Reference	Climbing Hallasan mountain is a piece of cake.
Naive Mistral	Climbing Hanloganseong Fortress is just like eating plain rice.
Transidiomation	Hiking Hanbansan is a piece of cake.
Original	나는 울며 겨자 먹기로 그 요구에 응했다.
Reference	I had to bite the bullet and backed down on the demand.
Naive Mistral	I responded by eating cucumber while crying.
Transidiomation	I went ahead and did it, despite crying over spilt milk.
Original	호랑이도 제말하면 나타난다더니만, 이제 오시는구만.
Reference	Speak of the devil, here he comes now.
Naive Mistral	The tiger appears if you call it out, but now it is coming towards us.
Transidiomation	And so, the tiger appears, but now you're coming too.

Table 5: Korean Translation on Mistral Example