sciendo
DOI 10.2478/jazcas-2025-0023

AI AND THE TRANSLATION OF IDIOMS – CHALLENGES, SUCCESS, AND A CORPUS PERSPECTIVE

FILIP KALAŠ[1] – PAVOL LIPTÁK[2]
[1]Department of Linguistics and Translation, Faculty of Applied Languages, University of Economics and Business, Bratislava, Slovakia (ORCID: 0009-0009-1802-1700)
[2]Department of Marketing, Faculty of Commerce, University of Economics and Business, Bratislava, Slovakia (ORCID: 0000-0003-2554-9465)

KALAŠ, Filip – LIPTÁK, Pavol: AI and the Translation of Idioms – Challenges, Success, and a Corpus Perspective. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 258 – 267.

Abstract: This study explores the capabilities of artificial intelligence in translating English idioms into Slovak, with and without contextual information. Using a dataset of 100 idioms and evaluating AI-generated translations against a validated bilingual dictionary of idioms, both qualitative and statistical analyses were employed. The results show an unexpectedly high accuracy in context-free translations, while context occasionally led to deterioration. McNemar's test and a t-test confirmed a statistically significant shift in performance. The study highlights key advantages and limitations of AI, suggesting further research into reverse and cross-linguistic translation as well as employment of corpus-based methods.

Keywords: AI translation, idiom translation, phraseology, machine translation evaluation, corpus linguistics

1 INTRODUCTORY REMARKS

The interpreting profession has undergone unprecedented transformation in the last few years, due to the impact of the COVID-19 pandemic and the introduction of generative artificial intelligence (Wang and Fantinuoli 2024). The pandemic triggered the widespread adoption of remote communication technologies, and remote interpreting became a mandatory practice. The crisis notwithstanding, the continued popularity of virtual and hybrid events has sustained the demand for such tools. Dong et al. (2019) underline that interpreting industry has been revolutionized by NLP in facilitating automatic text-based operations. These advancements have reshaped the interpreting industry, inspired by innovations among interpreters, software developers, and researchers (Rodriguez 2024).

Artificial intelligence (AI) has played a foundational role and been an impactful contribution towards the field of linguistics, particularly in translation. The progression and utilization of many AI-instrumented software, as Google Translate, DeepL, OpenAI, Mistral AI, Trados Studio etc., have helped streamline translation

across languages to a significantly greater extent. This has contributed towards real-world scenarios and scholarly pursuits within scientific works (Lund 2023). These advancements have not only improved machine translation to be more efficient and accurate but have also enabled the processing of complex linguistic structures, such as idiomatic expressions. Additionally, AI-powered translation software keeps evolving, with the integration of deep learning and large language models to improve contextual understanding and guarantee translation accuracy across different languages.

According to Rodriguez (2024, p. 118), phraseology serves as a fundamental intraparameter that ensures the proper transfer of the source language and its respective specialist terminology. For this purpose, whether or not AI is able to properly process and handle domain-specific phraseology is among the major parameters for evaluating AI-based translation vis-à-vis human interpreters. Yet, AI systems use pre-trained models and large databases to identify and generate fixed phrases, technical vocabulary, or idiomatic expressions in a specific context.

1.1 Research aims

The objective of this study is to examine the impact of context on the accuracy of AI-produced idiom translation. Specifically, it investigates whether context at the sentence level leads to more accurate and idiomatically better Slovak translations of English idioms. The following working hypotheses were postulated:

H_0: The presence of context does not significantly affect the quality of AI-generated idiom translation from English to Slovak.

H_1: The presence of context improves the quality of AI-generated idiom translation from English to Slovak.

The study combines quantitative statistical analysis with qualitative linguistic evaluation to test these hypotheses.

1.2 Methodology

This quantitative-qualitative analysis was conducted on a dataset of randomly selected 100 English idioms. To evaluate the accuracy of translations produced by AI (more specifically subscripted GPT-4o), a reliable source of Slovak equivalents was required. For this purpose, the *Prekladový anglicko-slovenský frazeologický slovník* (Kvetko 2014) was selected. This bilingual dictionary contains approximately 8,000 English idioms, accompanied by around 16,000 Slovak equivalents. A key criterion for its selections was the inclusion of real-context examples for each entry.

Given the stylistic variation typical of phraseological units, only stylistically neutral idioms were included in the analysis. The stylistic characteristics are explicitly indicated in each entry in the dictionary.

From both morphological and syntactic perspectives, phraseological units can vary significantly. To ensure representativeness and to focus on forms most

commonly occurring in spoken language, the dataset was limited to idioms proper – divided equally into 50 verbal phrases and 50 nominal phrases. Sentence-like phraseological units (e.g. proverbs, sayings), similes, and binomials were excluded.

The idioms were compiled in an Excel spreadsheet structured into the following columns: *Idiom without context, Translated by AI, Human evaluation, Idiom in context, Translated by AI, Human evaluation*, and *Improvement*. The meaning of the individual column titles is largely self-explanatory. However, some clarification may be required for the column *Improvement*. In certain cases, the AI correctly interprets and renders the idiom even without context; however, this equivalent may not be preserved in the contextual translation. Conversely, the AI may provide an improved rendering when context is available, offering more accurate or idiomatically appropriate Slovak equivalent. Thus, the *Improvement* column captures not only correction of previously inaccurate translations but also enhancements in idiomatic precision or naturalness.

A custom translation prompt was used to generate AI translations of the idioms, both in isolation and within contextual sentences. The prompt goes as follows:

Translate 100 English idioms into Slovak using the following spreadsheet structure. The idioms are located in cells B2 through B101. Write the Slovak translations in the corresponding cells D2 through D101 (i.e., the translation of B2 goes into D2, and so on). After translating all idioms, proceed to the contextual sentences in cells F2 through F101. Translate these sentences into Slovak and place the results in cells G2 through G101. Once all translations are complete, prepare the updated spreadsheet for download.

Following translation, a manual (human) evaluation of each output was conducted to assess the quality of the renditions. Subsequently, statistical methods were applied to determine whether the presence of context produced a statistically significant difference in translation quality and to address the hypothesis under investigation.

As for statistical analysis, McNemar's test was selected, as it is specifically designed for paired nominal data. This test is particularly suitable for assessing changes in categorical outcomes (e.g. correct vs. incorrect) before and after an intervention – in this case, the addition of context. It is commonly used when the same subjects (idioms) are evaluated under two different conditions, which makes it ideal for detecting shifts in translation accuracy.

$$x^2 = \frac{(b-c)^2}{b+c}$$

Fig. 1. Formula for McNemar's test

In addition, a one-tailed t-test was employed to examine whether the presence of context had a statistically significant effect on the overall quality of idiom translation. This allowed for a comparison of translation scores across conditions to determine the magnitude and direction of change.

2 THEORETICAL BACKGROUND

Translation has become an integral part of daily communication, frequently used in conversations with ChatGPT. Language consists of various expressions, inter alia, idioms, whose translation poses a significant challenge for both traditional neural machine translation systems and modern large language models due to the figurative nature of idiomatic expressions. In this paper, we proceed from Sinclair's (1991, p. 172) definition of an idiom, which is a "group of two or more words which are chosen together in order to produce a specific meaning or effect in speech or writing". As Baziotis et al. (2023) point out, "literal translation errors of idioms remain a major issue in automated translation, requiring novel evaluation metrics to assess their accuracy." In contrast to a literal translation, an idiom involves far more than a perfect semantic relation; it necessitates the integration of context and cultural customization that any AI powered translation systems must incorporate.

Since the emergence of AI chatbots and related technologies, a substantial body of research has focused on evaluating the effectiveness of AI-driven tools in translating idiomatic expressions (Hamood 2024; Mughal et al. 2024; Hakami and Abomoati 2024; Abjalova and Sharipova 2024; Obeidat et al. 2024). Some scholars concentrate on semantic and grammatical aspects of the translation process, while others focus more on its computational and informatics foundations.

Recent studies demonstrate that LLMs, or more specifically ChatGPT, have achieved significant improvements on idiomatic translation over baseline NMT models (Zhu et al. 2024). Castaldo and Monti (2024) also emphasize the importance of effective prompting strategies, stating that "the quality of LLM-generated translations is highly dependent on the structure and clarity of user prompts." This suggests that user interaction is crucial in guiding LLMs to produce more precise and context-aware translations of idioms.

From a computational perspective, the inclusion of knowledge bases like IdiomKB in translation models has been found to enhance the accuracy of translation by bringing back the figurative meanings of idioms rather than their literal meanings (Li et al. 2023). Donthi et al. (2024) highlight the potential of cosine similarity scoring in bringing idiomatic expressions in languages into alignment, with the point that "such methods enable LLMs to maintain linguistic style while guaranteeing semantic fidelity". Moreover, multilingual instruction tuning has been found to induce translation ability in LLMs even for low-resource languages (Li et al. 2024).

As AI-based translation models advance, the synthesis of linguistic and informatics methodologies is critical to improving idiomatic translation. Although LLMs have shown incredible ability, their probabilistic modeling foundation ensures that some mistranslations and biases continue to arise, calling for improvements in training practices and assessment frameworks.

While there are scholars (such as Jiao 2023) who see chatbots like ChatGPT as important tools for real-time and automated translation of texts, alongside machine translation, they acknowledge the frequent errors in their output. Nevertheless, more scholars (Derner and Batistič 2023; Sison et al. 2023; Artamonova 2023) contend that ChatGPT is extremely risky, citing its capacity to create misleading translations, disseminate misinformation, and cause ethical problems in language processing.

3 RESULTS

This study examined the impact of context on the quality of AI idioms translation. A total of 100 English idioms were evaluated under two conditions: without context and with context. Each translation's correctness was assessed, and statistical tests were conducted to determine the significance of differences obtained.

Quantitative results reveal that the addition of context led to a notable decline in translation accuracy. Without context, 91% of the idioms (n=91) were correctly translated. With context, accuracy declined to 77% (n=77).

A two-sample t-test with equal variance revealed that the difference was statistically significant, $t(196) = 2.82$, $p = 0.0054$ (two-tailed). The null hypothesis, which predicted no difference in translation quality between the two conditions, was therefore rejected.

In addition, McNemar's test was conducted to assess categorical change in translation accuracy. The test revealed that 17 idioms were correct without context but incorrect with context (Tab. 1), while only 3 showed clear-cut binary improvement.

Idiom	Translation by Kvetko	Translated by AI	Idiom in context	Translated by AI
cry wolf	robiť planý poplach; predstierať nebezpečenstvo	spôsobovať planý poplach	Moody had to be here somewhere. If he weren't, Judd knew what McGreavy would think. It would be the boy who ***cried wolf***.	Moody musel byť niekde nablízku. Ak nie, Judd vedel, čo by si McGreavy pomyslel. Že je to ako s chlapcom, ktorý ***kričal "vlk"***!

an old flame	stará láska	stará láska	You were luck to run into this girl. Who is she? Some ***old flame*** of yours?	Mal si šťastie, že si stretol to dievča. Kto to je? ***Starý plameň***?

Tab. 1. Examples of mistranslations with provided context

The remaining 75 idioms retained the same accuracy. The resulting chi-square value ($x^2 = 8.45$, $p = 0.0036$) also leads to the conclusion that the introduction of context had an effect on translation quality, but not in a positive direction. A detailed breakdown of translation outcomes is presented in Fig. 2.

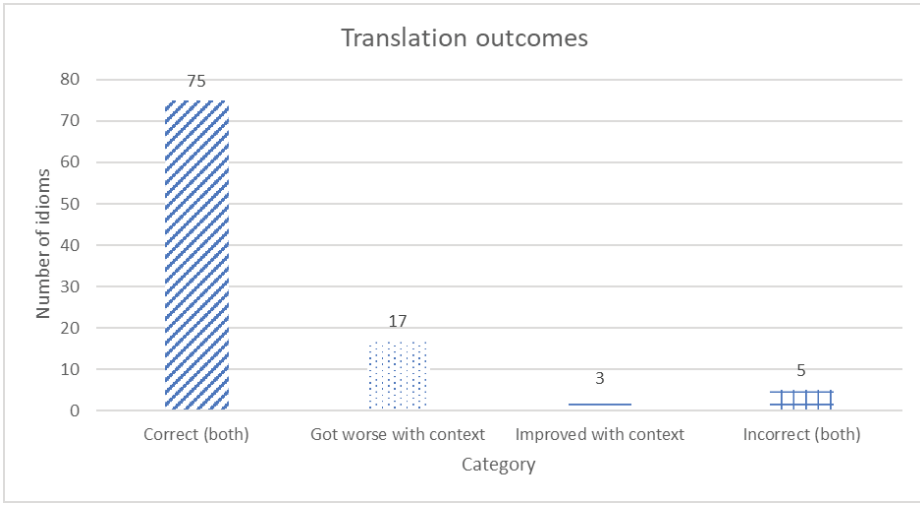

Fig. 2. Translation outcomes divided into four categories

A closer look at the individual examples supports the quantitative findings by showing how contextual information sometimes caused the AI to shift from accurate idiomatic expressions to less appropriate or overly literal renderings. For instance, the idiom *cry wolf* was initially rendered correctly by the AI and as such it was a semantically acceptable and idiomatic equivalent. However, when placed in a contextual sentence, the AI produced a literal back-translation. This rendering fails to convey the figurative meaning and results in a loss of communicative intent. Interestingly, it demonstrates the model's tendency to prioritize surface-level lexical matching over pragmatic interpretation when embedded in context.

As for the second example, *an old flame,* the situation was the same. The translation without context is correct, however, there is a literal translation

whencontext is provided. According to SSSJ (Jarošová et al. 2021), the Slovak word *plameň* ‘flame’ denotes figuratively to passion or zeal and is linked also to love in the collocation *plameň lásky* ‘flame of love’. As a result, the AI would have ended up with a correct translation if it had added the word *lásky* to *plameň*, however, again it failed to deliver the semantic information within its translation proposal.

In five cases, the AI produced incorrect translations regardless of whether contextual information was present or not (Tab. 2).

Idiom	Translation by Kvetko	Translated by AI	Idiom in context	Translated by AI
agree to differ	myslieť si svoje	zhodnúť sa, že sa nezhodneme	Sometimes in a close friendship, where important matters are concerned, people ***agree to differ***, and fall silent.	Niekedy v blízkom priateľstve, kde ide o dôležité veci, sa ***ľudia zhodnú, že sa nezhodnú***, a zmĺknu.

Tab. 2. Examples of mistranslations in both stages

This example illustrates a case in which the translation produced by the AI may initially appear acceptable, as it conveys a meaning that a Slovak reader can understand both in isolation and within context. However, it reveals two key shortcomings: it is fully literal and lacks the natural fluency characteristic of idiomatic Slovak. The resulting expression, while intelligible, is awkward and stylistically marked, deviating from conventional usage. Instances of this nature were observed sporadically and should be considered exceptions rather than representative of the overall translation patters.

Qualitative analyses revealed several patterns. In some cases, the AI system initially proposed a correct Slovak equivalent out of context but selected a less idiomatic or excessively literal rendition when context was added. In other instances, it could not integrate the idiom meaningfully into the wider sentence structure. This suggests limitations on contextual reasoning or phraseological awareness.

Nonetheless, particular attention was given to the factor of improvement (Tab. 3). If the translation with context was identical to the translation without context, it was deemed acceptable and adequate. However, qualitative analysis showed that in 20 cases, the contextual translation was considerably improved in terms of idiomaticity and naturalness. This highlights that while binary statistical methods identify only a small number of improvements, a more nuanced linguistic analysis reveals a greater degree of positive change due to context.

Idiom	Translation by Kvetko	Translated by AI	Idiom in context	Translated by AI
beat the air	hádzať hrach na stenu	márne sa snažiť; robiť niečo zbytočne	"Aren't we ***beating the air***, Blythe?"	***"Nehovoríme len do vetra***, Blythe?"
a body blow	ťažký úder, vážny neúspech	tvrdá rana	"That's your new assignment." It was ***a body blow*** and Liz winced.	"To je tvoja nová úloha." Bola to ***rana pod pás*** a Liz sa strhla.

Tab. 3. Examples of improved translations

4 DISCUSSION AND CONCLUSION

This study reveals several promising and surprising insights regarding the capabilities of AI in idiom translation. Interestingly, a substantial proportion of the 100 English idioms were rendered correctly by the AI, even without contextual support. This result challenges common assumptions that non-compositional, figurative expressions fall outside the reach of computational models. Given the longstanding view that idioms resist rule-based or literal translation, the high baseline accuracy demonstrated here is a compelling indication of the progress made in LLMs.

Nevertheless, the study also exposed critical limitations. The AI often produced literal, stylistically awkward, or semantically mismatched translations when context was introduced. These cases suggest that while surface-level idiomatic retrieval may be successful, deeper contextual and pragmatic integration remains a challenge. Furthermore, a major technical drawback emerged during batch translation attempts: when prompted with a list of idioms in spreadsheet format, the system processed only five, requiring the rest to be input manually. This underpins inefficiencies in AI interaction design for linguistic research.

While the present study was not corpus-driven in design, future work could benefit from a closer integration with corpus linguistics. For example, idiom translations generated by AI could be compared with those found in parallel corpora. However, this approach would be limited by the availability and structure of idioms in such corpora, because identifying and aligning idiomatic expressions remains complex.

It is also important to note that while statistical analysis showed only three improvements due to context, qualitative assessment found 20 cases with considerably improved idiomaticity. This suggests that broader evaluation criteria can offer a fuller picture of translation quality.

Future research could extend the current findings by exploring idiom translation in reverse direction – from Slovak into English – and further across other language

pairs, such as Slovak-German or English-German. Such studies would allow comparative insights into whether AI systems perform differently depending on the source and target language, especially in the case of structurally distant or closely related languages. In addition, future experiments could incorporate low-resource idioms, culturally bound expressions, or idioms with multiple transferred layers, which would further test the model's semantic awareness. Research on how prompt engineering and fine-tuning influence idiomatic output also remains a promising avenue.

ACKNOWLEDGEMENTS

The research has been elaborated within the project A-25-106/3020-18 *Artificial intelligence: challenges for linguistics and marketing*.

References

Abjalova, M., and Sharipova, S. (2024). Semantic and Grammatical Issues in Translating Idioms with Automatic Translation Systems. In 2024 9th International Conference on Computer Science and Engineering, pp. 58–63.

Artamonova, M. V. et al. (2023). Chatbot as a translation tool. In Litera 8, pp. 235–253.

Baziotis, Ch. et al. (2023). Automatic Evaluation and Analysis of Idioms in Neural Machine Translation. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Dubrovnik: Association for Computational Linguistics, pp. 3682–3700. Accessible at: https://aclanthology.org/2023.eacl-main.267/.

Derner, E., and Batistič, K. (2023). Beyond the Safeguards: Exploring the Security Risks of ChatGPT. Accessible at: https://arxiv.org/abs/2305.08005.

Dong, Y. et al. (2019): Acquisition of interpreting strategies by student interpreters. The Interpreter and Translator Trainer 13(4), pp. 408–425.

Donthi, S. (2025). Improving LLM Abilities in Idiomatic Translation. In Proceedings of the First Workshop on Language Models for Low-Resource Languages. Abu Dhabi: Association for Computational Linguistics, pp. 175–181. Accessible at: https://arxiv.org/abs/2407.03518.

Hakami, A. H., and Abomoati, G. S. (2024). Exploring the Impact of Prompt Formulation in AI Chatbots on the Translation of English-to-Arabic and Arabic-to-English Idioms: A Case-Study. Pakistan Journal of Life and Social Sciences 22(2), pp. 21371–21381.

Hamood, M. I. (2024). The Translation of English Food Idioms into Arabic Through ChatGPT: Problems and Solutions. Accessible at: https://shorturl.at/9TcUT.

Jarošová, A. et al. (2021). Slovník súčasného slovenského jazyka. Bratislava: Veda.

Jiao, W. et al. (2023). Is ChatGPT a Good Translator? Yes With GPT-4 As The Engine. Accessible at: https://arxiv.org/abs/2301.08745.

Li, J. et al. (2024). Eliciting the Translation Ability of Large Language Models via Multilingual Finetuning with Translation Instructions. In Transactions of the Association for

Computational Linguistics 12, pp. 576–592. Accessible at: https://aclanthology.org/2024.tacl-1.32/.

Li, S. et al. (2023). Translate Meanings, Not Just Words: IdiomKB's Role in Optimizing Idiomatic Translation with Language Models. In Computation and Language. Accessible at: https://arxiv.org/abs/2308.13961.

Lund, B. (2023). ChatGPT and a New Academic Reality: Artificial Intelligence-Written Research Papers and the Ethics of the Large Language Models in Scholarly Publishing. In Journal of the Association for Information Science and Technology. Accessible at: https://arxiv.org/abs/2303.13367.

Mughal, U. A. et al. (2024). The intersection of linguistics and artificial intelligence: A corpus-based study of idiom translation. Journal of applied linguistics and Tesol 7(4), pp. 1453–1460.

Obeidat, M. M. et al. (2024). Analyzing the Performance of Gemini, ChatGPT, and Google Translate into Rendering English Idioms into Arabic. Journal of Social Sciences 18(4), pp. 1–18.

Rodriguez, P. R. (2024). Phraseological evaluation of automatic interpretation assisted by Yandex. Translation Matters 6(2), pp. 115–130.

Sinclair, J. (1991). Corpus, Concordance, Collocation. Oxford: Oxford University Press.

Sison, A. J. et al. (2023). ChatGPT: More than a Weapon of Mass Deception, Ethical challenges and responses from the Human-Centered Artificial Intelligence (HCAI) perspective. Accessible at: https://arxiv.org/abs/2304.11215.

Wang, X., and Fantinuoli, C. (2024). Exploring the correlation between human and machine evaluation of simultaneous speech translation. In Proceedings of the 25th Annual Conference of the European Association for Machine Translation, Sheffield: EAMT, pp. 327–336. Accessible at: https://aclanthology.org/2024.eamt-1.28/.

Zhu, W. (2024). Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis. In Findings of the Association for Computational Linguistics: NAACL 2024. Mexico City: Association for Computational Linguistics, pp. 2765–2781. Accessible at: https://arxiv.org/abs/2304.04675.