Evaluating Machine Translation and Automatic Metrics in Subtitling: A Case Study on Spanish Multiword Expressions

María Miró Maestre and **Iván Martínez-Murillo**
Language Processing and Information Systems Group
University of Alicante
{maria.miro, ivan.martinezmurillo}@ua.es

Abstract

Evaluating the translation of multi-word expressions (MWEs) remains a major challenge for Machine Translation (MT), particularly in audiovisual subtitling, where idiomatic meaning and cultural context are essential for adequacy. This study investigates both the ability of state-of-the-art MT systems to translate Spanish MWEs into English and the extent to which current automatic evaluation methods reflect expert human judgment. We introduce ALMO-MWE, a dataset of 235 MWEs extracted from four films by Pedro Almodóvar to evaluate four MT systems using automatic metrics, LLM-as-a-judge approaches, and professional human assessment. Our results reveal a substantial mismatch between traditional automatic metrics and human judgments: n-gram-based metrics show near-zero correlation with expert evaluation and only limited discriminative capacity. In contrast, neural metrics and LLM-based judges exhibit substantially stronger agreement with human assessments, with GPT-OSS achieving the highest overall correlation. These findings highlight fundamental limitations of surface-form metrics for culturally and contextually sensitive translation phenomena and underscore the need for context-aware evaluation frameworks when assessing the translation quality of MWEs in audiovisual translation.

1 Introduction

The current landscape of Machine Translation (MT) systems offers increasingly sophisticated capabilities; however, persistent limitations remain in their ability to incorporate the cultural and contextual knowledge required to accurately adapt messages from a source language into a target language. These limitations become particularly evident in audiovisual translation, especially in subtitling, where cultural references, idiomatic expressions, and text-length constraints strongly influence translation adequacy. Among these challenges, the translation of multi-word expressions (MWEs) represents a particularly difficult task for MT systems. MWEs are often idiomatic, culturally embedded, and context-dependent, making their meaning difficult to infer from their individual components. In audiovisual contexts such as film subtitles, the correct interpretation of MWEs is further constrained by pragmatic and stylistic considerations, increasing the difficulty for automatic systems.

In light of these challenges, this study investigates two closely related questions:

- To what extent current MT systems can adequately translate Spanish MWEs into English?
- How effectively can automatic evaluation methods assess those translations?

To address these questions, we compiled ALMO-MWE, a dataset of 235 MWEs manually extracted from four films directed by Pedro Almodóvar, a filmmaker widely known for his extensive use of regional and colloquial language to shape character identity and dialogue. The expressions were translated automatically using four

broadly used MT systems: GPT-4, OPUS-MT, Gemini, and DeepL.

Translation quality was assessed using a threefold evaluation framework. First, five automatic evaluation metrics were used to compare the MT outputs with the official English subtitles professionally translated from Spanish. Second, a professional translator manually evaluated the translations to provide a human benchmark. Finally, we applied a recent evaluation paradigm in which a Large Language Model (LLM) acts as a judge to compare machine-generated translations with professional subtitles. This combined approach enables a comprehensive assessment of both translation performance and the extent to which different evaluation methods correlate with expert human judgment.

The main contributions of this study are as follows:

1. An empirical analysis of MWE translation from Spanish into English across four state-of-the-art MT systems, focusing on colloquial expressions extracted from film subtitles.

2. A threefold evaluation framework combining automatic metrics, LLM-as-a-judge evaluation, and expert human assessment to analyze the reliability of current evaluation approaches.

3. The creation of a new Spanish-English lexicon of 235 colloquial MWEs extracted from Pedro Almodóvar's films.

Through this investigation, we aim to contribute to the MT research community by providing insights into the challenges that culturally bound linguistic phenomena such as MWEs pose for both translation systems and evaluation methodologies. By examining the discrepancies between automatic metrics and expert human judgments, this work highlights the need for more context-aware evaluation frameworks capable of better capturing the linguistic and cultural complexity of audiovisual translation.

The remainder of this paper is structured as follows. Section 2 reviews recent advances in MT and discusses previous approaches to handling MWEs in translation. Then, Section 3 describes the methodology used to compile the MWE dataset, the MT systems evaluated, and the evaluation methods employed. Section 4 presents the results of the translation and evaluation experiments, while Section 5 discusses the main findings and analyzes the discrepancies between automatic metrics and human evaluation. Finally, Section 6 summarizes the conclusions and outlines directions for future research.

2 Related Work

In this section, we examine the steady evolution of MT systems and some of the current challenges that still need to be address to enhance their performance. We also review the state of the art regarding the integration and evaluation of MWEs within MT systems, underscoring the need for further specialised research on areas such as subtitling translation.

2.1 MT Evolution

MT has evolved considerably since its inception, achieving remarkable performance (Ranathunga et al., 2023). The earliest systems, developed in the 1950s, were rule-based and produced rigid and often unnatural translations (Bel'skaja, 1957). These systems were later refined through the emergence of example-based MT, which relied on bilingual corpora of parallel texts to guide the translation of new sentences (Somers, 1999). Subsequent progress led to statistical MT, leveraging large volumes of bilingual data to determine the most probable translation by analysing statistical relationships between source texts and their human-translated counterparts (Lopez, 2008). The expansion of statistical MT paved the way for neural MT (NMT), which employs neural networks to generate target-language output from source text. NMT systems can process vast datasets and operate with minimal supervision (Stahlberg, 2020), and many state-of-the-art translation tools today belong to this category. A timeline illustrating the evolution of these systems is presented in Figure 1.

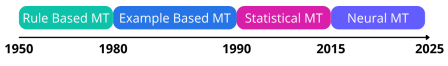

Figure 1: Evolution of MT systems.

However, existing NMT evaluation metrics exhibit several limitations. Most notably, they rely on reference translations for comparison, and often show limited correlation with human judgments. To better understand and improve this correlation, shared tasks such as WMT 2023 (Blain

et al., 2023) have been organized to develop more reliable and accurate evaluation metrics. Nevertheless, certain scenarios remain challenging for current approaches. A particularly prominent example is the translation of MWEs (Zaninello, 2020), which consist of two or more words whose combined meaning differs from the literal sum of their individual components. Addressing this phenomenon constitutes the central focus of the present study.

2.2 Multiword Expressions in MT

When discussing linguistic phenomena that strongly shape the idiosyncrasy of any language, MWEs are often among the most prominent examples, given their idiomatic richness and the deep cultural grounding that shape them. This complexity adds difficulty for language users in fully understanding their meanings and poses even greater challenges when attempting to accurately translate them into other languages (Almagsodi, 2025). MWEs constitute a heterogeneous set of linguistic units composed by two or more words that function as a single semantic unit exhibiting 'semantic idiomaticity'. This broad category includes a wide range of constructions, including idioms, collocations, sayings, and fixed expressions (Masini, 2019). Considering the diversity of linguistic structures encompassed by MWEs and the cultural influences that shape their meanings, it is unsurprising that they remain a significant challenge for MT systems seeking to improve the translation of these culturally rich expressions (Song and Xu, 2024; Zaitova et al., 2025). Their importance lies precisely in their cultural complexity—if MT systems can eventually identify appropriate cross-cultural equivalents of such expressions automatically, it would mark a significant advancement in their adaptative translation capabilities and cultural knowledge integration.

Numerous challenges arise when addressing the automatic translation of MWEs. These stem from their non-compositionality, where the meaning cannot be inferred by combining the definitions of their individual parts; their varying degrees of fixedness—for example, the fixed expression 'black and white' versus the more variable structure 'strong as an ox/horse/lion'); and their structural diversity, as some MWEs are syntactically irregular (e.g.,'by and large'), while others are more compositional and transparent (e.g., 'fresh air') (Miletić and Walde, 2024).

The Natural Language Processing (NLP) research community recognises the significant obstacles that MWEs pose for accurate automatic translation across different language pairs, largely due to their strong cultural and contextual dependencies. To address this challenge, many recent studies focus on developing techniques to improve automatic translation of these linguistic units and on creating high-quality MWE resources to train MT systems across a wide range of languages (Han et al., 2021; Garg et al., 2022; Hadj Mohamed et al., 2023; Dimakis et al., 2024; Ide et al., 2025).

However, the current state of the art in MT for MWEs reveals a lack of comprehensive research assessing the capabilities of contemporary NMT systems (Zaninello, 2020). Only a few studies have evaluated NMT performance on MWEs extracted from textual data (Rikters and Bojar, 2017; Corpas Pastor and Noriega-Santiáñez, 2024; Song and Xu, 2024) or have explored this linguistic phenomena in multimodal settings (Li et al., 2021). Research specifically focused on the Spanish-to-English language pair within the subtitling context using the latest MT systems is even scarcer. Only a few surveys analyse the performance of these systems on subtitling tasks—primarily with general text rather than MWEs (Koglin et al., 2023; Liang et al., 2024)—including longitudinal analyses of their evolution over the years for various languages (Etchegoyhen et al., 2014; Hagström and Pedersen, 2022; Karakanta, 2022).

3 Methodology

This section describes the methodology used to evaluate the translation of Spanish MWEs into English in a subtitling context. First, we present the the films that constitute the source corpus from which the MWEs were extracted. Next, we describe the automatic translation process applied to these expressions using the selected MT systems. Finally, we detail the evaluation framework adopted in this study, which combines automatic metrics, LLM-as-a-judge evaluation, and expert human assessment. An overview of the complete methodological approach is presented in Figure 2.

3.1 Data Collection

The decision to collect MWEs from Pedro Almodóvar's films stems from the director's dis-

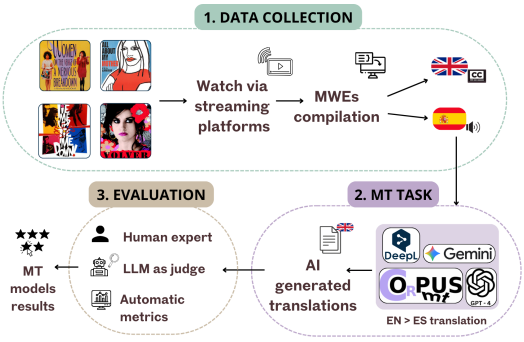

Figure 2: Methodology for evaluating MWEs in MT systems.

tinctive use of the Spanish language, which plays a crucial narrative and stylistic role in his works (Cintas and Remael, 2014). Almodóvar's preference for emotionally charged language in his films further encourages the presence of MWEs (Ciordia, 2016), employing language as a means of characterisation and identity construction. Given the considerable cultural and pragmatic weight of MWEs, their prevalence in Almodóvar's dialogues makes them a valuable resource for testing MT models on a linguistically and culturally demanding task. Considering the role language plays in defining his characters, we argue that the accurate translation of these MWEs can enhance the audiovisual experience and enrich the narrative across cultures by faithfully conveying character identity and dialectal nuances through non-standard expressions used in his films (Pérez, 2018).

Following this premise, four Almodóvar films were selected for this study: "Volver" (Almodóvar, 2006), "Tie me up! Tie me down!" (*"Átame"*) (Almodóvar, 1990), "All About My Mother" (*"Todo sobre mi madre"*) (Almodóvar, 1999), and "Women on the Verge of a Nervous Breakdown" (*"Mujeres al borde de un ataque de nervios"*) (Almodóvar, 1988). From these films we constructed ALMO-MWE, a Spanish–English lexicon of multiword expressions extracted from film subtitles. MWEs were manually identified by watching each film via official streaming platforms that provided both the original Spanish audio and the corresponding official English subtitles. In total, 235 MWEs were collected and organised into a parallel lexicon pairing their Spanish forms with their English equivalents—both the professional subtitle translations and the translations generated by the MT systems evaluated in this study, which are described in the following subsection.

3.2 Model Selection

To evaluate which model performs best in the task of MWE translation, we selected several of the most widely used models. Specifically, we included two task-agnostic LLMs that have achieved state-of-the-art results across multiple benchmarks: GPT-4 (Achiam et al., 2024) and Gemini (Team et al., 2025). These models have demonstrated strong performance in a wide range of tasks, including MT (Sinitsyna and Savenkov, 2024). In addition, we aimed to assess the effectiveness of a model specifically trained for MT. For this purpose, we selected OPUS-MT[1] (Tiedemann et al., 2023), using the version trained for Spanish–English translation. Finally, we included DeepL,[2] a widely used and popular translation tool that supports a broad range of languages.

For GPT-4[3] and Gemini,[4] the experiments were conducted through their publicly available interfaces. Since both models require a prompt to specify the task, we provided a concise instruction tailored for MT. The prompt used was: *"Translate the following sentence into English: [SENTENCE]"*.

In contrast, DeepL and OPUS-MT[5] did not require explicit prompts, as they are specifically designed for MT. These systems take the source sentence as input and automatically return the corresponding translation. More details about model versions can be seen in Appendix A.

It is important to note that comparisons between models are necessarily conducted at the system level, as proprietary tools such as DeepL do not disclose details of their internal processing pipelines.

3.3 Evaluation Methods

To evaluate how effectively models translate MWEs, we employed three complementary approaches: automatic metrics, LLMs as evaluative judges, and human assessment. This multifaceted design was intended to ensure robustness and reliability by validating results across different evaluation methodologies. This methodology also allowed us to analyse to what extent evaluation met-

[1] https://huggingface.co/Helsinki-NLP/opus-mt-es-en
[2] https://www.deepl.com/
[3] https://copilot.microsoft.com/
[4] https://gemini.google.com/
[5] The model was accessed via the Transformers library with the default parameter configuration.

rics correlate with human expert judgment when evaluating translation adequacy, to gain insights on which evaluation approaches can be beneficial when performing this type of comparative studies.

Automatic Metrics For the automatic evaluation, we selected a subset of metrics widely adopted within the MT community. These include both traditional surface-level measures and more recent embedding-based evaluation metrics. Specifically, we employed:

- **BLEU** (Papineni et al., 2002): a metric based on n-gram overlap between system outputs and reference translations.
- **NIST** (Doddington, 2002): an adaptation of BLEU that weights n-grams according to their informativeness, rather than treating all n-grams equally.
- **METEOR** (Lavie and Agarwal, 2007): a metric that extends beyond exact word matches by incorporating stemming and semantic similarity. It computes a harmonic mean of unigram precision and recall, assigning greated weight to recall than to precision.
- **BertScore** (Zhang et al., 2020): a semantic evaluation metric that leverages contextual embeddings from the BERT (Devlin et al., 2019) model to capture meaning beyond surface-level overlap.
- **COMET** (Rei et al., 2020): a recent metric using machine learning models to evaluate translations by incorporating information from both the source text and the target-language reference translation, providing a more-context aware estimation of translation quality.

Details of how we computed the metrics can be seen in Appendix A.

LLMs as Judges In addition to automatic metrics, we employed LLMs as evaluators to provide a more flexible and context-aware assessment of translation quality—particularly for MWEs. Previous studies have shown that automatic metrics often underperform when evaluated on translations allow for linguistic flexibility (Xiao et al., 2023; Martínez-Murillo et al., 2024). Moreover, recent research highlights that LLMs can serve as reliable evaluators, achieving accuracy levels comparable to human judgment (Gao et al., 2025). To this end, we used three state-of-the-art LLM models that have been used for evaluation purposes: Prometheus (Kim et al., 2024), GPT-oss (Agarwal et al., 2025) and Atla Selene mini (Alexandru et al., 2025). All models have demonstrated strong correlation with human assessments in generative evaluation settings.

Each evaluation model was provided with the source sentence, the system-generated translation, and the reference translation. For all models, the input also included evaluation rubrics consisting of a predefined 1–5 scale to guide the model's scoring process. These rubrics instructed models to assess translation adequacy, fluency, and fidelity to the original meaning, with particular emphasis on the accurate handling of MWEs.[6]

Human Evaluation A professional translator also conducted a manual evaluation of the generated translations by assessing their accuracy in the target language based on whether the original meaning of the MWEs in the Spanish source sentences was correctly understood and conveyed. To this end, the translator applied a binary classification, assigning a score of '1' to correct translations and '0' to incorrect ones that failed to identify the MWE or provided an inaccurate equivalent in English. The main evaluation criterion was to consider as correct those translations that successfully preserved the MWE original meaning, regardless of whether this was achieved through an equivalent English MWE or a more literal yet contextually adequate expression. We considered this approach more appropriate than restricting correctness only to cases involving equivalent MWEs, since the ultimate goal of translation is to preserve meaning and ensure adaptation to the target context. Following this criterion, the Spanish MWE '*¡Qué morrazo tienes, Pepa!*' would be correctly translated as either 'You're so cheeky, Pepa!' or 'You've got some nerve, Pepa!'. Conversely, an incorrect translation would be 'You've got a nose, Pepa!', which represents a literal rendering of the Spanish phrase '*morrazo*' ('nose') and ignores its idiomatic structure and non-compositionality. This literal approach results in a loss of meaning, as the Spanish MWE refers to someone being cheeky or

[6]Rubrics are available in the repository at `https://github.com/gplsi/ALMO-MWE`

bold–not to the English idiom 'to have a nose for something', which denotes perceptiveness or intuition.

4 Results & Evaluation

The automatically generated translations from each MT system, along with the original Spanish dialogues and the officially subtitled English versions created by a professional translator, are available in our public repository.[7] This repository comprises 940 translations corresponding to 235 original MWEs extracted from Almodóvar's films, accompanied by manual evaluations performed by a professional translator indicating whether each translation was correct or incorrect. With all MWEs translated automatically, the following subsections present the results obtained from the different evaluation approaches employed in this study.

4.1 Automatic & LLM as judge scores

The results derived from the different models using automatic evaluation metrics are summarised in Table 1. It reports the scores from both the automatic metrics and the LLM-based evaluations, which acted as evaluators across the four datasets. This dual approach was designed to analyse the degree of correlation between traditional automatic metrics and LLM-based evaluations, with the expectation that both methods would consistently identify the highest-performing model.

Metric	GPT-4	Gemini	OPUS-MT	DeepL	Ori_Trans
BLEU	0.326	0.258^{*}	0.251^{*}	0.282	-
METEOR	0.381	0.291^{*}	0.269^{*}	0.321^{*}	-
NIST	0.902	0.719	0.711	0.799	-
BERTSCORE	0.920	0.909^{*}	$0.902*$	$0.910*$	-
COMET	0.703	0.680^{*}	0.624^{*}	0.672^{*}	-
GPT-OSS	3.043	2.894^{*}	2.360^{*}	2.843^{*}	4.931^{*}
Prometheus	3.183	$3.077^{\dagger}$	2.953^{*}	$3.077^{\dagger}$	3.826^{*}
Atla Selene Mini	3.421	3.292^{*}	2.860^{*}	3.25^{*}	4.522^{*}

Table 1: Results obtained for the ALMO-MWE corpus. Paired T-test of the models was computed against GPT-4 scores, with statistical significance indicated by: * <0.05, † <0.15.

The results in Table 1 illustrate the outcomes of two complementary evaluation paradigms: reference-based automatic metrics (BLEU, METEOR, NIST, BERTScore, COMET) and LLM-as-a-judge evaluations (GPT-OSS, Prometheus, Atla Selene Mini). This dual evaluation approach allows us to contrast surface-level, reference-dependent measures with more holistic, context-aware assessments of translation quality.

A notable pattern emerges in the ranking of systems across these metrics. GPT-4 consistently achieves the best performance among the evaluated systems across both automatic metrics and LLMs-as-judges evaluations, underscoring its strong capabilities in MWE translation. In contrast, a smaller and more specialised MT system such as OPUS-MT yields the lowest overall performance among the four tested models in all automatic metrics and LLMs-as-judges, suggesting that lightweight or domain-limited systems struggle with the semantic and contextual complexity of MWEs.

A notable discrepancy emerges with respect to the second-best system. According to automatic metrics, DeepL ranks second, whereas in the LLM-as-a-judge evaluation, Gemini obtains the second-highest score. This divergence indicates that reference-based automatic metrics, particularly those relying on surface-level word overlap or semantic similarity to a single reference, are not always consistent with LLM judgments. Since many of the evaluated translations are not strictly literal, acceptable paraphrastic or contextually appropriate renderings may be penalised by automatic metrics despite being judged favourably by LLM evaluators.

More broadly, the results suggest that general-purpose LLMs, such as GPT-4 and Gemini, outperform specialised MT systems like DeepL and OPUS-MT when translating MWEs. One plausible explanation is that large, task-agnostic LLMs are trained on broader and more diverse corpora, enabling them to encode richer world knowledge and contextual representations. This broader knowledge base may facilitate better handling of non-literal meanings and idiomatic expressions, leading to more accurate and contextually appropriate translations.

Despite GPT-4's relative superiority, its performance remains far from optimal. Across both automatic metrics and LLM-based evaluations, it does not approach ceiling performance, averaging approximately 3 out of 5 in the judge-based assessment and obtaining relatively low scores on metrics such as BLEU, METEOR, and COMET. These findings highlight the persistent difficulty of MWE translation and emphasise the need for fur-

[7] https://github.com/gplsi/ALMO-MWE

ther research into systems specifically optimised for such phenomena.

Finally, we evaluated the human-produced translations using the LLM-as-a-judge framework to assess how closely each judge correlates with gold-standard outputs. As expected, GPT-OSS and Atla Selene Mini demonstrate high agreement, assigning near-perfect scores to the human translations. However, Prometheus assigns a lower average score of 3.8 out of 5. Given that human translations serve as the reference standard and are assumed to be of high quality, this discrepancy suggests that Prometheus does not fully correlate with human judgment in this evaluation setting and may apply stricter or differently calibrated assessment criteria.

To corroborate the outcomes, we conducted paired t-tests comparing each system against GPT-4 across all evaluation metrics. The tests were computed segment-wise, assessing whether the mean paired difference between each model and GPT-4 was statistically different from zero. Results reported that GPT-4 significantly outperforms with statistical significance Gemini, OPUS-MT, and DeepL across most automatic metrics. The only exception is NIST, where differences did not reach statistical significance. Regarding LLM-based evaluators (GPT-OSS, Prometheus, and Atla Selene Mini), most systems show statistically significant differences when compared to GPT-4. Notably, even the original human translation differs significantly from GPT-4 in several LLM-based metrics. Overall, the paired t-test analysis suggests that GPT-4 achieves statistically superior performance across the majority of automatic and neural evaluation metrics.

4.2 Human evaluation

The results of the professional translator's evaluation for the 235 MWEs are presented in Table 2. The percentages reported correspond to the proportion of MWEs whose translations were judged correct by the professional translator (see Section 3 for details of the evaluation protocol). As shown, GPT-4 achieves the highest proportion of correctly translated MWEs (77%), followed closely by Gemini with 75%. This ranking is broadly consistent with the tendencies observed in the automatic evaluation, particularly in the LLM-as-a-judge approaches, which also identified GPT-4 as the top-performing system and Gemini as the second best.

By contrast, the human evaluation reveals a noticeable performance gap between these two systems and DeepL. Although DeepL has been widely reported as a competitive MT system, its performance in translating MWEs in this dataset is substantially lower than that of GPT-4 and Gemini. This finding contrasts with the results suggested by several automatic metrics (see Table 1), which tend to assign comparatively higher scores to DeepL outputs. Finally, OPUS-MT achieves less than 40% correctly translated MWEs, clearly emerging as the weakest system among those evaluated.

	GPT-4	Gemini	DeepL	OPUS-MT	Total MWEs
MWEs accurately translated	182 (77%)	177 (75%)	152 (64%)	87 (37%)	235

Table 2: Professional translator evaluation of MWEs translation with each MT system (% of correct answers expressed in parenthesis).

4.3 Metric - Human correlation

The discrepancies observed between automatic evaluation scores and professional human judgments motivate a closer examination of the extent to which current translation evaluation metrics correlate with expert assessments. In this section, we therefore analyse how reliably these metrics reflect the criteria used by professional translators when evaluating the adequacy of MWE translations in an audiovisual subtitling context.

To systematically assess the extent to which automatic evaluation metrics reflect expert human judgments, we computed two complementary statistical indicators: point-biserial Pearson correlation (r) and the Area Under the ROC Curve (AUC), reported in Table 3. Pearson's r measures the strength of the linear association between metric scores and the binary human evaluation labels (0/1), indicating whether higher metric scores tend to correspond to translations judged as correct. Since the human evaluation is binary, we additionally report AUC, which measures the discriminative capacity of each metric—namely, the probability that the metric assigns a higher score to a correct translation than to an incorrect one. Confidence intervals for both statistics were estimated using non-parametric bootstrap resampling (5,000 iterations).

The results reveal a clear hierarchical pattern

Metric	Pearson r	95% CI	AUC	95% CI
BLEU	0.017	[-0.047, 0.082]	0.646	[0.610, 0.681]
METEOR	0.003	[-0.066, 0.060]	0.665	[0.630, 0.699]
NIST	0.007	[-0.063, 0.060]	0.625	[0.588, 0.661]
BERTScore	0.336	[0.276, 0.392]	0.723	[0.689, 0.755]
COMET	**0.480**	[0.430, 0.527]	**0.792**	[0.763, 0.820]
GPT-OSS	**0.508**	[0.459, 0.555]	**0.800**	[0.771, 0.827]
Prometheus	0.261	[0.200, 0.320]	0.645	[0.612, 0.676]
Selene	0.487	[0.436, 0.533]	0.771	[0.743, 0.797]

Table 3: Correlation between automatic evaluation metrics and human translator judgments in MWE translation.

among evaluation metrics. Traditional n-gram-based metrics (BLEU, NIST, and METEOR) exhibit near-zero global Pearson correlations, with confidence intervals overlapping zero. This indicates the absence of a stable linear relationship between surface-level overlap and expert-assessed translation adequacy in MWE subtitling. Although their AUC values remain slightly above chance (0.62–0.67), suggesting limited discriminative ability, their correlation with human evaluation is weak and inconsistent. This finding reinforces the long-standing concern that surface-level lexical overlap is insufficient for evaluating translations of non-compositional and culturally bound expressions such as MWEs.

In contrast, embedding-based metrics demonstrate substantially stronger consistency. BERTScore achieves a moderate global correlation ($r \approx 0.34$) and improved discriminative capacity ($AUC \approx 0.72$), indicating that contextual semantic similarity better reflects human adequacy judgments. However, the most significant improvement emerges with supervised neural metrics and LLM-based evaluators. COMET reaches a global correlation of approximately 0.48 and an AUC close to 0.79, reflecting strong and stable agreement with expert assessments.

Among the LLM-as-judge approaches, GPT-OSS achieves the strongest correspondence with expert human judgments ($r \approx 0.51$, $AUC \approx 0.80$), slightly outperforming COMET. Selene follows closely, while Prometheus displays substantially lower correlation and discriminative performance ($r \approx 0.26$, $AUC \approx 0.65$), performing closer to traditional automatic metrics than to other LLM-based evaluators. This divergence highlights that not all LLM-as-judge systems exhibit equivalent effectiveness as evaluators, and that model architecture and evaluation design may play a crucial role in agreement with expert human criteria.

Overall, the results suggest that evaluation metrics incorporating contextual semantic modelling—either through supervised quality estimation (COMET) or LLM-based judgment—are substantially better suited for assessing culturally nuanced translation phenomena such as MWEs. In contrast, n-gram-based metrics show limited capacity to account for the interpretative and pragmatic considerations that professional translators apply in audiovisual subtitling contexts. However, although GPT-OSS and COMET exhibit markedly stronger correlations with expert judgments than traditional metrics, these correlations remain moderate rather than strong. This indicates that, while recent evaluation approaches represent a significant step forward, current automatic metrics still do not adequately capture the complex criteria that professional translators employ when evaluating idiomatic and culturally embedded MWEs.

5 Qualitative Analysis of Metric-Human Disagreement

While the previous sections quantitatively examined MT system performance and the correlation between automatic metrics and expert human evaluation, a qualitative analysis helps clarify how these metrics behave when evaluating the translation of Spanish MWEs into English and which signals they prioritize when assigning scores.

A comparison of the results in Tables 1 (automatic metrics) and 2 (human evaluation) reveals a clear discrepancy between automatic and human assessments for certain systems. In particular, OPUS-MT receives comparatively favorable scores from several automatic metrics despite performing poorly according to the professional translator. Human evaluation shows that OPUS-MT correctly translates fewer than 40% of the 235 MWEs in the dataset, whereas GPT-4 and Gemini exceed 70%. Nevertheless, the scores assigned by several metrics often differ only marginally from those of the best-performing systems.

One explanation for this discrepancy lies in the inherent translational asymmetry of MWEs. Such asymmetry frequently manifests as one-to-many correspondences, where a single source-language expression may be translated by several valid target-language equivalents, or more complex many-to-many mappings between semantically related MWEs (Constant et al., 2017; Monti et al., 2020). In our dataset, many expressions follow the one-to-many pattern, meaning that multi-

Original sentence	Reference translation	System	Translation	BLEU	METEOR	NIST	BERTScore	COMET	Prometheus	GPT OSS	Selene	Human
1. *Con que me dejes en el centro* ***me viene bárbaro***.	If you drop me in the centre that'd be great.	**OPUS**	With you leaving me in the center I come barbarian.	0.400	0.375	1.328	0.872	0.408	3	1	2	×
		GPT	If you drop me off in the center, it's perfect for me.	0.500	0.577	1.661	0.923	0.771	4	4	4	✓
2. *Así tengo la impresión de que no ha* ***pasado el tiempo***.	It's as if time has stood still.	**OPUS**	So I get the impression that time has not passed.	0.200	0.256	0.561	0.887	0.631	3	3	3	✓
		GPT	That way I get the impression that time hasn't passed.	0.100	0.068	0.280	0.900	0.687	3	3	3	✓
3. *Bueno,* ***al loro con*** *la puerta.*	Keep an eye on the door.	**OPUS**	Well, the parrot with the door.	0.333	0.312	0.861	0.881	0.400	3	1	1	×
		GPT	Well, watch out with the door.	0.333	0.312	0.861	0.910	0.603	3	2	3	✓
4. *¡Pues* ***estáis buenos*** *los dos!*	You're quite a pair.	**OPUS**	Well, you're both good!	0	0.125	0.500	0.882	0.689	3	3	3	×
		GPT	Well, you two are quite a pair!	0.285	0.436	0.571	0.934	0.901	4	4	4	✓
5. ***Estoy harta de*** *ser buena.*	I'm sick of being good.	**OPUS**	I'm sick of being good.	0.800	0.793	1.857	0.984	0.976	3	5	4	✓
		GPT	I'm tired of being good.	0.800	0.750	1.857	0.994	0.963	3	4	4	✓
6. *Y* ***mira la que armaste***.	And look at the fuss you caused!	**OPUS**	And look at the one you armed.	0.714	0.691	2.005	0.916	0.468	3	1	2	×
		GPT	And look at the mess you made.	0.714	0.691	2.005	0.948	0.661	4	4	3	✓

Table 4: Comparison of scores obtained for translations generated by OPUS-MT and GPT-4 across several automatic metrics, LLM-as-judge evaluators, and human evaluation.

ple English renderings can legitimately convey the intended meaning. During human evaluation, lexical divergence from the reference translation was therefore not penalized; instead, adequacy judgments were based on whether the translation preserved the intended meaning in context. By contrast, most automatic metrics primarily rely on lexical overlap or similarity-based semantic signals when comparing candidate translations with reference sentences.

To illustrate how these limitations affect the evaluation of MWE translation, Table 4 presents several representative examples comparing translations generated by the weakest-performing system (OPUS-MT) and the strongest-performing system (GPT-4). This comparison allows us to observe both cases in which evaluation metrics fail to detect incorrect translations and cases in which they correctly reward adequate renderings.

With the exception of the second and fifth MWEs (i.e., *pasar el tiempo* and *estar harta de*), OPUS-MT generally fails to translate the expressions accurately, producing literal word-for-word renderings rather than conveying the intended idiomatic meaning. Despite this poor performance, several automatic metrics assign relatively similar scores to both correct and incorrect translations. For example, BERTScore yields nearly identical scores across the MWEs shown in Table 4, sometimes assigning higher scores to incorrect translations than to accurate ones (see MWEs 2 and 6). A comparable pattern appears for COMET, which occasionally assigns similar scores to both correct and incorrect renderings (e.g., MWEs 2 and 4). In contrast, BLEU and METEOR show inconsistent behaviour, sometimes assigning low scores to correct translations while producing their highest scores for other correct MWEs.

A direct comparison between GPT-4 and OPUS-MT further highlights the limited discriminative capacity of several metrics. With the exception of COMET, most metrics produce very similar scores for both systems even when translation quality differs substantially according to the human evaluation (see MWEs 1, 2, 3 and 6). In these cases, GPT-4 generates translations that correctly convey the intended meaning, whereas OPUS-MT produces literal or semantically inaccurate renderings. Nevertheless, the resulting metric scores differ only marginally, suggesting that metrics relying primarily on lexical overlap or embedding similarity struggle to clearly separate adequate translations from incorrect ones when some degree of surface or semantic similarity with the reference translation is present.

The LLM-as-judge evaluators show different levels of sensitivity to translation quality. Prometheus displays the weakest discriminative

behaviour: although it does not assign identical scores to all translations, it systematically gives OPUS-MT a score of 3 across the examples, regardless of whether the translation is judged correct or incorrect by the human evaluator. In contrast, GPT-OSS and Selene show a more coherent evaluation pattern, assigning higher scores to GPT-4 than to OPUS-MT (see MWEs 1, 4 and 6). However, these judges do not fully correlate with human evaluation in all cases. For instance, MWE 2 is judged correct for both systems by the human evaluator, yet GPT-OSS and Selene assign identical mid-range scores to both translations. Overall, these results suggest that LLM-as-judge evaluators—particularly GPT-OSS and Selene—provide a closer approximation to human judgments than traditional automatic metrics, although they still do not fully replicate expert evaluation criteria.

In light of these results, it becomes clear that the computational principles underlying many evaluation metrics—such as n-gram overlap or embedding-based similarity—are only partially capable of capturing the cultural and contextual knowledge required to accurately assess MWE translation. Although LLM-based evaluators represent a promising direction for improving MT evaluation, our findings indicate that they still do not fully align with professional translators.

Consequently, relying exclusively on automatic MT evaluation metrics for complex tasks such as subtitle translation entails a substantial methodological risk. Audiovisual translation requires sensitivity to contextual interpretation, cultural references, and pragmatic adequacy, all of which are closely tied to the communicative function of MWEs in situated language use. These results therefore highlight the need for hybrid evaluation frameworks that combine automatic metrics with expert human assessment when analysing culturally complex translation phenomena.

6 Conclusions and Future Work

This study pursued a twofold objective: first, to examine the performance of state-of-the-art MT systems in translating Spanish MWEs into English; and second, to evaluate the extent to which current automatic evaluation metrics can capture the pragmatic and cultural nuances associated with such linguistic phenomena. To this end, we constructed a bilingual parallel MWE lexicon extracted from a corpus of four Pedro Almodóvar films and used it to evaluate several prominent LLMs and specialised MT systems. The resulting translations were assessed through three complementary approaches: automatic evaluation metrics, LLM-as-a-judge evaluation, and expert human assessment.

The results show that none of the evaluated MT systems achieved consistently satisfactory performance in translating MWEs. Moreover, the comparison between automatic metrics and human judgments revealed important discrepancies, highlighting the limitations of traditional evaluation approaches for culturally and pragmatically complex translation phenomena. Among the evaluation methods considered, GPT-OSS demonstrated the highest correlation with human judgments, suggesting that LLM-based evaluation may represent a promising direction for improving automatic assessment of nuanced translation tasks.

These findings further underscore the continuing importance of professional human translators in contexts involving complex linguistic and cultural phenomena such as MWEs. In particular, several MT systems struggled with expressions containing colloquial, offensive, or culturally sensitive language. For instance, Gemini failed to translate certain MWEs containing insults or obscene terms, illustrating current limitations in handling pragmatically marked expressions. More broadly, our results indicate that traditional automatic evaluation metrics often fail to capture the criteria that professional translators apply when assessing the adequacy of idiomatic translations.

Future research may extend this work in several directions. First, the analysis could be expanded to additional language pairs in order to examine how translational asymmetries of MWEs manifest across different linguistic combinations. Second, since the MWEs analyzed in this study were extracted from audiovisual content, an especially promising avenue involves exploring the potential of multimodal models that integrate visual context to support more accurate interpretation and translation of MWEs in subtitling scenarios.

Acknowledgements

This research is funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project Desarrollo Modelos ALIA (Resol. SEDIA 19.08.2024).

Carbon Impact Statement

The computational experiments in this study can be divided into two parts. For the models used to generate translations, we relied on publicly available platforms, and therefore, no carbon impact data is available.

For the evaluation using LLMs as judges, each judge required approximately 1.5 hours to evaluate all machine translation models (4.5 hours in total). These evaluations were performed on a single NVIDIA DGX A100 GPU, with an estimated power consumption of 0.88 kW.

Assuming electricity consumption from the Spanish grid, with an approximate carbon intensity of 0.113 $kgCO_2eq$ per kWh, the total emissions are estimated at 0.1 $kgCO_2eq$, with no direct offsets applied.

These estimates were calculated using the MachineLearningCO_2 Impact calculator (Lacoste et al., 2019).[8]

References

Achiam, Josh, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2024. GPT-4 technical report. https://arxiv.org/abs/2303.08774.

Agarwal, Sandhini, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Alexandru, Andrei, Antonia Calvi, Henry Broomfield, Jackson Golden, Kyle Dai, Mathias Leys, Maurice Burger, Max Bartolo, Roman Engeler, Sashank Pisupati, Toby Drane, and Young Sun Park. 2025. Atla selene mini: A general purpose evaluation model.

Almagsodi, Aysha. 2025. Idioms across cultures: A corpus-based study of translation strategies in Amy Tan's novels. *International Journal of Linguistics, Literature and Translation*, 8(9):146–158.

Almodóvar, Pedro. 1988. Mujeres al borde de un ataque de nervios. Motion picture. Spain: El Deseo S.A. English title: Women on the Verge of a Nervous Breakdown.

Almodóvar, Pedro. 1990. ¡Átame! Motion picture. Spain: El Deseo S.A. English title: Tie Me Up! Tie Me Down!

Almodóvar, Pedro. 1999. Todo sobre mi madre. Motion picture. Spain: El Deseo S.A. English title: All About My Mother.

Almodóvar, Pedro. 2006. Volver. Motion picture. Spain: El Deseo S.A. English title: Volver.

Bel'skaja, Izabella K. 1957. Machine translation of languages. *Research*, 10(10):383–389.

Blain, Frederic, Chrysoula Zerva, Ricardo Rei, Nuno M. Guerreiro, Diptesh Kanojia, José G. C. de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh Azadi, Constantin Orasan, and André Martins. 2023. Findings of the WMT 2023 shared task on quality estimation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 629–653, Singapore, December. Association for Computational Linguistics.

Cintas, Jorge Díaz and Aline Remael. 2014. *Audiovisual translation: subtitling*. Routledge.

Ciordia, Leticia Santamaría. 2016. A contrastive and sociolinguistic approach to the translation of vulgarity from Spanish into English and Polish in the film Tie Me Up! Tie Me Down! (Pedro Almodóvar, 1990). *Translation and Interpreting Studies. The Journal of the American Translation and Interpreting Studies Association*, 11(2):287–305.

Constant, Mathieu, 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.

Corpas Pastor, Gloria and Laura Noriega-Santiáñez. 2024. Human versus neural machine translation creativity: A study on manipulated MWEs in literature. *Information*, 15(9):530.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, Jill, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Dimakis, Antonios, Stella Markantonatou, and Antonios Anastasopoulos. 2024. Dictionary-aided translation for handling multi-word expressions in low-resource languages. In Ku, Lun-Wei, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2588–2595, Bangkok, Thailand, August. Association for Computational Linguistics.

Doddington, George. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second*

[8] https://mlco2.github.io/impact#compute

International Conference on Human Language Technology Research, HLT '02, page 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Etchegoyhen, Thierry, Lindsay Bywood, Mark Fishel, Panayota Georgakopoulou, Jie Jiang, Gerard van Loenhout, Arantza del Pozo, Mirjam Sepesy Maučec, Anja Turner, and Martin Volk. 2014. Machine translation for subtitling: A large-scale evaluation. In Calzolari, Nicoletta, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 46–53, Reykjavik, Iceland. European Language Resources Association (ELRA).

Gao, Mingqi, Xinyu Hu, Xunjian Yin, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2025. LLM-based NLG evaluation: Current status and challenges. *Computational Linguistics*, 51(2):661–687, 06.

Garg, Kamal Deep, Shashi Shekhar, Ajit Kumar, Vishal Goyal, Bhisham Sharma, Rajeswari Chengoden, and Gautam Srivastava. 2022. Framework for handling rare word problems in neural machine translation system using multi-word expressions. *Applied Sciences*, 12(21):11038.

Hadj Mohamed, Najet, Malak Rassem, Lifeng Han, and Goran Nenadic. 2023. AlphaMWE-Arabic: Arabic edition of multilingual parallel corpora with multiword expression annotations. In Mitkov, Ruslan and Galia Angelova, editors, *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 448–457, Varna, Bulgaria, September. INCOMA Ltd., Shoumen, Bulgaria.

Hagström, Hanna and Jan Pedersen. 2022. Subtitles in the 2020s: The influence of machine translation. *Journal of Audiovisual Translation*, 5(1):207–225.

Han, Lifeng, Gareth Jones, Alan Smeaton, and Paolo Bolzoni. 2021. Chinese character decomposition for neural MT with multi-word expressions. In Dobnik, Simon and Lilja Øvrelid, editors, *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 336–344, Reykjavik, Iceland (Online), May 31–2 June. Linköping University Electronic Press, Sweden.

Ide, Yusuke, Joshua Tanner, Adam Nohejl, Jacob Hoffman, Justin Vasselli, Hidetaka Kamigaito, and Taro Watanabe. 2025. CoAM: Corpus of all-type multiword expressions. In Che, Wanxiang, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27004–27021, Vienna, Austria, July. Association for Computational Linguistics.

Karakanta, Alina. 2022. Experimental research in automatic subtitling: At the crossroads between machine translation and audiovisual translation. *Translation spaces*, 11(1):89–112.

Kim, Seungone, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. In Al-Onaizan, Yaser, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, Miami, Florida, USA, November. Association for Computational Linguistics.

Koglin, Arlene, Willian Henrique Cândido Moura, Morgana Aparecida De Matos, and João Gabriel Pereira da Silveira. 2023. Quality assessment of machine-translated post-edited subtitles: an analysis of Brazilian translators' perceptions. *Linguistica Antverpiensia, New Series–Themes in Translation Studies*, 22.

Lacoste, Alexandre, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Lavie, Alon and Abhaya Agarwal. 2007. Meteor: an automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, page 228–231, USA. Association for Computational Linguistics.

Li, Jiaoda, Duygu Ataman, and Rico Sennrich. 2021. Vision matters when it should: Sanity checking multimodal machine translation models. In Moens, Marie-Francine, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8556–8562, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Liang, Chuqiao, Mozhgan Ghassemiazghandi, and Marlina Jamal. 2024. Post-editing challenges in Chinese-to-English neural machine translation of movie subtitles. *Social Sciences & Humanities Open*, 10:100949.

Lopez, Adam. 2008. Statistical machine translation. *ACM Computing Surveys*, 40(3):1–49, August.

Martínez-Murillo, Iván, Paloma Moreda, and Elena Lloret. 2024. Analysing the problem of automatic evaluation of language generation systems. *Procesamiento del Lenguaje Natural*, 72(0):123–136.

Masini, Francesca. 2019. Multi-word expressions and morphology. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press.

Miletić, Filip and Sabine Schulte im Walde. 2024. Semantics of multiword expressions in transformer-based models: A survey. *Transactions of the Association for Computational Linguistics*, 12:593–612.

Monti, Johanna, Mihael Arcan, and Federico Sangati. 2020. Translation asymmetries of multiword expressions in machine translation: An analysis of the TED-MWE corpus. In *Computational Phraseology*, pages 23–42. John Benjamins Publishing Company.

Papineni, Kishore, 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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Pérez, Francisco Javier Díaz. 2018. Language and identity representation in the English subtitles of Almodóvar's films. *Cultus*, 2035:96–121.

Ranathunga, Surangika, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11), February.

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In Webber, Bonnie, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, November. Association for Computational Linguistics.

Rikters, Matīss and Ondřej Bojar. 2017. Paying attention to multi-word expressions in neural machine translation. In Kurohashi, Sadao and Pascale Fung, editors, *Proceedings of Machine Translation Summit XVI: Research Track*, pages 86–95, Nagoya Japan, September 18 – September 22.

Sinitsyna, Daria and Konstantin Savenkov. 2024. Comparative evaluation of large language models for linguistic quality assessment in machine translation. In Martindale, Marianna, Janice Campbell, Konstantin Savenkov, and Shivali Goel, editors, *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 2: Presentations)*, pages 154–183, Chicago, USA, September. Association for Machine Translation in the Americas.

Somers, Harold. 1999. Example-based machine translation. *Machine translation*, 14(2):113–157.

Song, Huacheng and Hongzhi Xu. 2024. Benchmarking the performance of machine translation evaluation metrics with Chinese multiword expressions. In Calzolari, Nicoletta, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2204–2216, Torino, Italia, May. ELRA and ICCL.

Stahlberg, Felix. 2020. Neural machine translation: A review. *Journal of Artificial Intelligence Research*, 69:343–418.

Team, Gemini, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2025. Gemini: A family of highly capable multimodal models. `https://arxiv.org/abs/2312.11805`.

Tiedemann, Jörg, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato Yves Scherrer, Raul Vazquez, and Sami Virpioja. 2023. Democratizing neural machine translation with OPUS-MT. *Language Resources and Evaluation*, (58):713–755.

Xiao, Ziang, Susu Zhang, Vivian Lai, and Q. Vera Liao. 2023. Evaluating evaluation metrics: A framework for analyzing NLG evaluation metrics using measurement theory. In Bouamor, Houda, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10967–10982, Singapore, December. Association for Computational Linguistics.

Zaitova, Iuliia, Vitalii Hirak, Badr M. Abdullah, Dietrich Klakow, Bernd Möbius, and Tania Avgustinova. 2025. Attention on multiword expressions: A multilingual study of BERT-based models with regard to idiomaticity and microsyntax. In Chiruzzo, Luis, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4083–4092, Albuquerque, New Mexico, April. Association for Computational Linguistics.

Zaninello, Andrea. 2020. Multiword expression aware neural machine translation. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 3816–3825, Marseille, France. European Language Resources Association (ELRA).

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *Proceedings of the Eighth International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia. Curran Associates, Inc.

A Models and Metrics Details

Model Versions. We used the following models in our experiments: `gemini-2.5-pro`, `gpt-4-0613`, and `DeepL`, all queried in February 2025. In addition, we used `opus-mt-es-en` with default inference parameters from the HuggingFace Transformers library.

Evaluation Metrics. Following SacreBLEU recommendations, we report multiple automatic evaluation metrics. BLEU-1, METEOR, and NIST scores are computed using the NLTK python package with version 3.9.2, with tokenisation performed by whitespace. We also report BERTScore (version 0.3.13) and COMET (version 2.2.7, `Unbabel/wmt22-comet-da`).