Research Journal in Advanced Humanities

RJAH

https://doi.org/10.58256/4442fg06

Open Access

Check for updates

RESEARCH ARTICLE

Section: *Literature, Linguistics & Criticism*

Beyond the literal: Machine translation performance and strategies in rendering audiovisual political idioms

Bilal Alsharif[1]* & **Razan Khasawneh**[1]

[1]Al-Zaytoonah University of Jordan, Jordan

*Correspondence: Bilal Alsharif, b.alsharif@zuj.edu.jo

ABSTRACT

This is a mixed-method research that integrates quantitative and qualitative research methodologies. It aims to identify and categorize the translation strategies employed by four selected MT systems in rendering audiovisual (AV) political idioms as categorized by Helleklev (2006). Moreover, it aims to evaluate the quality of the MT systems' translations of audiovisual political idioms by comparing them to a human translation reference. By comparing the four MT systems, the study found that the system's most frequently used translation strategy is "using an explanatory everyday expression," which is used mainly by ChatGPT-4 (56.89%) compared to other systems. Notably, "an everyday expression is translated by an idiom" occurred only once in ChatGPT-4 translation (1.72%). The human translator exhibits a lower frequency of word-for-word translation than the four MT systems, demonstrating a lower frequency of translating an idiom with an equivalent one. Correspondingly, the analysis reveals that Gemini employed strategies that align most closely with those of the human translator among all three other systems. Comparing the four systems to the human reference, the study also found that the BLEU scores vary significantly, with Matecat outperforming the other systems with a BLEU score 55.77. This indicates that its translation output is the closest to the human reference among all systems. Meanwhile, Microsoft Bing scored the lowest score of 24.39, suggesting that it is unreliable in translating idioms. All systems, however, have scored comparatively close to one another in terms of length ratio. This demonstrates that, in comparison to the human reference, their translations are frequently of an appropriate length.

KEYWORDS: audiovisual translation (AVT), bilingual evaluation understudy (BLEU), idioms, machine translation (MT), translation strategies

Research Journal in Advanced Humanities

Volume 6, Issue 2, 2025

ISSN: 2708-5945 (Print)

ISSN: 2708-5953 (Online)

ARTICLE HISTORY

Submitted: **18 May 2025**

Accepted: **19 June2025**

Published: **29 June 2025**

HOW TO CITE

Alsharif, B., & Khasawneh, R. (2025). Beyond the literal: Machine translation performance and strategies in rendering audiovisual political idioms. *Research Journal in Advanced Humanities*, *6*(2). https://doi.org/10.58256/4442fg06

Published in Nairobi, Kenya by Royallite Global, an imprint of Royallite Publishers Limited

Introduction

In today's interconnected, globalized world, the demand for translation has expanded drastically. Meeting this demand through human translation alone far outstrips the available human capacity, time, and cost, which led to the emergence of machine translation (MT) systems. By translating text between languages, MT simplifies transferring knowledge (Rawling & Wilson, 2021). With big data and artificial intelligence, MT technologies have advanced from rule-based to neural approaches, providing real-time solutions. However, there are still concerns regarding the output quality of these MT translations (Sabtan et al., 2021). In order to ensure that appropriate and efficient translations are produced, an excellent translation system must have strong language comprehension and generation skills. Several studies have indicated that the quality of MT is more deficient than human translation (Rivera-Trigueros, 2022; Li & Chen, 2019).

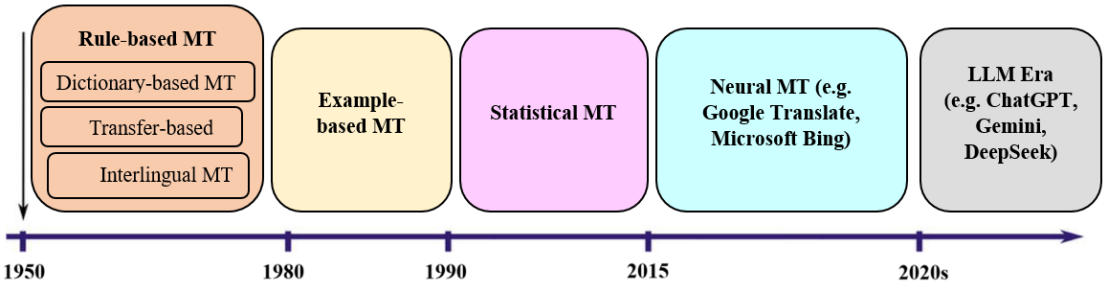

Figure (1): Timeline of MT evolution.

Translation is not merely a linguistic task of converting words; it is a cross-cultural process that navigates the intricate layers of a language. Language is not limited to its literal sense; it also contains deeper, figurative layers of meaning (Khasawneh & Alsharif, 2025a). Translating idioms is one of the most challenging and important tasks for translators, as understanding a language necessitates knowing its idioms and their precise meaning. For machine translation (MT), this presents a significant challenge since idioms are culturally bound, and an accurate translation requires cultural and contextual understanding that goes beyond literal word equivalence. Baker (2018, p. 67) states that idioms are "frozen patterns of language" that accept only minor or no changes in form and convey meanings that cannot be derived from their constituent elements. Many researchers have focused their studies on investigating the strategies used to translate idioms from English into Arabic in different contexts. Mounadil (2023) investigated the strategies used to translate idioms in the literary text *Animal Farm* by George Orwell, employing Baker's (2011) strategies. Al-kaabneh (2023) examines how Royal Court translators render metaphors and idioms in King Abdullah II of Jordan's political speeches, focusing on Nida's (1969) equivalence theory. In their study, Ali and Sayyiyed Al-Rushaidi (2016) aimed to identify the difficulties and strategies encountered by undergraduate students when rendering idioms. Abu Ain (2014) explored how metaphors and idioms in King Hussein of Jordan are translated into English, focusing on cultural and ideological fidelity by applying Newmark's (1988) framework. However, few studies have investigated the translation strategies employed by MT and AI systems. Obeidat et al. (2024) analyzed the performance of three MT systems in conveying the style and sense-based features of randomly selected idioms, employing Baker's taxonomy framework. Meanwhile, Aldelaa and Malkawi (2024) explored the problems Neural Machine Translation (NMT) systems face when translating idioms from Arabic into English, focusing on literary texts (novel).

In contrast to prior work, the present study aims to identify and categorize the translation strategies employed by four selected MT systems in rendering audiovisual (AV) political idioms. In addition, it aims to evaluate the quality of the MT systems' translations of audiovisual political idioms by comparing them to human translation as a reference. Audiovisual translation (AVT) presents several challenges to MT systems since these systems are trained on a large dataset of grammatically correct written texts, not spoken language (Burchardt et al., 2016). Besides, AVT has synchronization and space constraints and is rich in slang, idioms, humor, and informal speech. Accordingly, the current study aims to answer the following questions:

1. What are the predominant translation strategies employed by all four MT systems to translate audiovisual political idioms, as categorized by Helleklev (2006)?
2. Which MT system aligns most closely with human translator strategies?
3. How do BLEU scores and length ratios vary across systems compared to the human reference?

Literature Review

Machine translation

Machine translation (MT) is a sub-field of computational linguistics that uses software to translate a text or speech from one language into another. The fundamental concept of MT is automating the translation process and minimizing or eliminating the need for human intervention. Like other translators, it aims to produce a translation that preserves the original text's meaning (Irfan, 2017). However, while MT and human translators seek precision and faithfulness, their outputs often require refinement to meet quality standards. According to Alsharif et al., (2025), MT entails pre-processing, which refers to preparing the text for the MT system while leaving its linguistic analysis intact, or post-editing, which involves human revision of the output once the MT system has produced its translation.

Despite the continued role of human translators, the growing demand for large-scale translation services highlights the necessity and advantages of MT. According to Benyahia (2024), there is a high demand for translation that exceeds the human translator's capacity. Consequently, MT systems offer a more efficient, cost-effective, and beneficial solution, particularly where human-quality translation is unnecessary and a general understanding—of the "gist"— of the ST is needed (Benyahia, 2024). However, the effectiveness of online tools, including machine translation, is significantly impacted by 'inadequate infrastructure' and the translators'/students' lack of a positive learning or working environment (Khasawneh et al., 2025). Melby (1996) suggests that in the future, computers could potentially make judgments and decisions that align with the non-linguistic knowledge people often use in their daily lives. While MT seeks to preserve meaning, human translation inherently serves as a bridge for communication, helping to overcome differences in culture, customs, and values between languages (Khader, 2023). This cultural mediation remains a challenge for MT systems, especially when translating idioms.

Several MT systems have been developed, each using a unique methodology to achieve automated translation. Over time, MT has evolved from Rule-Based MT (RBMT) Systems (e.g., SYSTRAN) and computer-assisted translation (CAT) tools (e.g., Matecat) that incorporate different MT engines to Neural machine translation (NMT) with AI enhancements (e.g., Microsoft Bing), and Large Language models (LLMs) representing the new generation of translation tools (e.g., Gemini, ChatGPT).

Matecat is a free CAT tool with integrated MT suggestions that offer different features such as glossary management, translation memory (TM), and real-time collaboration (Nigam, 2025). However, Matecat can be slow in the project's last stages (Karpina, 2024), requires an ongoing internet connection, and may have consistency and file size limits. On the other hand, Microsoft Bing is user-friendly and has been developed for a wide range of languages. According to Arthur et al. (2016), when translating low-frequency content phrases essential to understanding the meaning of the sentence, NMT commonly makes mistakes. Meanwhile, both Gemini and ChatGPT perform several tasks, including text summarizing, writing assistance, and translation between several languages. Gemini outperformed other LLM-based translation techniques and performed exceptionally when translating from English to other languages (Team et al., 2023). ChatGPT comprehends and generates human-like texts with a high fluency level (Bender et al., 2021). However, both can make translation errors, lack human touch, and privacy concerns (Al-Salman & Haider, 2024; Mukhtar, 2025).

Machine Translation Evaluation

The term MT evaluation denotes the various methods used to assess how well a machine translation system performs. Various methods have been used to evaluate translation output, and whether the translation is human or machine-generated, this process is important to ensure accuracy, fluency, and adequacy. There are two main approaches to MT evaluation: human (manual) and automatic. Rivera-Trigueros (2022) states that human evaluation involves identifying errors, annotating, and classifying. Despite being the most dependable approach, human evaluation has several drawbacks. It is slow, expensive, inflexible, and it cannot be replicated (Khasawneh & Alsharif, 2025b), which can be due to the subjective matter and reliance on individual judgment. Human evaluators must meet specific standards, including adequate training, clear evaluation criteria, and familiarity with the text's subject matter (Rivera-Trigueros, 2022). Evaluators' personal judgments, language preferences, or cultural backgrounds can also affect consistency. Yet, automatic evaluation, on the other hand, uses computational metrics without human intervention. This method is less expensive, more objective, and requires less human effort. It also relies on human-created "reference" translation, whose quality is often

presumed but not verified (Castilho et al., 2018). However, automatic evaluation metrics originated from systems with old architectures, which are occasionally not modified to fit modern paradigms (Way, 2018). The most widely used conventional measures are METEOR, TER, and BLEU. BLEU, which was developed by Papineni and his colleagues in 2002, measures the difference between an automatic translation and a human-created reference translation. It compares the overlap between MT output and reference translations regarding unigrams (single words) and high-order n-grams (Maučec & Donaj, 2019). Accordingly, BLEU mainly uses n-gram precision ranging from unigram (single word) to more extended sequences (bigrams, trigrams, and four-grams) (Bronsdon, 2025). It also adds a brevity penalty for translations shorter than the reference translations to avoid. The output is a number between 0 and 1, or 0 to 10, or 0 to 100, with higher values representing more similar text, i.e., closer the translation to the human translation reference.

This method is criticized for its poor performance at the sentence level and weak handling of recall (Callison-Burch, Osborne, & Koehn, 2006). Moreover, specific languages have lower scores than others because BLEU penalizes sentence length and word order and does not capture fluency or semantic similarity (Segonne & Mickus, 2023; Haque et al., 2022). However, it is still the most popular and widely used method for MT evaluation due to its minimal processing needs and cost. It also would provide a prompt indicator of the accuracy and vocabulary employed.

Translation of Idioms

Every language has a different culture, and idioms are deeply rooted in every culture as part of the language. Idioms are phrases and expressions whose meaning cannot be understood from the literal meaning of the individual words. They add vibrancy, color, nuance, and depth to daily conversations. Idioms cannot be easily classified; however, the existence of different types suggests variations in how they are learned and understood. According to Kvetko (2005), idioms can be classified into unchangeable idioms (fixed expressions that cannot be modified) and unchangeable idioms (allow variation in grammar, lexis, and Orthography). Another classification is proposed by Fernando (1996), who classify idioms into three types: pure Idioms (non-literal expressions consisting of multiple words and opaque meaning), semi-idioms (a mix of literal and non-literal components), and literal idioms (transparent with minimal variation). These classifications emphasize the difficulties in comprehending idioms based on their semantic transparency, which is particularly helpful in translation, language teaching, and lexicography.

Since the meaning of an idiom is not the product or result of its components or constitutes (Hussein et al., 2000), idioms pose several challenges not only for non-native speakers to master but also for translators to render from one language to another as often a literal translation does not make any sense. These difficulties are compounded by the complex interactions between languages that even bilinguals exhibit (Alsharif & Khasawneh, 2025), which highlights the subtle and often non-literal nature of cross-linguistic communication. Li et al. (2024) suggest that translating idioms presents difficulties due to their non-literal nature, necessitating careful contextual analysis to interpret and convey their intended meaning accurately. Strakšienė (2009) has a different point of view; he mentions that a translator's major problem when translating idioms is the absence of equivalence on the idiom level. The challenges noted by Khasawneh et al. (2025) in translating Arabic figures of speech, such as metaphor and metonymy, resonate with the complexities of rendering political idioms, where cultural context plays a pivotal role. A critical task for translators is the transposition of cultural components from the SL into the TL, which frequently necessitates the replacement of textual elements with their functional equivalents (Alazzam et al., 2024). Correspondingly, translators should be careful when using idioms in the TL, which might seem superficially similar, yet their meaning often goes beyond their literal words. Baker (1992) enlists four strategies for translating idioms, including utilizing an idiom with a similar meaning and form in the TL, using an idiom of similar meaning but dissimilar form, using paraphrase to convey the meaning of the idiom, and, in some cases, omitting the idiom in translation. On the other hand, Helleklev (2006) lists four ways to translate idioms, which are the primary focus of the current study:

a. Translating an idiom by an equivalent idiom.
b. Word-for-word.
c. With an explanatory everyday expression.
d. An everyday expression is translated by using an idiom.

Methodology

This is a mixed method research that integrates quantitative research methodology to investigate some trends that subsequently be further examined using qualitative data (Saldanha & O'Brien, 2014). Figure 2 illustrates the methodological steps undertaken in the study to compare AI and human translations of English idioms translated into Arabic. The former American President Joseph Biden's audiovisual political speeches delivered in English were analyzed to identify all idiomatic expressions used with their professional human-translated Arabic version to achieve the study aims. The idioms are translated using four MT systems: ChatGPT-4 and Gemini (LLMS), Matecat (CAT tool), and Microsoft Bing (NMT). These platforms are selected for translation analysis to assess a broad and representative range of modern translation technologies. These platforms each represent a different paradigm in machine-assisted translation today.

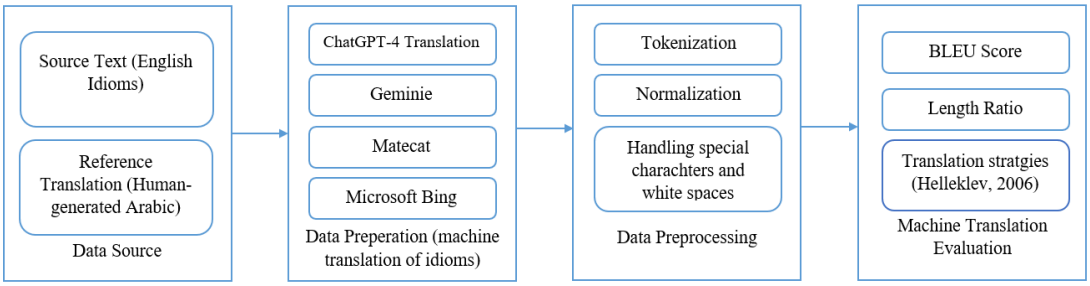

Figure (2): Design of the Study.

Before calculating the BLEU score, data undergo "data preprocessing," which involves: 1. tokenization, segmenting text into analyzable units (words/ subwords); 2. normalization: standardizing text format (e.g., lowercase, punctuation); 3. handling special characters and white spaces to ensure clean input and avoid parsing errors. Finally, in the MT evaluation, the quality of MT is assessed using several metrics, including the BLEU score, which is used to measure accuracy and similarity to the human reference translation. The higher the BLEU score, the closer the translation to the human translation reference. The length ratio, which flags omissions and additions, is used to compare the length of the MT to the reference. The length ratio is calculated per idiom to understand length deviation at a more fine-grained level. A value close to 1.0 means the machine translation's length is very similar to the human reference's length. Python, a high-level, general-purpose programming language, is used to calculate BLEU and length ratio. Figure 3 shows a general interpretation of the BLEU score:

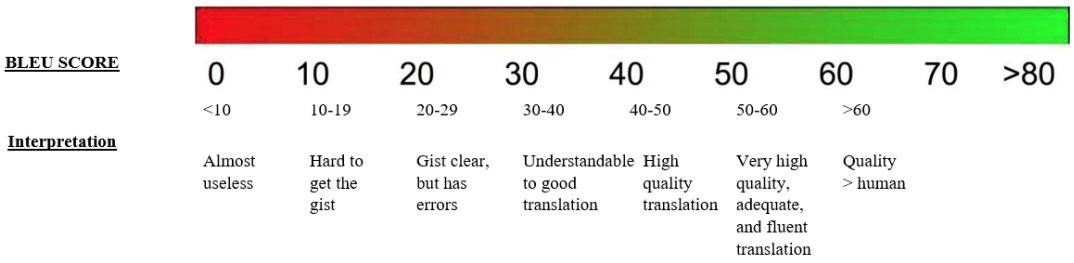

Figure (3): Interpretation of the BLEU Score.

After that, the translation strategies used by the human translator and every MT system were identified by incorporating Helleklev's (2006) framework of translating idioms: translating an idiom by an equivalent idiom, word for word, with an explanatory everyday expression, or an everyday expression is translated by using an idiom. The simplicity and clarity of Helleklev's (2006) categorization make it easier to apply consistently in comparative analyses, particularly for studies that concentrate exclusively on idioms and figurative language.

Results and Discussion

This section presents a comparative quantitative analysis of idiom translation strategies used by human translators and four different machine translation systems: ChatGPT-4, Gemini, Matecat, and Microsoft Bing, based on Helleklev's (2006) four methods.

Table (1): The Number and Percentage of Strategies Observed in the Human Translation Vs. Four AI tools

No.	Strategy	Human Translator	ChatGPT-4	Gemini	Matecat	Microsoft Bing
1	Translating an idiom by an equivalent idiom	44.73 %	17.24 %	22.41 %	18.96 %	15.51 %
2	Word-for-word	12.06 %	25.86 %	25.86 %	27.58 %	31.03 %
3	With an explanatory everyday expression	58.62 %	55.17 %	51.72 %	53.44 %	53.44 %
4	An everyday expression is translated by using an idiom	0	1.72% %	0	0	0

According to Table (1), the human translator predominantly uses explanatory everyday expression (58.62%), followed by translating an idiom by an equivalent idiom (44.73%). It rarely resorts to a word-for-word strategy (12.06%) or translation using an idiom for everyday expression (0%).

In contrast, ChatGPT-4 tends to use an explanatory everyday expression strategy (56.89%) more frequently than other strategies. Word-for-word (25.86%) comes next, and then translating using an equivalent idiom (17.24%). It also uniquely used the substitution of an everyday expression using an idiom (1.72%). Gemini's shows a similar trend, favoring explanatory everyday expression (51.72%) and word-for-word (25.86%) strategies, with minimal use of equivalent idioms (22.41%). Matecat and Microsoft Bing follow similar patterns, leaning heavily on explanatory everyday expression (53.44%, 53.44, respectively) and word-for-word (27.58%, 31.03%, respectively) strategies, and a smaller percentage for equivalent idioms (18.96%, 15.51%, respectively).

It can be noted that the human translator and all four machine translation systems have the highest tendency for explanatory everyday expression strategy. In addition, the human translator uses word-for-word less frequently, while the four translation systems use the fewest equivalent idioms. By comparing the percentages for each strategy among the human translator and the four MT systems, Gemini aligns most closely with the human translator strategies, scoring the lowest total difference. Like the human translator, Gemini favors the use of an explanatory everyday expression and employs word-for-word less than ChatGPT-4, Matecat, or Bing. Its equivalent idiom usage is more than Bing's, although it is still lower than the human translator.

In addition, the BLEU score is used to evaluate the quality of the four MT systems by comparing them with the Arabic human-generated translation, known as reference text. According to Table (2), Matecat achieved the highest BLEU score among all four AI tools, with a score of 55.77. Gemini comes next with a BLEU score of 32.73. Meanwhile, ChatGPT-4 obtained a BLEU score of 27.77, and Microsoft Bing achieved the lowest score at 24.39. Hence, Matecat outperforms all other systems, while the other three systems had considerably lower and relatively closer scores to each other. This suggests that Matecat translation is the closest to human quality in terms of n-gram overlap and fluency. As for the length ratio, ChatGPT-4 and Matecat have the highest length ratio, suggesting that their translations are closer in length to the human reference translation. Following is Gemini, which has a 0.96 length ratio and is also close to the human reference. Microsoft Bing has the lowest length ratio score with 0.93, indicating that its translation is slightly shorter than the human translation reference on average compared to other systems. Generally, all systems have an appropriate length compared to the human reference.

Table (2): BLEU Score

	AI tool	BLEU	Length ratio
1	ChatGPT-4	27.77	0.98
2	Gemini	32.73	0.96
3	Matecat	56.77	0.98
4	Microsoft Bing	24.39	0.93

The section below demonstrates how every MT system translated each idiom, highlighting the chosen translation stratgies in comparison to human translation reference.

Example (1)

	Idiom	Strategy	BLEU	Length ration
Source	in lockstep			
Human Translation	بخطى ثابتة	Translating an Idiom by an Equivalent Idiom	1	1
ChatGPT-4	بخطى متناسقة	With an explanatory everyday expression	0.73	0.67
Gemini	على خط واحد	With an explanatory everyday expression	0	1.00
Matecat	على اتصال	With an explanatory everyday expression	0	0.67
Microsoft Bing	التوافق	With an explanatory everyday expression	0	0.33

In the above example, the idiom "in lockstep" describes an organized work in which entities closely align their decisions or actions. Only the human translator used an equivalent idiom to translate the idiom in the ST, rendering it into "بخطى ثابتة," which means "with firm steps." On the other hand, all four machine translations used an explanatory everyday strategy. Nevertheless, their translations and BLEU scores varied significantly. ChatGPT-4 translated the idiom into "خطى متناسقة," which literally means "harmonious steps," resulting in a moderately accurate translation in comparison to the human translation with a low BLEU score of 0.73 and a length ratio of 0.67. Gemini translated the idiom as "على خط واحد," which literally means "on one line," showing no overlap with the reference human translation with a BLEU score of 0 and 1 as length ratio. Matecat translated the idiom into "على اتصال," meaning "in contact," scoring a BLEU score of 0 and 0.67 length ratio. Finally, Microsoft Bing used "التوافق," which literally means "the agreement," to translate the idiom with a BLEU score of 0 and a significantly shorter length ratio of 0.33.

The human translator successfully used an equivalent idiom in Arabic, resulting in a high-quality translation. Meanwhile, the MT systems failed to produce translations that closely match the human translation, as can be noted from their near zero BLEU score, despite attempting to explain the idiom's meaning through everyday expressions.

Example (2)

	Idiom	Strategy	BLEU	Length ratio
Source	look the other way			
Human Translation	سنغض النظر في الاتجاه الآخر	With an explanatory everyday expression	1	1
ChatGPT-4	تغاضى	Translating an idiom by an equivalent idiom	0.22	0.50
Gemini	تجاهل عمدًا	With an explanatory everyday expression	0.13	0.50
Matecat	سنغض النظر في الاتجاه الآخر	With an explanatory everyday expression	1	1
Microsoft Bing	تجاهل	With an explanatory everyday expression	0.13	0.25

In example (2), the idiom refers to ignoring anything improper or illegal. It frequently suggests a conscious decision to ignore a problem rather than face it. The human translator rendered the idiom "look the other way" as "سنغض النظر في الاتجاه الآخر," which literally means "we will look the other direction," using an explanatory everyday expression strategy. ChatGPT-4 successfully translated the idiom using an equivalent idiom in the TL, "تغاضى," meaning "we will turn a blind eye." However, its translation achieved a 0.22 BLEU score and a length ratio of 0.50, showing a significant difference from the human translation.

Meanwhile, Gemini translated the idiom as "تجاهل عمدًا," which literally means "look away," utilizing the explanatory everyday expression. This results from a BLEU score of 0.13 and a length ratio of 0.50, which indicates a considerable divergence from the human translation. Matecat produced an identical translation to the human reference one, achieving a perfect BLEU score of 1 and a length ratio of 1. Microsoft Bing produced the shortest translation using "تجاهل," meaning "ignore." The explanatory everyday expression strategy resulted from a BLEU score of 0.13 and a length ratio of 0.25.

Both the human translator and Matecat provided a similar translation using explanatory everyday expressions; however, ChatGPT-4 produced an equivalent idiom but with a lower BLEU score and length ratio. Gemini and Microsoft Bing offered explanatory expressions, resulting in a low BLEU score and length ratio. This suggests how different translation systems handle idioms and how their success rates vary.

Example (3)

	Idiom	Strategy	BLEU	Length ratio
Source	**welcome** them here **with open arms**			
Human Translation	**ةحوتفم عرذأب** انه مهب **بحرن**	word-for-word	1	1
ChatGPT-4	**بحرن عرذأب** انه مهب **ةحوتفم**	word-for-word	1	1
Gemini	**بحرن عرذأب** انه مهب **ةحوتفم**	word-for-word	1	1
Matecat	**بحرن عرذأب** انه مهب **ةحوتفم**	word-for-word	1	1
Microsoft Bing	**نسنتبقبلهم** انه **بحافوة**	Translating an idiom by an equivalent idiom	0	0.60

The idiom " **welcome with open arms**" is used to greet someone warmly and enthusiastically. The human translator used a word-for-word strategy to translate the idiom "welcome with open arms" as "عرذأب بحرن ةحوتفم." The same strategy is used by ChatGPT-4, Gemini, and Matecat, producing the exact translation and achieving a perfect BLEU score of 1 and a length ratio of 1, indicating a perfect match with the human reference translation. In contrast, Microsoft Bing produced a translation using an equivalent idiom in the TL, "نسنتبقبلهم بحافوة," meaning "welcome them here warmly." However, this translation received a BLEU score of 0 and a 0.60 length ratio, showing no overlap with the human reference translation.

Although the human translators ChatGPT-4, Gemini, and Matecat favored a word-for-word translation strategy, Microsoft Bing successfully produced a related and understandable translation that is appropriate for contexts where natural Arabic is more important than adherence to the original text image.

Example (4)

	Idiom	Strategy	BLEU	Length ratio
Source	to squeeze			
Human Translation	للضغط	With an explanatory everyday expression	1	1
ChatGPT-4	لتضييق الخناق	An everyday expression is translated using an idiom	0.43	2.00
Gemini	طغض الضغط لزيادة	With an explanatory everyday expression	0.73	2.00
Matecat	للضغط طغض على	With an explanatory everyday expression	0.73	2.00
Microsoft Bing	للضغط	With an explanatory everyday expression	1	1

In the above example, the idiom "to squeeze" is used to exert pressure on something. The human translator used explanatory everyday expression to translate the idiom "to squeeze" into "للضغط," which literally means "to pressure/ press." ChatGPT-4 rendered the above everyday expression using an idiom "لتضييق الخناق," meaning "to tighten the grip." This resulted in a BLEU score of 0.43 and a length ratio of 2.00, suggesting a noticeable difference in wording and twice as long a translation from the human reference translation. Gemini used a similar strategy to the human translation, adding the word "لزيادة" to emphasize the intent of increasing or intensifying pressure implied by the original. Both Gemini and Matecat achieved a BLEU score of 0.73 and a 2.00 length ratio, indicating a partial match to the reference and a longer but still clear translation. On the other hand, Microsoft Bing matched the human translation output, producing a perfect match with a BLEU score of 1 and a length ratio of 1.

The human translators, Gemini, Matecat, and Microsoft Bing, used a direct and concise translation suitable for technical and literal needs, while ChatGPT-4 successfully produced a translation that is idiomatic and more suitable for contexts where figurative meaning is needed.

Example (5)

	Idiom	Strategy	BLEU	Length ratio
Source	pay the price			
Human Translation	يدفع الثمن	Translating an idiom by an equivalent idiom	1	1
ChatGPT-4	يدفع الثمن	Translating an idiom by an equivalent idiom	1	1
Gemini	يدفع الثمن	Translating an idiom by an equivalent idiom	1	1
Matecat	يدفع الثمن	Translating an idiom by an equivalent idiom	1	1
Microsoft Bing	يدفع الثمن	Translating an idiom by an equivalent idiom	1	1

In example (5), the idiom is used to experience the consequences of one's actions or misdeeds. The human translator and all four MT systems used the same translation strategy, producing a similar output for the idiom "pay the price." All translations agreed on "يدفع الثمن" as an equivalence for the original achieving a perfect BLEU score of 1 and a length ratio of 1. This denotes that all MT systems perfectly matched the human reference translation regarding lexical choice and length. Consequently, there is a complete agreement among the human translator, and all four MT systems successfully produce an equivalence for the original idiom that is clear and effective. This indicates that the MT systems can perform similarly to the human translator. However, this is likely a result of the idiom's simplicity and the presence of a very direct and widely accepted equivalent in Arabic. In addition, both English and Arabic expressions are widely used in every language, making them easily recognizable for humans and translation systems.

Conclusion

Despite the significant advancement in MT systems, they still fail to deliver a high-quality, publishable translation for all content types. Translating idioms can be challenging for human translators and MT systems due to their non-literal nature. However, MT struggles even more with idioms since they require deep cultural understanding and contextual awareness. This paper aims to identify and categorize the translation strategies employed by four MT systems—ChatGPT-4, Gemini, Matecat, and Microsoft Bing—in rendering audiovisual (AV) political idioms and evaluate their translation quality by comparing them to human translation references. By comparing the translations of the four systems, the study found that the predominant translation strategy employed by all four MT, as categorized by Helleklev (2006), is "using an explanatory everyday expression," which is used by ChatGPT-4 (56.89%) more frequently than other systems. The second most common strategy, "word-for-word", is observed most often in the outputs of Microsoft Bing (31.03%). Following is " using an equivalent idiom," which Gemini (22.41 %) utilized more than other systems. Notably, "an everyday expression is translated by using an idiom" occurred only once in ChatGPT-4 translation (1.72%), suggesting its rarity in the dataset. Human translators and all four machine translation systems predominantly employ "translating an idiom with an explanatory everyday expression."

On the other hand, a notable divergence is observed in the secondary strategies. The human translator demonstrates a lower frequency of word-for-word translation than the four MT systems, which demonstrate a lower frequency of translating an idiom with an equivalent one. This aligns with Obeidat et. al., (2024) study who noted that despite significant technological progress, machine translation still struggles to understand nonliteral language, such as idioms. Hence, by comparing the total absolute differences of the employed translation strategies, the analysis reveals that Gemini has the lowest total differences, suggesting that its employed strategies align most closely with those of the human translator among all other three systems.

The study also revealed that the BLEU score varies significantly among all four systems compared to the human reference. Matecat outperformed the other systems, scoring the highest BLEU score at 55.77, suggesting that its translation output is the closest to the human reference among all systems. Meanwhile, the other systems — Gemini, ChatGPT, and Microsoft Bing have noticeably lower scores with minimal difference from each other, with Microsoft Bing (24.39) at the bottom, indicating that it is unreliable for translating idioms. However, in terms of the length ratio, all systems scored relatively closer to each other. This shows that their translations are often of a suitable length in relation to the human reference.

By closely comparing MT translations of idioms, this study serves as a benchmark of traditional MT, NMT, and LLMs within this context. Additionally, it offers practical recommendations for developers about areas that need improvement and for translators about post-editing strategies. Our future direction is to investigate the translation of idioms in different contexts and across languages.

Authors' Contribution

Bilal Alsharif: study design, data collection, drafting the manuscript, reviewing and editing.
Razan Khasawneh: study design, data collection, drafting the manuscript, reviewing and editing.

Disclosure Statement

No potential conflict of interest is reported by the author(s).

Data Availability Statement

The data will be available on request.

References

Alazzam, T. S., Alzghoul, M. A., & Alzghoul, R. (2024). The Translation of Medical and Health-Related Idioms by University Students from English into Arabic: Challenges and Strategies. Jordan Journal of Modern Languages & Literatures, 16(4), 941-954.

Aldelaa, A. S., & Malkawi, M. A. K. M. (2024). Investigating problems related to the translation of idiomatic expressions in the Arabic novels using neural machine translation. Theory and Practice in Language Studies, 14(1), 71-78.

Ali, H., & Sayyiyed Al-Rushaidi, S. M. (2017). Translating idiomatic expressions from English into Arabic: Difficulties and strategies. Arab World English Journal (AWEJ) Volume, 7.

Al-kaabneh, B. (2023). Equivalence in the translation of metaphors and idioms in King Abdullah II's political speeches. International Journal of Humanities, Philosophy and Language, 6(22), 01-08.

Al-Salman, S., & Haider, A. S. (2024). Assessing the accuracy of MT and AI tools in translating humanities or social sciences Arabic research titles into English: Evidence from Google Translate, Gemini, and ChatGPT. International Journal of Data and Network Science, 8(4), 2483-2498.

Alsharif, B., & Khasawneh, R. (2025). The Production of Laterals by Arabic-English Bilingual Children. World Journal of English Language, 15(5), 301–310.

Alsharif, B., Khasawneh, R., & Alzghoul, M. (2025). Strategies of Rendering Metaphor from Arabic into English: A Comparative Study of ChatGPT and Matecat. World Journal of English Language, 16(1), 45–51.

Arthur, P., Neubig, G., & Nakamura, S. (2016). Incorporating discrete translation lexicons into neural machine translation. arXiv preprint arXiv:1606.02006.

Baker, M. (1992). A coursebook on translation. Routledge.

Baker, M. (2018). In other words: A coursebook on translation. Routledge.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big?. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 610–623).

Benyahia, M. (2024). Examining the Efficiency of Machine Translation in Translating English Idioms used in American Media. Journal of Translation and Language Studies, 5(2), 43–55.

Bronsdon, C. (2025). BLEU Metric: Evaluating AI Models and Machine Translation Accuracy | Galileo Blog. BLEU Metric: Evaluating AI Models and Machine Translation Accuracy - Galileo AI

Burchardt, A., Lommel, A., Bywood, L., Harris, K., & Popović, M. (2016). Machine translation quality in an audiovisual context. Target, 28(2), 206-221.

Callison-Burch, C., Fordyce, C. S., Koehn, P., Monz, C., & Schroeder, J. (2007). (Meta-) evaluation of machine translation. In Proceedings of the Second Workshop on Statistical Machine Translation (pp. 136-158).

Castilho, S., Doherty, S., Gaspari, F., & Moorkens, J. (2018). Approaches to human and machine translation quality assessment. Translation quality assessment: From principles to practice, 9–38.

Fernando, C. (1996). Idioms and Idiomaticity. Oxford: Oxford University Press.

Haque, S., Eberhart, Z., Bansal, A., & McMillan, C. (2022, May). Semantic similarity metrics for evaluating source code summarization. In Proceedings of the 30th IEEE/ACM International Conference on Program Comprehension (pp. 36–47).

Helleklev, C. (2006). Metaphors and terminology in social science: A Translation and an analysis.

Hussein, R. F, Khanji, R. and Makhzoomy, K.F. (2000). The acquisition of idioms: Transfer or what?. Journal of King Saud University, Lang. & Transl, (12), 23–34.

Irfan, M. (2017). Machine Translation. Department of Computer Science Bhria University Islamabad. Retrieved from https://www.researchgate.net/publication/320730405

Karpina, O. (2024). Comparative study of modern CAT tools: Smartcat vs Matecat. In Пріоритети германської і романської філології. Волинський національний університет імені Лесі Українки.

Khader, K. T. (2023). Challenges of Translating Selected Verses from Al-Baqarah Surah into English: An Analytical Study. Al-Zaytoonah University of Jordan Journal for Human & Social Studies, 4(3).

Khasawneh, R. R., Alsharif, B. I., & Khasawneh, R. R. (2025). Online Learning Amidst Crisis: Perceptions of Gazan University Students. In Frontiers in Education (Vol. 10, p. 1659256). Frontiers.

Khasawneh, R. R., Moindjie, M. A., & Kasuma, S. A. A. (2025). Diachronic translation of figures of speech in

Antara's Mu'allaqa. World Journal of English Language, 15(3), 290-290.
Khasawneh, R., & Alsharif, B. (2025a). Translation Strategies of Idiomatic Expressions: A Systematic Review. Forum for Linguistic Studies, 7(10).
Khasawneh, R., Alsharif, B. (2025b). A Comparative Study of AI-powered Tools for Arabic-English and English-Arabic Translation. Journal of Language Teaching and Research, 16(6).
Kvetko, P. (2005). English lexicology: in theory and practice. Univerzita sv. Cyrila a Metoda.
Li, H., & Chen, H. (2019). Human vs. AI: An assessment of the translation quality between translators and machine translation. International Journal of Translation, Interpretation, and Applied Linguistics (IJTIAL), 1(1), 1-12.
Maučec, M. S., & Donaj, G. (2019). Machine translation and the evaluation of its quality. In Recent trends in computational intelligence. IntechOpen.
Melby, A. (1996). Machine translation and other translation technologies. Annual review of applied linguistics, 16, 86–98.
Mounadil, T. (2023). Strategies for translating idioms and proverbs from English into Arabic. British Journal of Translation, Linguistics and Literature, 3(2), 02–09.
Mukhtar, I. A. (2025). Translation and Technology. Transcultural Journal of Humanities and Social Sciences, 6(2), 269–283.
Nigam, P. (2025). MateCat vs Smartcat: Top Differences & Similarities| Doctor Elearning Blog. MateCat vs Smartcat: Top Differences & Similarities - Doctor Elearning
Obeidat, M. M., Haider, A. S., Tair, S. A., & Sahari, Y. (2024). Analyzing the Performance of Gemini, ChatGPT, and Google Translate in Rendering English Idioms into Arabic. FWU Journal of Social Sciences, 18(4).
Rawling, P., & Wilson, P. (2021). The Routledge Handbook of Translation and Philosophy. Abingdon, Oxon: Routledge, Taylor & Francis Group.
Rivera-Trigueros, I. (2022). Machine translation systems and quality assessment: a systematic review. Language Resources and Evaluation, 56(2), 593-619.
Sabtan, Y. M. N., Hussein, M. S. M., Ethelb, H., & Omar, A. (2021). An evaluation of the accuracy of the machine translation systems of social media language. International Journal of Advanced Computer Science and Applications, 12(7).
Saldanha, G., & O'Brien, S. (2014). Research methodologies in translation studies. Routledge.
Segonne, V., & Mickus, T. (2023). " Definition Modeling: To Model Definitions." Generating Definitions With Little to No Semantics. arXiv preprint arXiv:2306.08433.
Strakšienė, M. (2009). Analysis of idiom translation strategies from English into Lithuanian. Kalbų studijos, (14), 13-19.
Team, G., Anil, R., Borgeaud, S., Alayrac, J. B., Yu, J., Soricut, R., ... & Blanco, L. (2023). Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805.
Way, A. (2018). Quality expectations of machine translation. Translation quality assessment: From principles to practice, 159-178.