Evaluating the Impact of Verbal Multiword Expressions on Machine Translation

Anonymous ACL submission

Abstract

Verbal multiword expressions (VMWEs) present significant challenges for natural language processing due to their complex and often non-compositional nature. While machine translation models have seen significant improvement with the advent of language models in recent years, accurately translating these complex linguistic structures remains an open problem. In this study, we analyze the impact of three VMWE categories—verbal idioms, verb-particle constructions, and light verb constructions—on machine translation quality from English to multiple languages. Using established multiword expression datasets and standard machine translation datasets, we evaluate how state-of-the-art translation systems handle these expressions. Our experimental results consistently show that VMWEs negatively affect translation quality, with deeper analysis indicating that this degradation is primarily attributable to the VMWE itself rather than general sentence-level difficulty.

1 Introduction

Machine translation (MT) has achieved remarkable progress with the advent of neural architectures and large language models (Lu et al., 2024; Team et al., 2024), yet certain linguistic phenomena continue to challenge advanced systems, such as structural differences (Wang et al., 2022) and morphologically complex words (Weller-Di Marco and Fraser, 2022). *Verbal multiword expressions* (VMWEs)—fixed or semi-fixed combinations of words containing a verb with non-compositional meanings—represent one of these persistent challenges. Unlike regular word combinations whose meanings derive from their constituent parts, multiword expressions require contextual understanding and often language-specific knowledge to translate accurately. For example, when reading "*spill the beans*," humans understand this idiom means to reveal a secret, yet translation systems render it literally leading to loss of the intended meaning, as depicted in Figure 1. Similarly, when people "*land on their feet*" after a setback, or "*check in*" at a hotel, the meanings emerge from the word combinations rather than individual components. Despite their prevalence in everyday language and their importance for meaning preservation in translation, VMWEs remain understudied in the context of MT systems (Ramish et al., 2013; Cap et al., 2015).

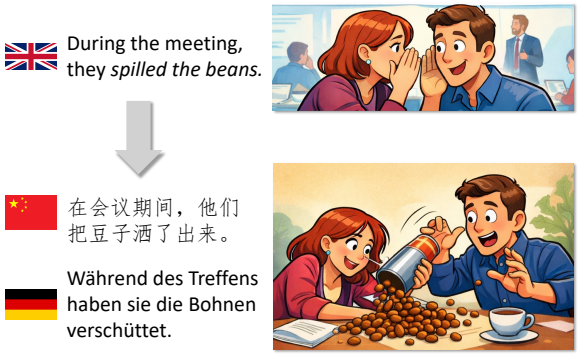

Figure 1: Illustrative example of incorrect machine translation of a sentence containing the verbal multiword expression *spill the beans* ("reveal a secret"), where literal translations into Chinese and German lead to a semantically incorrect interpretation.

Verbal multiword expressions have long been recognized as an important challenge in NLP (Fellbaum et al., 2006; Boukobza and Rappoport, 2009; Rohanian et al., 2020). Following previous work on the most widely recognized verbal multiword expressions (Constant et al., 2017; Savary et al., 2017), we focus on three linguistically distinct VMWE types: (1) verbal idioms, whose meanings cannot be derived from component words (e.g., *take the lion's share*), (2) verb-particle constructions, where particles modify verb meanings in non-trivial ways (e.g., *figure out*), and (3) light verb constructions, where verbs primarily serve grammatical functions while nouns carry more semantic weight (e.g., *give a speech*).

In this work, we present a comprehensive analysis of how these different categories of VMWEs affect translation quality across multiple state-of-the-art MT systems. Our investigation employs a novel dual-dataset approach, examining both established VMWE datasets (e.g., Saxena and Paul, 2020; Tu and Roth, 2012) and VMWE sentences extracted from the WMT (Workshop on Machine Translation) datasets (Koehn and Monz, 2006; Kocmi et al., 2024). Specifically, we use large language models to extract sentences containing verbal multiword expressions from the WMT datasets. We evaluate translation quality through two complementary methods: (1) reference-free quality estimation models (Juraska et al., 2024; Guerreiro et al., 2024) for VMWE datasets, which predict translation adequacy and fluency without relying on human-generated reference translations, and (2) human direct assessment (DA) scores for WMT data, providing insights into human perception of translation quality. We conduct experiments on seven language pairs and eight state-of-the-art MT systems: four specific neural MT models, three LLM-based models and Google Translate API. Our experimental results demonstrate that verbal multiword expressions consistently degrade translation quality, with the degree of degradation correlating with non-compositionality. We perform deeper analysis and show that the presence of the VMWE is the primary cause for the translation degradation rather than confounding factors such as linguistic complexity. In summary, the key contributions are:

- We provide the first comprehensive study of the impact of verbal multiword expressions (VMWEs) on machine translation (MT) quality, using both VMWE-specific datasets and MT datasets, with automatic evaluation methods and human scores.
- Our experimental results quantitatively demonstrate that VMWEs consistently degrade translation quality across multiple MT systems and language pairs, with the degree of degradation correlating with the level of non-compositionality.
- We provide a fine-grained analysis of VMWE-related translation difficulty using quality estimation and regression-based controls, showing that the observed degradation is primarily attributable to VMWE presence rather than standard sentence-level difficulty factors.

All our code and experimental results will be made publicly available upon acceptance.

2 Related Work

Verbal Multiword Expressions in NLP. Multiword expressions (MWEs) have long been recognized as important challenges in natural language processing due to their idiosyncratic, non-compositional, and often ambiguous nature (Sag et al., 2002; Baldwin and Kim, 2010; Constant et al., 2017). Among MWEs, verbal multiword expressions (VMWEs) such as light verb constructions, verb-particle constructions, and verbal idioms have drawn considerable attention (Ramisch, 2014; Pasquer et al., 2020; Rohanian et al., 2020; Savary et al., 2023). These observations have led to considerable work toward constructing annotations and datasets for individual VMWE types. For example, Tu and Roth (2011) and Tu and Roth (2012) compiled datasets for English light verbs and verb-particles. Saxena and Paul (2020) and Haagsma et al. (2020) introduced datasets of idiomatic expressions. VMWEs have also been widely studied for their cross-linguistic variability, as languages differ both in the lexicon and syntactic patterns employed to express them (Cordeiro and Candito, 2019; Rácz et al., 2014; Khoshtab et al., 2025). While English uses light verbs like *make* and *give*, other languages may use unique verbs not equivalent in English for similar constructions (Butt, 2010). Other studies have explored idiomatic expressions of low resource languages (Dimakis et al., 2024; Ohuoba et al., 2024). Annotation efforts like the PARSEME shared task (Savary et al., 2017, 2023) have standardized VMWE identification across languages.

MWE in Machine Translation. Machine Translation (MT) has seen significant progress with the transition from phrase-based statistical MT (Koehn et al., 2003; Koehn and Knowles, 2017) to neural MT (Bahdanau et al., 2016; Vaswani et al., 2017). More recently, large language models (LLMs) have been applied to MT tasks, bringing improvement in fluency and contextual understanding (Manakhimova et al., 2023; Kocmi et al., 2024). Researchers have known for a long time that MWE is essential for MT systems. Earlier work by Carpuat and Diab (2010) examined idiom translation in statistical MT and highlighted the mismatch between compositional meaning and phrase-based alignment. With the advent of neural MT, numerous studies have focused on improving MWE translation with neural models (Koehn and Knowles, 2017; Zaninello and Birch, 2020; Baziotis et al., 2023). More re-

cently, Song and Xu (2024) analyzed the effects of MWEs and named entities in Chinese–English MT, finding that errors involving idioms significantly affect overall quality. However, despite these advances, there remains a notable lack of systematic evaluation of VMWEs in state-of-the-art machine translation systems. Our work provides the first comprehensive study of the impact of verbal multiword expressions (VMWEs) on machine translation (MT) quality.

3 Evaluating Translation on VMWE Datasets

Here we detail our use of existing VMWE datasets: the VMWE categories under study, the datasets themselves, the MT systems we evaluate, and the metrics used to assess performance. We use the MT systems to translate the VMWE datasets to the target languages. We then get the translation score with quality evaluation (QE) models.

3.1 Verbal Multiword Expression Types

Following previous research (Constant et al., 2017; Savary et al., 2017), we focus on three types of common verbal multiword expressions in English: verbal idioms, verb-particle constructions and light verb constructions. These categories challenge natural language processing due to their semantic opacity, context dependence and idiomaticity.

Verbal idiom (VID): A fixed phrase with a verb whose meaning cannot be guessed from the individual words (e.g., *spill the beans* means to *reveal the secret*). These expressions are highly non-compositional and often require contextual or cultural understanding.

Verb-Particle Construction (VPC): Consists of a verb and a particle (typically an adverb or preposition) that together form a new meaning (e.g., *give up* means *quit*). These expressions are generally semi-compositional, as the meaning is not fully predictable from the parts.

Light Verb Construction (LVC): Combines a verb with little meaning on its own (e.g., *take*) and a predicative noun (e.g., *walk*) to form a full expression (e.g., *take a walk*). These constructions are semi-compositional, with the noun carrying most of the meaning.

3.2 Datasets

Verbal multiword expressions have long been recognized as a challenge in natural language processing. Therefore, there have been multiple corpora dedicated to VMWE and its subcategories.

Verbal Idiom datasets. We use these two datasets: 1) EPIE (Saxena and Paul, 2020): This dataset contains annotated formal idiomatic English sentences extracted from the British National Corpus (BNC) (Leech, 1992). 2) MAGPIE (Haagsma et al., 2020): A large sense-annotated corpus based on the BNC and Parallel Meaning Bank (PMB) (Abzianidze et al., 2017). We retain only idiom sentences from these two corpora that include a verb. This filtering step ensures our dataset accurately reflects verbal idiomatic expressions. After filtering, we sample 2,000 sentences, 1,000 from each corpus.

Verb-Particle Construction datasets. We use the Tu and Roth (2012) VPC dataset consisting of 1,348 sentences sampled from BNC corpus, of which 876 sentences with true verb-particle construction and 472 sentences with simplex verb-preposition combination sentences, focusing on combinations of six verbs *do*, *get*, *give*, *have*, *make* and *take* with nineteen common prepositions or particles. We only keep the 876 sentences that contain the true verb-particle construction.

Light Verb Construction datasets. We use the Tu and Roth (2011) LVC dataset. It is a balanced, annotated dataset of English light verb constructions using the BNC, includes 1,039 true light verb construction sentences and 1,123 simplex verb-noun sentences. This dataset focuses on the same six common English light verbs used in the VPC dataset. We used all 1,039 true light verb construction sentences for evaluation.

Non-VMWE Sentences. To accurately assess the impact of VMWEs on translation quality, we constructed a comparison dataset consisting of sentences without verbal multiword expressions. We obtained these non-VMWE sentences by applying rule-based heuristics to filter the BNC. Verbal idioms are excluded using idiom dictionaries from the EPIE and MAGPIE corpora. Verb-particle constructions are filtered using the spaCy dependency parser (Honnibal et al., 2020) where a particle has a *prt* relation to a verb. Light verb constructions are detected using the light verb lists provided by Tu and Roth (2012) and Huddleston et al. (2002). Finally we randomly sampled 2,000 sentences as our non-VMWE dataset from the filtered result.

3.3 QE Systems & MT Models Used

For machine translation quality assessment, we utilize two state-of-the-art quality estimation (QE)

models from the WMT24 Metrics Shared Task (Freitag et al., 2024). Unlike other standard evaluation models, both models have reference-free mode, meaning they can assess translation quality using only the source sentence and the MT hypothesis.

1) MetricX-24-QE (Juraska et al., 2024): Based on mT5 (Xue et al., 2021), it is fine-tuned on a combination of human evaluation scores and synthetic data designed to capture common translation errors. In QE mode, it generates a score in the range from 0 to 25, where lower scores indicate better translation quality.

2) xCOMET-QE (Guerreiro et al., 2024): Built on a large pre-trained encoder model XLM-RoBERTa-XL (Goyal et al., 2021), it is trained using a multi-task setup that combines sentence-level quality assessments and word-level severity tags, in order to predict both a sentence-level score and identify specific error spans with their severity levels. In QE mode, it produces a score between 0 and 1, with higher scores indicating better translation quality.

MT Models: We used 8 MT systems in this work: 1) SeamlessM4T (Communication et al., 2023), 2) Madlad400 (Kudugunta et al., 2023), 3) M2M100 (Fan et al., 2020), 4) Opus-MT (Tiedemann and Thottingal, 2020), 5) LLaMAX3 Alpaca (Lu et al., 2024), 6) Phi-4-multimodal (Microsoft et al., 2025), 7) GemmaX2 (Cui et al., 2025), 8) Google Translate API. See further details about these MT systems in Appendix A.

Language scope. For VMWE-dataset experiments, we translate from English into seven targets: German (de), Czech (cs), Russian (ru), Chinese (zh), Spanish (es), Japanese (ja), and Turkish (tr). Selected to span major language families and to be supported by our QE models and MT systems.

4 Evaluating Translation on WMT Datasets

The experiment design above uses reference-free evaluation on VMWE-specific datasets. To test whether the same pattern holds under human evaluation, and to strengthen external validity, we turn to WMT (Workshop on Machine Translation) datasets (Koehn and Monz, 2006), which provide system outputs with human DA scores. Details about human DA scores in Appendix G.

In the WMT setting, we (i) locate VMWE instances in English source sentences, (ii) collect the corresponding system translations and human DA scores, and (iii) measure the difference in human-rated quality between VMWE and non-VMWE subsets. This lets us ask the same question as before—do VMWEs hurt translation quality?—but now answer with human judgments.

4.1 Dataset

The WMT datasets are widely used for benchmarking MT models. The datasets across multiple years contain English sentences along with their corresponding translations in other languages by MT systems and humans, accompanied by human judgment (direct assessment scores) (Bojar et al., 2016) evaluating the translation quality.

To maintain a balance between diversity and consistency, we select the four most frequently used language pairs from 2017 to 2024: English to Czech (*en-cs*), English to German (*en-de*), English to Chinese (*en-zh*) and English to Russian (*en-ru*).

Different MT systems are evaluated and reported by the WMT Findings of each year—(Bojar et al., 2017), (Bojar et al., 2018), (Barrault et al., 2019), (Barrault et al., 2020), (Akhbardeh et al., 2021), (Kocmi et al., 2022), (Kocmi et al., 2023), (Kocmi et al., 2024)—from 2017 to 2024. We selected the top four MT systems for each year's specific language pair to reflect the highest level of MT performance in that year.

4.2 VMWE Extraction and Classification Using Large Language Models (LLMs)

We construct VMWE and non-VMWE subsets from the WMT data as follows: For non-VMWE sentences, we apply the same procedure as in Section 3.2 into the WMT datasets. To obtain VMWE sentences, we apply a two-step extraction pipeline:

1) Heuristic retrieval. For verbal idioms: Starting from filtered idiom lists in EPIE and MAGPIE, we retrieve sentences containing similar phrases, requiring a BLEU-4 score of at least 0.6 (Papineni et al., 2002). We use dependency parser for potential verb-particles and light verbs.

2) LLM-based disambiguation. Following the PARSEME guidelines (Savary et al., 2023), we design a chain-of-thought prompt (Wei et al., 2023) and since there are no gold labels in WMT to compute accuracy, we test the performance of the LLMs on the existing VMWE datasets. GPT-4o (OpenAI et al., 2024b) is the best performing model, so we use it for extracting VMWE sentences from WMT. The performance of GPT-4o is given in Table 1, and the results for all LLMs and full pipeline details are provided in Appendix H. Finally, we compare

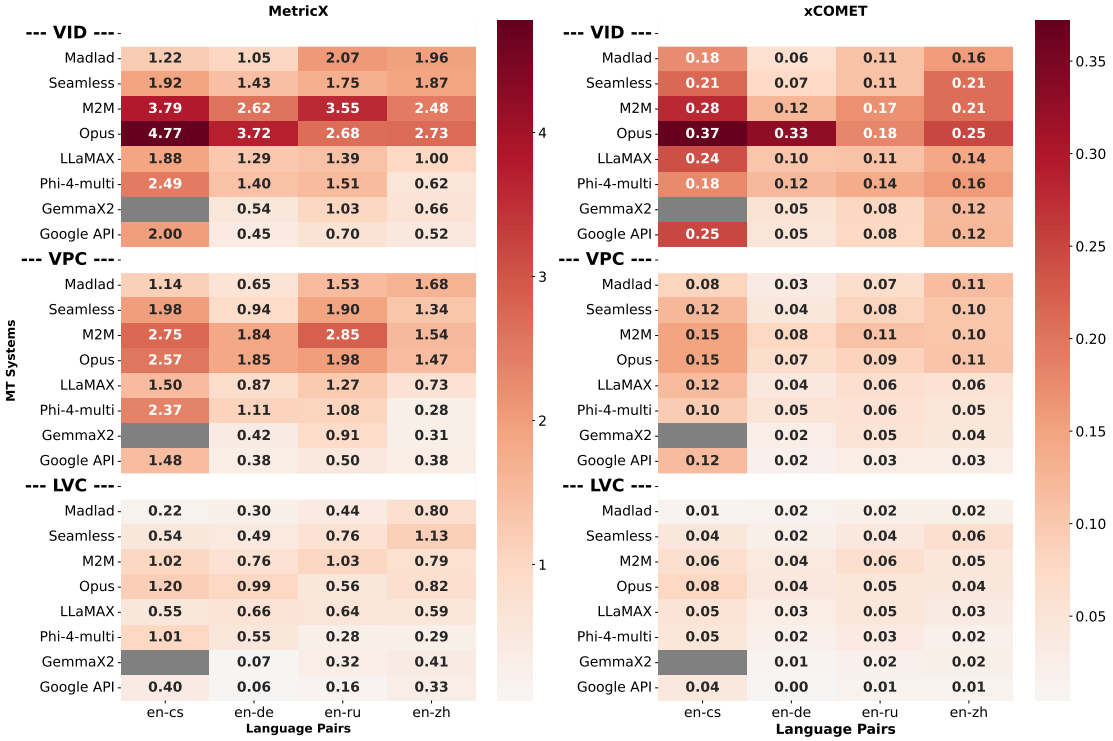

Figure 2: Comparison of the translation quality between sentences with and without VMWE, using MetricX-24-QE (*range from 0 to 25, lower the better*) and xCOMET-QE (*range from 0 to 1, higher the better*) scores. For each cell, the values are the difference between the score of VMWE and non-VMWE sentences. Larger numbers (darker colors) mean worse translation performance on VMWE sentences.

Type	VID	VPC	LVC
Pos	81.8	80.0	81.6
Neg	78.9	84.4	78.0

Table 1: F1-score of GPT-4o performance for identifying VMWE sentences on specific balanced sampled dataset. Pos: VMWE sentences. Neg: non-VMWE sentences.

translation quality between the VMWEs and non-VMWEs using their respective human DA scores.

5 Results

We report the translation quality comparing sentences with and without verbal multiword expressions on both the specific VMWE datasets and the machine translation datasets (WMT). Our experimental results suggest that VMWE consistently degrade machine translation quality.

5.1 Performance on VMWE Datasets

Figure 2 provides a comprehensive comparison of VMWE effects on MT system's performance using existing VMWE datasets, evaluated by MetricX-24 and xCOMET as discussed in Section 3.2. We evaluate the QE scores of sentences with and without the language phenomenon (VMWE vs. non-VMWE), and report the difference δ between them. For MetricX-24 (lower the better),

$$\delta_{\text{MetricX24}} = QE_{\text{VMWE}} - QE_{\text{non-VMWE}}.$$

Since xCOMET is higher the better, we reverse it to make it consistent: $\delta_{\text{xCOMET}} = QE_{\text{non-VMWE}} - QE_{\text{VMWE}}$. Positive values indicate VMWE sentences have a worse translation quality than non-VMWE sentences. We observe that VMWE consistently degrade translation quality across multiple MT systems and languages. Verbal idioms experience the largest performance gap, followed by verb-particles and light verb constructions.

A language effect is also visible: In MetricX view, penalties tend to be larger for en–cs and en–ru than for en–zh, with en–de in between—especially pronounced for VID. xCOMET has a different view: en–cs and en–zh tend to bear larger penalties than en–de. Overall, the two metrics agree in direction and relative magnitude: VID shows the largest degradation, moderately on VPC, then LVC. The performance of remaining three language pairs is consistent with this trend, details in Appendix F.

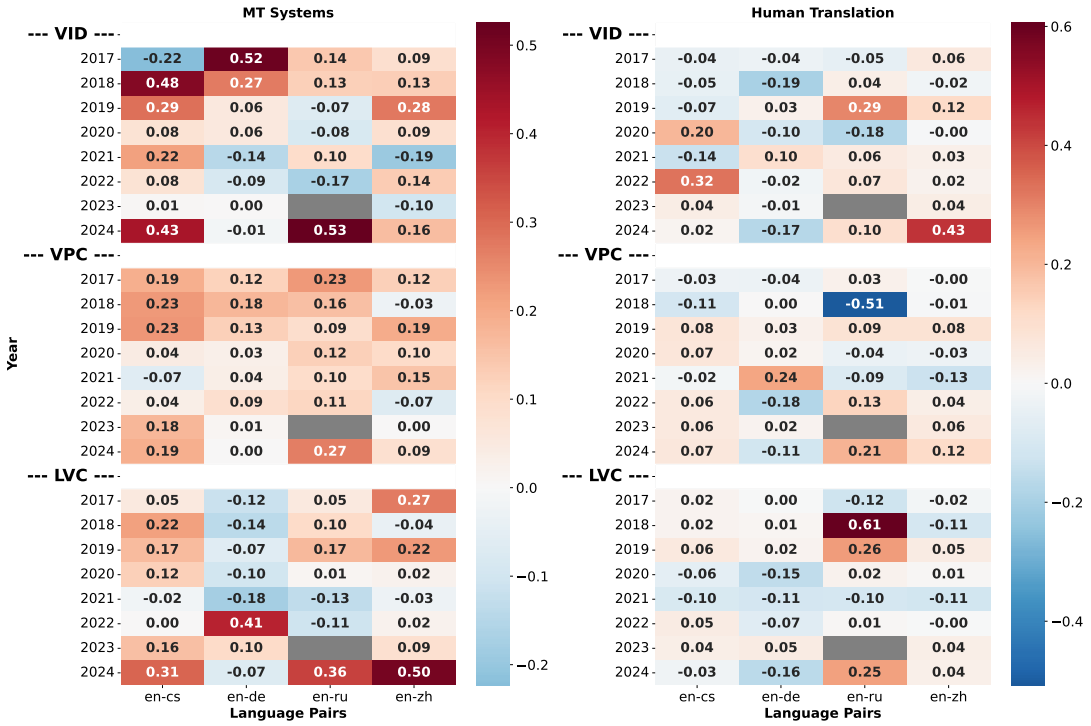

Figure 3: Comparison of the translation quality between sentences with and without VMWE, using the human DA scores (*typically range from -1 to 1, higher the better*). Left is machine translation and right is human translation (both evaluated by humans). For each cell, the values are the difference between the score of VMWE and non-VMWE sentences. Grey cell indicates missing language pair in that year. Anomalies for Human DA scores on en–ru 2018 are explained in appendix D.

Invalid Translations. Some MT models, particularly the LLM-based ones, frequently produce invalid outputs—wrong target language, empty translation, or repetitive text. We exclude these instances from our reported scores. GemmaX2 mistranslates over 75% of *en-cs* sentences, so we omit its results for that pair (grey cell). We further discuss these mistranslations in Appendix B, and provide error cases in Appendix J.

5.2 Performance on WMT Dataset

MT Systems' Performance on Sentences w/ and w/o VMWEs. After extracting the VMWE sentences from the WMT dataset, we compare their human DA scores with those of non-VMWE sentences of each year's top four MT systems from 2017 to 2024. Positive scores mean that VMWE sentences received lower score than non-VMWE. In the left heatmap of Figure 3, top MT systems are still negatively affected by the presence of VMWE. Notably, this trend aligns with our earlier findings using MetricX-24 and xCOMET, though the previous results are more salient and consistent. It suggests that both automatic evaluation and human judgment confirm the translation degradation.

Human Translators' Performance on Sentences w/ and w/o VMWEs. The WMT datasets also provide multiple human translators' translations (also evaluated by human judges). From the right heatmap in Figure 3, English to Russian is most affected overall. In contrast to machine translation, human translations appear to be less affected by VMWEs. Most cells show a small delta, indicating that humans translate VMWE sentences with comparable quality to non-VMWE ones. This suggests that human translators implicitly understand the figurative meanings of VMWEs, which themselves encode this semantic knowledge. But MT systems fail to capture this, underscoring the room for improvement.

6 Analysis

Our results show VMWE sentences demonstrate degradation in overall translation quality across languages and MT systems. However, an important question remains: are there any confounding factors that contribute to this degradation? It may be possible that the VMWE sentences might just be harder sentences in general, so the observed degradation may not be due to multiword expression. In

MT System	VID	VPC	LVC	xCOMET
Opus	78.64	65.51	62.21	66.89
M2M	82.15	67.67	62.19	73.02
Phi-4-multi	66.50	54.43	56.36	74.68
Madlad	69.88	52.79	50.14	78.22
Seamless	67.91	56.30	55.75	76.43
LLaMAX	66.82	53.11	55.89	78.08
GemmaX2	55.41	46.77	43.90	85.69
Google API	**52.98**	**40.77**	**39.25**	**87.19**

Table 2: Percentage of sentences in which xCOMET-QE error spans overlap the source VMWE span. Values are averaged over seven target languages. Lower overlap indicates errors are less concentrated on the verb multiword expression itself. Stronger systems show relatively lower overlap, while weaker systems exhibit higher overlap, especially on verbal idioms.

this section, we perform deeper analysis and show the presence of the VMWE candidate is the primary cause for the translation quality degradation.

6.1 xCOMET Error Span

xCOMET provides token-level error spans in the translated text marked with severity labels. For this experiment, we first find the overlap percentage between these error spans and the VMWE candidate in the translated text. We then show that this overlap percentage has a high correlation with the QE score assigned by xCOMET. This will help identify if the degradation in translation quality is primarily due to the presence of the VMWE candidate. Since xCOMET outputs the error span in the translated text, a key part of this analysis is to align the tokens between the English source sentence and its translation. For this, we use SIMALIGN (Jalili Sabet et al., 2020) to obtain bilingual word alignments. SIMALIGN computes contextual embeddings for source and target tokens with a multilingual Transformer, constructing a similarity matrix via cosine similarity between every source–target token pair, then extracting alignments using heuristic matching. We then use these alignments to trace errors back to the source-side VMWE candidate.

Table 2 shows the error span overlap for the eight models across seven language pairs. A high error span overlap indicates that the MT system struggles with the VMWE candidate, showcasing its weakness dealing with VMWE sentences. The two best performing models, GemmaX2 and Google API consistently have low overlap percentage, while the MT systems that suffer the most degradation like Madlad or Opus show a high score, especially for VIDs. Further analysis in Appendix K (Figure 7 shows that the Pearson correlation score between the percentage of sentences in which the error span overlaps with the VMWE candidate and the overall QE score is consistently high across languages. This confirms that the VMWE degrades translation quality. Furthermore, the slope of the best fit line indicates in Figure 7 that verbal idioms have most significant impact on translation quality, followed by VPC and then LVC. See appendix K for more details about error span overlap.

6.2 Confounding Factors

A potential concern in comparing VMWE and non-VMWE segments is that the observed quality differences may reflect general sentence difficulty rather than effects specific to VMWE phenomena. VMWE-containing sentences may be longer, more lexically ambiguous, or syntactically more complex, all of which are known to negatively impact machine translation quality. To account for this confound, we perform a segment-level regression analysis that provides controls for standard source-side difficulty factors.

Following Araghi and Palangkaraya (2024), we compute four features for each English source sentence: (i) if the sentence contains VMWE (I_{vmwe}), (ii) sentence length (S_{len}): number of tokens, (iii) degree of polysemy (P_{deg}): average number of WordNet senses per content word (Fellbaum et al., 2006), and (iv) structural complexity (T_{cmp}): sum of dependency arc lengths parsed by spaCy (Honnibal et al., 2020). All continuous features are z-normalized. We then fit linear regression models with QE scores as dependent variables and VMWE presence as the main predictor, alongside the three difficulty features. Specifically, we estimate segment-level linear models of the form:

$$\text{score}_i = \beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i,$$

where $\mathbf{x}_i = [I_{vmwe,i}, S_{len,i}, P_{deg,i}, T_{cmp,i}]$, and ε_i denotes the residual error term.

The regression dataset consists of $N = 305{,}428$ segment-level observations drawn from our QE experiments. Each observation corresponds to a single translation output produced by a particular MT system, for a given English source sentence, into a specific target language. We pool across all MT systems and all language pairs used in our experiments, and retain only segments for which both QE scores (xCOMET and MetricX) and all three difficulty features are available. This yields

Predictor	xCOMET β (SE)	MetricX β (SE)
β_0	0.7908 (0.0010)***	5.6889 (0.0183)***
I_{vmwe}	-0.0813 (0.0012)***	0.9954 (0.0240)***
S_{len}	-0.0359 (0.0012)***	0.6348 (0.0240)***
P_{deg}	-0.0126 (0.0006)***	0.0244 (0.0110)*
T_{cmp}	-0.0120 (0.0009)***	0.0059 (0.0210)

Table 3: Segment-level linear regression of QE scores on VMWE presence and source-side difficulty features. Standard errors (SE) are in parentheses. The β_0 is the expected score for non-VMWE segments at mean difficulty. For MetricX, higher score$_i$ indicates worse quality. Significance codes: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

205,565 VMWE outputs and 99,863 non-VMWE outputs from the BNC baseline.

Table 3 reports the estimated coefficients β with standard errors in parentheses; for the z-scored difficulty predictors, β corresponds to a one-standard-deviation increase in the feature. Model fit statistics indicate modest but expected explanatory power at the sentence level, with $R^2 = 0.063$ (adjusted $R^2 = 0.063$) for xCOMET and $R^2 = 0.027$ (adjusted $R^2 = 0.027$) for MetricX. As expected, longer, more polysemous and structurally complex sentences are harder for MT, exhibiting lower xCOMET scores and higher MetricX scores. Importantly, VMWE presence remains a strong predictor of translation degradation even after controlling for these difficulty factors. On average, VMWE sentences receive xCOMET scores that are approximately 0.08 points lower on the 0–1 scale, and MetricX scores that are about 1 point higher (worse) on the 0–25 scale. Taken together, VMWE presence remains a significant predictor of QE scores conditional on other factors.

6.3 Error Analysis

Table 4 provides some examples of poor translations. Sentence (a) contains the VMWE expression *get out*. In this context, *get out* does not mean just to leave physically–it is used idiomatically for vacating a position or situation. Google API mistranslated this as another "retire", losing the intended meaning. The Chinese translation roughly says "*Once you're retired, you have three months to retire*." making the translated text awkward and incoherent.

Sentence (b) contains the VMWE expression: *breaking up*—meaning to end a romantic relationship. It does not mean actually dead. "Romper" in Spanish means "to break". The Spanish sentence when back translated says "*To break, as Neil Sedaka once said, is difficult to do.*" losing its intended meaning.

Type	Text
Sentence (a)	Once you're retired you have three months to get out.
Translation	一旦退休，您就有三个月的时间退休。
Sentence (b)	Breaking up, as Neil Sedaka once put it, is hard to do.
Translation	Romper, como dijo una vez Neil Sedaka, es difícil de hacer.
Sentence (c)	"He's dead to the world," said Mabel, over her shoulder.
Translation	他已经死了，"梅布尔回头说道。"

Table 4: Illustrative examples of incorrect translations.

In Sentence (c), the VMWE expression: *dead to the world* means to be deeply asleep, completely unaware of what's happening around. It does not mean actually dead. The translated sentence has a severe misinterpretation of the idiom—"他已经死了" meaning "*He is already dead*." This is a literal translation, completely losing the intended idiomatic meaning. The contextual mismatch means anyone reading this in Chinese would think a character died, changing the entire tone of the scene. Meanwhile, in the English version, it's a light, perhaps even humorous or casual remark about someone being fast asleep.

7 Conclusion

Verbal multiword expressions systematically degrade machine translation quality. Across languages and systems, we observe a consistent gradient: verbal idioms are most disruptive, verb–particle constructions show moderate degradation, and light-verb constructions are least affected. This trend holds under both MetricX-24 and xCOMET. The degradation is primarily attributed to the presence of the VMWE and not the overall semantics. Together, these findings suggest that improving robustness on everyday text requires explicit handling of non-compositionality—through targeted VMWE detection, VMWE-aware training, and selective pre-editing or normalization—to produce translations that are more faithful and reliable in the presence of multiword expressions.

Limitations

While we conduct a large-scale evaluation of VMWEs, our study comes with some limitations. The WMT datasets have human reference scores for the translations, but the translated sentences in VMWE datasets are solely judged by QE models and do not involve human reference scores. While the QE translation scores align with the human reference scores according to our experiments, it still remains a limitation in our work.

Our work involves a large-scale analysis of the negative impact of verbal multiword expressions (VMWEs) on machine translation quality. While our findings highlight important details about how different VMWE categories and languages are affected, we do not propose any new method to improve translation quality. We leave this for future work.

Finally, our scope is limited from English to Czech, German, Spanish, Japanese, Russian, Turkish, and Chinese. We do not explore low-resource languages and other translation directions.

References

Lasha Abzianidze, Johannes Bjerva, Kilian Evang, Hessel Haagsma, Rik van Noord, et al. 2017. The Parallel Meaning Bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 242–247, Valencia, Spain. Association for Computational Linguistics.

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, et al. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Sahar Araghi and Alfons Palangkaraya. 2024. The link between translation difficulty and the quality of machine translation: a literature review and empirical investigation. *Lang. Resour. Eval.*, 58(4):1093–1114.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. *Handbook of natural language processing*, 2:267–292.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, et al. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, et al. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Christos Baziotis, Prashant Mathur, and Eva Hasler. 2023. Automatic evaluation and analysis of idioms in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3682–3700, Dubrovnik, Croatia. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, et al. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, and Others. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, et al. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Ram Boukobza and Ari Rappoport. 2009. Multi-word expression identification using sentence surface features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 468–477, Singapore. Association for Computational Linguistics.

Miriam Butt. 2010. The light verb jungle: Still hacking away. *Complex Predicates: Cross-Linguistic Perspectives on Event Structure*.

Fabienne Cap, Manju Nirmal, Marion Weller, and Sabine Schulte im Walde. 2015. How to account for idiomatic German support verb constructions in statistical machine translation. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 19–28, Denver, Colorado. Association for Computational Linguistics.

Marine Carpuat and Mona Diab. 2010. Task-based evaluation of multiword expressions: a pilot study

in statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 242–245, Los Angeles, California. Association for Computational Linguistics.

Antonio Castaldo and Johanna Monti. 2024. Prompting large language models for idiomatic translation. In *Proceedings of the 1st Workshop on Creative-text Translation and Technology*, pages 32–39, Sheffield, United Kingdom. European Association for Machine Translation.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, et al. 2023. Seamless: Multilingual expressive and streaming speech translation.

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

Silvio Ricardo Cordeiro and Marie Candito. 2019. Syntax-based identification of light-verb constructions. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 97–104, Turku, Finland. Linköping University Electronic Press.

Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan, and Bin Wang. 2025. Multilingual machine translation with open large language models at practical scale: An empirical study.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

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

Sundesh Donthi, Maximilian Spencer, Om Patel, Joon Doh, Eid Rodan, Kevin Zhu, and Sean O'Brien. 2025. Improving llm abilities in idiomatic translation.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2020. Beyond english-centric multilingual machine translation.

Christiane Fellbaum, Alexander Geyken, Axel Herold, Fabian Koerner, and Gerald Neumann. 2006. Corpus-based studies of german idioms and light verbs. *International Journal of Lexicography*, 19(4):349–360.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. 2021. Larger-scale transformers for multilingual masked language modeling. *CoRR*, abs/2105.00572.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, et al. 2024. The llama 3 herd of models.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 279–287, Marseille, France. European Language Resources Association.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python.

Rodney Huddleston, Rodney Huddleston, and Geoffrey K. Pullum. 2002. *The clause: complements*, chapter 4. Cambridge University Press.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in c++.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Paria Khoshtab, Danial Namazifard, Mostafa Masoudi, Ali Akhgary, Samin Mahdizadeh Sani, and Yadollah Yaghoobzadeh. 2025. Comparative study of multilingual idioms and similes in large language models.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, et al. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, et al. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, et al. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Philipp Koehn and Christof Monz, editors. 2006. *Proceedings on the Workshop on Statistical Machine Translation*. Association for Computational Linguistics, New York City.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset.

Geoffrey Leech. 1992. 100 million words of english:the british national corpus (bnc). *Second Language Research*, 28:1–13.

Linfeng Liu, Saptarshi Ghosh, and Tianyu Jiang. 2025. Evaluating the impact of verbal multiword expressions on machine translation.

Yinquan Lu, Wenhao Zhu, Lei Li, Yu Qiao, and Fei Yuan. 2024. LLaMAX: Scaling linguistic horizons of LLM by enhancing translation capabilities beyond 100 languages. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10748–10772, Miami, Florida, USA. Association for Computational Linguistics.

Shushen Manakhimova, Eleftherios Avramidis, Vivien Macketanz, Ekaterina Lapshinova-Koltunski, Sergei Bagdasarov, and Sebastian Möller. 2023. Linguistically motivated evaluation of the 2023 state-of-the-art machine translation: Can ChatGPT outperform NMT? In *Proceedings of the Eighth Conference on Machine Translation*, pages 224–245, Singapore. Association for Computational Linguistics.

Microsoft, : Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras.

Adaeze Ohuoba, Serge Sharoff, and Callum Walker. 2024. Quantifying the contribution of MWEs and polysemy in translation errors for English–Igbo MT. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 537–547, Sheffield, UK. European Association for Machine Translation (EAMT).

OpenAI. 2025. GPT-5.1 System Card.

OpenAI, Josh Achiam, Steven Adler, et al. 2024a. Gpt-4 technical report.

OpenAI, : Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024b. Gpt-4o system card.

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

Caroline Pasquer, Agata Savary, Carlos Ramisch, and Jean-Yves Antoine. 2020. Verbal multiword expression identification: Do we need a sledgehammer to crack a nut? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3333–3345, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Anita Rácz, István Nagy T., and Veronika Vincze. 2014. 4FX: Light verb constructions in a multilingual parallel corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 710–715, Reykjavik, Iceland. European Language Resources Association (ELRA).

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

Carlos Ramisch. 2014. Multiword expressions acquisition: A generic and open framework. In *Multiword Expressions Acquisition*.

Carlos Ramish, Laurent Besacier, and Alexander Kobzar. 2013. How hard is it to automatically translate phrasal verbs from English to French? In *Proceedings of the Workshop on Multi-word Units in Machine Translation and Translation Technologies*, Nice, France.

Omid Rohanian, Marek Rei, Shiva Taslimipoor, and Le An Ha. 2020. Verbal multiword expressions for identification of metaphor. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2890–2895, Online. Association for Computational Linguistics.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing*, pages 1–15, Berlin, Heidelberg. Springer Berlin Heidelberg.

Agata Savary, Cherifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, et al. 2023. PARSEME corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35, Dubrovnik, Croatia. Association for Computational Linguistics.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, et al. 2017. The PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47, Valencia, Spain. Association for Computational Linguistics.

Prateek Saxena and Soma Paul. 2020. Epie dataset: A corpus for possible idiomatic expressions.

Huacheng Song and Hongzhi Xu. 2024. A deep analysis of the impact of multiword expressions and named entities on Chinese-English machine translations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6154–6165, Miami, Florida, USA. Association for Computational Linguistics.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, et al. 2024. Gemma 2: Improving open language models at a practical size.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Yuancheng Tu and Dan Roth. 2011. Learning English light verb constructions: Contextual or statistical. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 31–39, Portland, Oregon, USA. Association for Computational Linguistics.

Yuancheng Tu and Dan Roth. 2012. Sorting out the most confusing english phrasal verbs. In *International Workshop on Semantic Evaluation*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Haifeng Wang, Hua Wu, Zhongjun He, Liang Huang, and Kenneth Ward Church. 2022. Progress in machine translation. *Engineering*, 18:143–153.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Marion Weller-Di Marco and Alexander Fraser. 2022. Test suite evaluation: Morphological challenges and pronoun translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 458–468, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer.

Andrea Zaninello and Alexandra Birch. 2020. Multiword expression aware neural machine translation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3816–3825, Marseille, France. European Language Resources Association.

A Machine Translation Systems Used

We used the following widely used machine translation systems in this work:
1) Madlad400 (Kudugunta et al., 2023): A Google multilingual machine translation model, based on the T5 architecture (Raffel et al., 2023) that was trained on 250 billion tokens covering over 450 languages. The model we used here is madlad400-10b-mt.
2) SeamlessM4T (Communication et al., 2023): Meta AI's massively multilingual and multimodal machine translation model, supporting an impressive range of translation capabilities with 96 languages for text input/output. The model we used here is seamless-m4t-v2-large, model size 2.3B.
3) M2M100 (Fan et al., 2020): A Meta-powered multilingual encoder-decoder model, primarily designed for translation tasks, supporting direct translation between 100 languages without requiring English as an intermediate language. The model we used here is m2m100-1.2B.
4) Opus-MT (Tiedemann and Thottingal, 2020): Provides open translation services built on the Marian neural machine translation framework (Junczys-Dowmunt et al., 2018), trained on Opus data, and later converted to PyTorch models for the Hugging Face ecosystem. The models we used here are from Helsinki-NLP—opus-mt-en-xx, "xx" is the target language code, mdoel size are all less than 1B.
5) LLaMAX3 Alpaca (Lu et al., 2024): An LLM-based machine translation model, LLaMAX combines powerful multilingual translation capabilities with instruction-following abilities. This model extends Meta's LLaMA 3 architecture (Grattafiori et al., 2024) to support translation between over 100 languages without sacrificing its ability to follow complex instructions. The model we used here is LLaMAX3-8B-Alpaca.
6) Phi-4-multimodal (Microsoft et al., 2025): Built upon the pretrained Phi-4-mini model, Phi-4-multi can processes text, image, and audio inputs to generate text outputs. While primarily known for its multimodal capabilities, it can handle translation tasks as part of its broader language understanding abilities, making it another LLM-based MT system for our task. The model we used here is Phi-4-multimodal-instruct, model size 5.6B.
7) GemmaX2 (Cui et al., 2025): A very recent multilingual LLM-based translation model, that achieved state-of-the-art performance across 28 languages. Based on Google's Gemma2 architecture (Team et al., 2024), it consistently outperforms other LLM-based MT models like TowerInstruct and XALMA, achieving competitive results with Google Translate and GPT-4-turbo. The model we used here is GemmaX2-28-9B-v0.1.
8) Google Translate API[1]: A widely used multilingual neural machine translation service developed by Google, offering translations for 249 languages and language varieties as of March 2025.

We run all MT models on a server with eight NVIDIA RTX 6000 Ada Generation GPUs.

B Invalid Translations by LLM-Based MT Models

During evaluation, we observed several instances of invalid outputs from LLM-based MT models. The most common cases were translations in the wrong language (e.g., translating to Russian text when asked for Chinese), and untranslated sentences where the model outputs the English source sentence. We also observed some cases of malformed or repetitive output, such as repeated characters or phrases in the translated text. To identify invalid translations, we use the lingua-language-detector-2.1.0[2] to automatically detect the language of the translated output and match it with the intended target language. If there is a mismatch, the translation is marked as invalid. We filter out these invalid translations when reporting the scores for VMWE vs non-VMWE sentences.

Table 5 shows all MT systems and corresponding languages pairs where the percentage of mistranslated sentence in all existing VMWE datasets is greater than 10. LlaMAX fails to convert more than 99% of the sentences to the Turkish and Japanese. While GemmaX2 achieves SOTA performance in terms of QE score, it suffers consistently from mistranslating—all 7 languages have significant error rates, with Japanese, Czech and Turkish have over 70% errors. Interestingly, Opus is the only MT system with a high error rate that is not LLM-based (we use model Helsinki-NLP/opus-mt-en-trk to translate English to Turkish).

[1] https://cloud.google.com/translate?hl=en
[2] https://pypi.org/project/lingua-language-detector/

MT System	Language Pair	Percentage of Errors
LlaMAX	en-tr	99.89
LlaMAX	en-ja	99.80
LlaMAX	en-de	16.03
LlaMAX	en-cs	14.01
GemmaX2	en-ja	78.68
GemmaX2	en-cs	75.69
GemmaX2	en-tr	74.91
GemmaX2	en-es	53.16
GemmaX2	en-de	14.57
GemmaX2	en-zh	14.06
GemmaX2	en-ru	11.44
Opus	en-tr	62.74

Table 5: Percentage of mistranslated sentences.

C Comparison with top-tier LLMs.

We additionally evaluate a stronger LLM baseline. While prior work discussed in our setup indicates that recent LLM-based translators (e.g., GemmaX2) and commercial MT APIs can be competitive with GPT-4-class systems on general MT benchmarks (Cui et al., 2025), directly verifying VMWE robustness on top-tier LLMs strengthens the main claim of this paper.

We run GPT-4.1 (OpenAI et al., 2024a) and GPT-5.1 (OpenAI, 2025) as direct translators on the same controlled subset: 100 sentences per category (NON-VMWE, VID, VPC, LVC), translating from English into Czech, German, Russian, and Chinese. We score all outputs with the same MetricX-24 pipeline and compute the degradation gap as $\Delta = Metricx24_{\text{VMWE}} - Metricx24_{\text{non-VMWE}}$, where lower MetricX-24 is better and thus larger Δ indicates a larger VMWE-specific degradation. See prompt for translation in appendix M.

Table 7 compares GPT-4.1, GPT-5.1, and Google Translate API on this same 100-sentence-per-category sample. Two patterns stand out. First, VID remains consistently harder than NON-VMWE for all three systems across all four languages, indicating that the core VMWE difficulty persists even for top-tier LLMs used directly as translators. Second, GPT-5.1 reduces the average degradation gap relative to GPT-4.1 across all three VMWE types (most clearly for VPC and LVC), with several small or negative Δ values for VPC/LVC. Since these results are computed on a fixed subset of only 100 sentences per category, occasional negative gaps may reflect sampling variance (i.e., the VMWE subset being slightly easier than the non-VMWE subset under MetricX-24), but the VID gaps remain positive and stable. Overall, the data suggest that newer LLMs narrow the VMWE degradation gap, but do not eliminate it, especially for idiomatic VID cases.

We also report xCOMET-based degradation gaps on the same controlled subset of 100 sentences per category. Since higher xCOMET indicates better quality, we define $\Delta_{\text{xCOMET}} = xCOMET_{\text{non-VMWE}} - xCOMET_{\text{VMWE}}$, where larger Δ_{xCOMET} indicates a larger drop on VMWE sentences. Table 8 shows that GPT-4.1, GPT-5.1, and Google Translate API all exhibit clear degradation for VID, with Google showing the largest gaps and the largest drops on en–cs. For VPC and LVC, gaps are close to zero for GPT-4.1/GPT-5.1 but consistently positive for Google, suggesting that the VMWE effect is more pronounced for Google under xCOMET even when MetricX-24 gaps are smaller for some language pairs. Together with Table 7, these results indicate that the VMWE effect is robust across metrics, while the magnitude and the relative gains of GPT-5.1 over GPT-4.1 depend on the VMWE type and the evaluation metric.

D Anomalies in WMT Human Translation Score

On closely examining Figure 3, we identify English–Russian (en–ru) in 2018 as an outlier in the human translation scores.

We investigated these specific anomalies and found that they are largely due to data sparsity and uneven sampling in the WMT human evaluation logs for that year. In the 2018 en–ru dataset, only a very small number of reference translations received human DA scores: 243 evaluated reference segments in total, of which just 11 contain LVCs and 10 contain VPCs. By contrast, the 2019 en–ru dataset has 1,499 evaluated reference segments, including 37 LVC sentences and 105 VPC sentences.

E Qualitative Discussion

A consistent finding across our experiments is that VID is the most affected VMWE category in machine translation, followed by VPC, while LVC is affected the least. This ordering is stable across our main experimental settings. In a supplementary check on a small controlled subset (100 sentences per category), GPT-4.1 and GPT-5.1 exhibit the same qualitative pattern under both MetricX-24 and xCOMET, which is consistent with the broader

Models	LVC		VPC		VID	
	Pos	Neg	Pos	Neg	Pos	Neg
GPT-4o	**81.6**	**78.0**	80.0	84.4	81.8	78.9
Phi-4	71.3	70.6	**80.5**	**84.9**	81.7	**80.1**
Llama-3.3-70B	56.7	44.7	**80.5**	**84.9**	**83.1**	79.4
DeepSeek-R1-70B	55.8	57.1	77.6	83.3	80.5	79.3

Table 6: F1-score comparison of different LLMs. Pos: VMWE sentences. Neg: non-VMWE sentences.

Type	Lang	GPT-4.1	GPT-5.1	Google API
VID	en–cs	+0.95	+0.48	+2.00
	en–de	+0.68	+0.34	+0.45
	en–ru	+0.70	+0.42	+0.70
	en–zh	+0.62	+0.45	+0.52
	Avg	+0.74	+0.42	+0.92
VPC	en–cs	+0.61	−0.05	+1.48
	en–de	+0.43	+0.20	+0.38
	en–ru	+0.50	−0.01	+0.50
	en–zh	+0.63	+0.23	+0.38
	Avg	+0.54	+0.09	+0.68
LVC	en–cs	+0.10	−0.86	+0.40
	en–de	+0.06	−0.14	+0.06
	en–ru	+0.38	+0.15	+0.16
	en–zh	+0.68	+0.24	+0.33
	Avg	+0.31	−0.15	+0.24

Table 7: ΔMetricX-24 degradation gaps (higher = worse) for three VMWE categories on controlled sample. We report $\Delta = Metricx24_{\text{VMWE}} - Metricx24_{\text{non-VMWE}}$, so positive values indicate worse performance on VMWE sentences relative to non-VMWE sentences under the same evaluation pipeline.

Type	Lang	GPT-4.1	GPT-5.1	Google API
VID	en–cs	+0.18	+0.22	+0.25
	en–de	+0.01	+0.02	+0.05
	en–ru	+0.05	+0.06	+0.08
	en–zh	+0.16	+0.14	+0.12
	Avg	+0.10	+0.11	+0.13
VPC	en–cs	+0.01	+0.00	+0.12
	en–de	−0.01	+0.00	+0.02
	en–ru	+0.01	+0.00	+0.03
	en–zh	+0.04	+0.02	+0.03
	Avg	+0.01	+0.00	+0.05
LVC	en–cs	+0.00	−0.01	+0.04
	en–de	−0.02	+0.00	+0.00
	en–ru	+0.01	+0.01	+0.01
	en–zh	+0.04	+0.02	+0.01
	Avg	+0.01	+0.00	+0.02

Table 8: ΔxCOMET degradation gaps (higher = worse) on the same controlled 100-sentence-per-category sample. We report $\Delta_{\text{xCOMET}} = xCOMET_{\text{non-VMWE}} - xCOMET_{\text{VMWE}}$, so positive values indicate worse performance on VMWE sentences relative to non-VMWE sentences.

trends observed for the primary baselines, but we do not treat this small-scale LLM comparison as a standalone conclusion.

We interpret the overall pattern through the lens of compositionality. LVCs are often more compositional: much of the semantic content is carried by the noun while the verb is semantically light (e.g., "*take a walk*"). As a result, translation is less sensitive to the particular light verb choice and errors tend to resemble minor lexical substitutions rather than meaning shifts. VPCs are typically semi-compositional, where the particle may be literal in some contexts but idiomatic in others, leading to intermediate difficulty. VIDs are strongly non-compositional, and their meanings cannot be recovered from their parts (e.g., "*beat the clock*" meaning "finish before the deadline"). Correct translation therefore requires selecting a target-language expression that conveys the intended meaning rather than a literal rendering, which remains challenging across systems and languages. Overall, our results support the conclusion that **VMWE-related degradation in machine translation correlates with the degree of non-compositionality**.

F Heatmap for Additional Languages

Reading Figure 4, sentences containing VMWEs are consistently penalized across systems and language pairs, with a clear gradient on both Metricx-24 and xCOMET.

Language effects. Across both metrics, en–es tends to show the largest gaps, en–tr is moderate, and en–ja often attenuates or flips some systems, especially for VID and VPC. Overall, the two metrics agree in direction and relative magnitude, reinforcing the non-compositionality gradient.

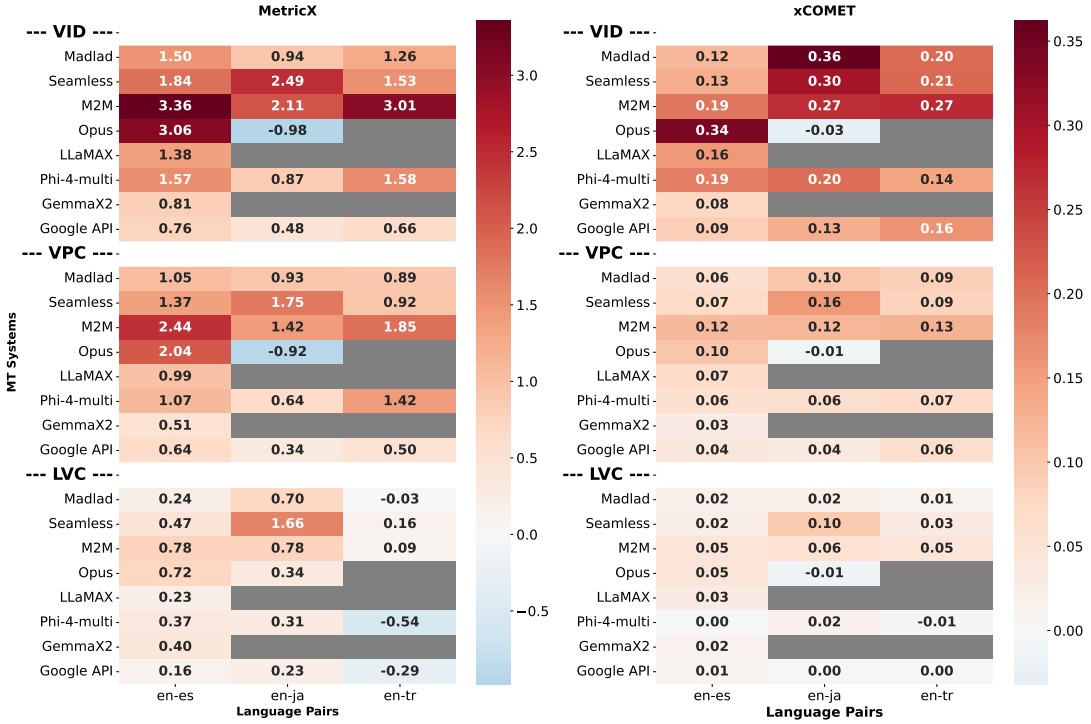

Figure 4: Comparison of the translation quality between sentences with and without VMWE, using MetricX-24-QE (*range from 0 to 25, lower the better*) and xCOMET-QE (*range from 0 to 1, higher the better*) scores. For MetricX-24 (left), the cell values are calculated as the difference of averaged score (VMWE minus non-VMWE). For xCOMET (right), the scores are opposite as non-VMWE minus VMWE. Larger numbers (darker colors) mean worse translation performance on VMWE sentences.

G Human Direct Assessment (DA)

Direct Assessment (DA) is WMT's segment-level human evaluation protocol. Raters view a source sentence, a single system translation and human-provided reference translation. Then assign a scalar quality score on a continuous scale in 0–100, reflecting overall adequacy/fluency.

How WMT collects DA. Each segment receives multiple independent ratings. Items are shown one-at-a-time; raters use a slider to score quality. WMT includes control items and standard filters to remove unreliable ratings.

How DA scores are handled. Raw scores are standardized *per rater* to remove scale bias:

$$z_{i,j} = \frac{s_{i,j} - \mu_j}{\sigma_j},$$

where $s_{i,j}$ is rater j's score for segment i, and μ_j, σ_j are that rater's mean and standard deviation. We obtain subset or system scores by averaging $z_{i,j}$ over relevant segments (and raters), optionally with stratified bootstrap for uncertainty.

How we use DA in this paper. We automatically identify English source sentences containing a VMWE (VID/VPC/LVC) and construct a non-VMWE comparison subset. For each language pair, year, and system, we collect corresponding DA separately for VMWE and non-VMWE sentences and report their difference to assess VMWE impact under human judgment.

H Comparing LLM performance in Extracting VMWE Sentences

The WMT (Workshop on Machine Translation) dataset contains English sentences along with their corresponding translations in other languages by MT systems and humans, accompanied by human judgment (direct assessment scores) (Graham et al., 2013) evaluating the translation quality. Our goal in section 4 is use this dataset to evaluate if human translation scores align with the MT evaluation models we used in section 3. The VMWE candidate sentences are collected by these rules: For verbal idioms, we use BLEU-4 (Papineni et al., 2002), threshold of 0.6 to find the verbal idiom candidates based on the verbal idiom dictionary from EPIE and MAGPIE. For verb-particle con-

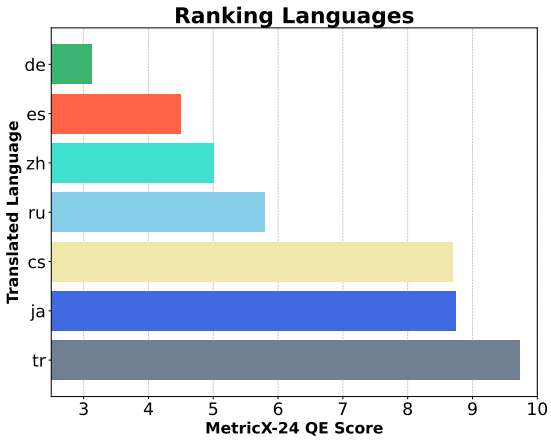

Figure 5: Ranking languages based on translation quality for sentences with VMWE using MetricX-24. Lower score indicates better performance.

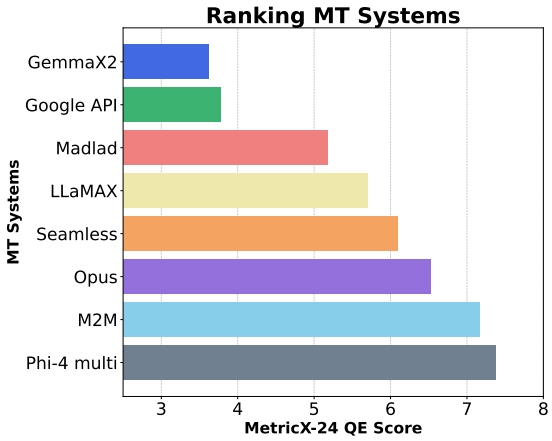

Figure 6: Ranking MT systems based on translation quality for sentences with VMWE using MetricX-24. Lower score indicates better performance.

structions, we use spaCy dependency parser (Honnibal et al., 2020) where a particle has a *prt* relation to a verb to extract the verb-particle construction candidates. And light verb constructions are detected using the light verb lists (*have, take, make, get, put, give, pay, do, offer, raise*), the related noun are extracted by spaCy dependency parser. After that, we use GPT-4o to extract VMWE sentences.

For the extraction, we tested different LLMs on the existing VMWE datasets. Table 6 shows the performance of different LLMs in classifying VMWE on existing datasets. We test 4 different models, GPT-4o (OpenAI et al., 2024b), Phi-4 (Microsoft et al., 2025), Llama 3.3-70B (Grattafiori et al., 2024) and DeepSeek-R1-Distill-70B (DeepSeek-AI, 2025). We use the same prompts and few-shot examples for all models, given in Appendix O. All the LLMs have nearly identical performance in VPC and VID categories, with Phi-4 and Llama having exact same F1 scores in VPC. However, GPT-4o establishes dominance in LVC category, with all other LLMs performing significantly worse. Due to its balanced performance across all categories, we chose GPT-4o for the main experiment.

I Comparing Translation Quality across Languages & MT Systems

In this section, we rank translation quality across languages and MT systems using their MetricX-24 QE scores. xCOMET rankings appear in Appendix L. For languages, we rank by QE scores averaged over all three VMWE categories and eight MT systems; Figure 5 presents the results. Lower scores indicate better translation quality. The rankings reveal consistent cross-linguistic trends: high-resource languages such as German (de) and Spanish (es) perform best, whereas morphologically rich or low-resource languages such as Turkish (tr), Japanese (ja), and Czech (cs) perform worst. For MT systems, we rank by QE scores averaged over the three VMWE categories and seven languages; Figure 6 presents the results. GemmaX2 achieves the best translation quality, closely followed by Google API. Because our results exclude outputs in the wrong target language (Appendix B) and GemmaX2 produces many such errors (Table 5), its score is not directly comparable to systems with fewer failures. Phi-4-multi and M2M exhibit the largest degradation in presence of VMWEs. Interestingly, the best- and worst-performing systems are both LLM-based, while traditional neural MT systems occupy the middle ground, suggesting that scale alone does not guarantee VMWE robustness.

J Analyzing Translations With Low QE Scores (Error Analysis)

Example Error 1 - Sentence Fluency and Coherency

Model: Google-API

Source Sentence: “Once you’re retired you have three months to get out.”

Chinese Translation: “一旦退休，您就有三个月的时间退休。”

Explanation: The english sentence contains the verbal multiword expression “get out”, which is often idiomatic. In this context, “get out” does not mean just to leave physically–it is often idiomatic

for vacating a position, property, or situation. For the Chinese translated text, the MT system did not interpret "get out" idiomatically. In this context, it should mean something like "vacate," "leave the premises," or "exit the position". But it got mistranslated as another "retire", which loses the intended meaning. The original says "Once you're retired…", but then the Chinese translation says "三个月的时间退休"—i.e., "three months to retire"—which repeats the idea of retiring, making the translated text something like "Once you're retired, you have three months to retire."

Example Error 2 - Literal Translation
Model: LlaMAX
Source Sentence: "Young women were beginning to get on his nerves "
Czech Translation: "Young women začínaly být jeho nervy ."
Explanation: This sentence contains the verbal multiword expression: "get on his nerves"—expressing growing annoyance. The MT system literally translates the phrase to Czech—"začínaly být jeho nervy" is grammatically broken and semantically incorrect, failing to convey the emotional content.

Example Error 3 - Literal Translation
Model: GemmaX2
Source Sentence: "He 's dead to the world , ' said Mabel , over her shoulder ."
Chinese Translation: "他已经死了，"梅布尔回头说道。"
Explanation: This sentence contains the verbal multiword expression: "dead to the world"—meaning to be deeply asleep, completely unaware of what's happening around. It does not mean actually dead. The translated sentence has a severe misinterpretation of the idiom—"他已经死了" means "He is already dead.". This is a literal translation, completely losing the intended idiomatic meaning. The contextual mismatch means anyone reading this in Chinese would think a character died, changing the entire tone of the scene. Meanwhile, in the English version, it's a light, perhaps even humorous or casual remark about someone being fast asleep.

Example Error 4 - Literal Translation
Model: GemmaX2
Source Sentence: "Breaking up, as Neil Sedaka once put it, is hard to do."
Spanish Translation: "Romper, como dijo una vez Neil Sedaka, es difícil de hacer."
Explanation: This sentence contains the verbal multiword expression: "breaking up"—meaning to end a romantic relationship. It does not mean actually dead. "Romper" in Spanish means "to break", the Spanish sentence when back translated says "To break, as Neil Sedaka once said, is difficult to do." losing its intended meaning.

Example Error 5 - Invalid Output (Mistargeted Translation)
Model: GemmaX2
Source Sentence: "At about the same time the aliens department of the Home Office took on extra staff and moved to Cleveland House on Thorney Street."
Turkish Translation: "À peu près à la même époque, le département des étrangers du ministère de l'Intérieur a embauché du personnel supplémentaire et a déménagé à Cleveland House sur Thorney Street."
Explanation: Although the source sentence was intended to be translated into Turkish, the translation was mistakenly produced in French, demonstrating correct meaning transfer but in the wrong target language.

Example Error 6 - Invalid Output (Incomplete Translation)
Model: Seamless
Source Sentence: "Now the formula one grand prix circus is getting ready to hit the road for another season … first race is in South Africa in five weeks time … what 's new … is it going to be a head to head between Mansell and Senna."
Japanese Translation: "フォミュラ1グランプリサカスシズン1のスタトをめました"
Explanation: In this instance, the verbal multiword expression "hit the road" means to get started. In this situation, the MT system does get the multiword expression correct, as the back-translation reads "The Formula 1 Grand Prix circus has decided to start Season 1.", however this is an example of incomplete translation as the MT sytem omitted the rest of the information in the source sentence.

Example Error 7 - Invalid Output (Incomplete Translation)
Model: Google API
Source Sentence: "So in effect, if you write this

long, long, rambling press release, you won't get over the point you were trying to do, you'll get publicity okay, but you won't get the kind of thing you wanted.."
Japanese Translation: "ゝゝゝ"
Explanation: In this instance, the Japanese text reads "long long", only translating a a very small part of the sentence.

Example Error 8 - Invalid Output (No Translation)
Model: LlaMAX
Source Sentence: "' Warm in winter and cool in summer, ' Fen said."
German Translation: "' Warm in winter and cool in summer, ' Fen said."
Explanation: In this instance, the LLM outputs the English source sentence as its output without translating it.

Example Error 9 - Invalid Output (Partial Translation)
Model: LlaMAX
Source Sentence: "Prior to becoming a Volunteer, I was a cutter/draper/pattern maker in a professional theatre."
Czech Translation: "Prior to becoming a Volunteer, I was a řezník/sušíř/šikovník v profesionálním divadle."
Explanation: In this instance, the LLM partially translates the sentence.

Example Error 10 - Invalid Output (Character Repetition)
Model: Google API
Source Sentence: "I 'm , I 'm , quite amazed that, Poland was occupied by Germany and Czechoslovakia and France , that 's okay er and in fact you know we should just turn a blind eye to it and just let it carry on forever ."
Japanese Translation: "私は、この...、この、この、この、この、この、この、この、、この、この、この、この、、この、この、この、、この、この、、この、この、、この、この、、この、この、、この、この、、この、この、、この、....."
Explanation: In this instance, the MT system repeats a character multiple times.

K Error Span Overlap Details

We further analyze the xCOMET error-span overlap. Figure 7 reports the Pearson correlation between the share of sentences where the error span overlaps the VMWE candidate and the overall QE score for Chinese and German. Strong correlations are found for all three VMWE categories, indicating that xCOMET spans covering the candidate are closely linked to lower translation quality rather than reflecting broader semantic or syntactic difficulty. Moreover, the fitted slope suggests that verbal idioms have the largest effect, followed by VPCs and then LVCs, consistent with the pattern reported by Liu et al. (2025).

L Comparing Translation Quality across Languages & MT Systems - xCOMET Result

In this section, we rank translation quality across languages and MT systems. For languages, we rank them using their xCOMET QE scores averaged over all three VMWE categories and eight MT systems. Figure 8 shows the result. Higher scores indicate better translation quality. The rankings reveal consistent cross-linguistic trends: high-resource languages such as German (de) and Spanish (es) achieve the best performance, while morphologically rich or low-resource languages such as Turkish (tu), Japanese (ja), and Czech perform the worst. For MT systems, we rank them using the MetricX-24 QE scores averaged over three VMWE categories and seven languages. Figure 10 shows the result. Google API is the best performing model, closely followed by GemmaX2. In contrast, Phi-4-multi and M2M exhibit the largest degradation in presence of VMWEs.

M Prompt for GPT-based Translation

Table 9 reports the exact instruction prompt used for GPT-4.1 and GPT-5.1.

N LLM Paraphrasing

We evaluate previous proposed strategies for improving MT performance on VMWEs, particularly source side paraphrasing prior to translation (Donthi et al., 2025; Castaldo and Monti, 2024). Our experiments show that while this approach proves effective for verbal idioms, it negatively affects the performance for light verbs, underscoring an inherent trade-off between overall generalization and improving specific VMWE categories.

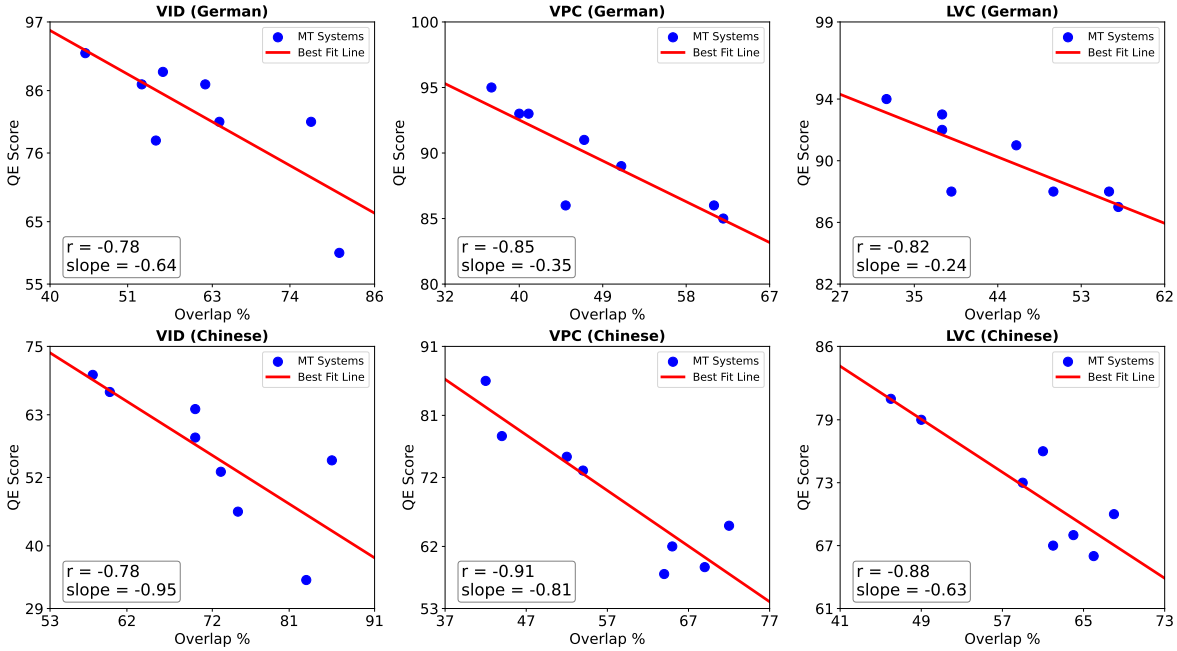

Figure 7: Error–span overlap vs. quality for German and Chinese. Each panel shows, for one VMWE type (VID, VPC, LVC), the relationship between the share of sentences whose xCOMET-QE error spans intersect the source VMWE (x-axis; lower is better) and the mean QE score (y-axis; higher is better). Blue points are MT systems; the red line is the least-squares fit. Insets report Pearson r and the fitted slope. Across languages and VMWE types, greater overlap is strongly associated with lower quality (negative r), with the steepest effects for idioms.

v

Model	Prompt
GPT-4.1 / GPT-5.1	`You are a professional machine translation system. Your job is to translate text from English into {target_lang_name}. For every input you receive, respond with ONLY the translated sentence in the target language. Do not include the source language, do not say "Translation", do not add quotes, explanation, or any extra text.`

Table 9: Instruction prompt used to obtain GPT-4.1 and GPT-5.1 translations in our experiments.

LLM paraphrasing has been a tool proposed to help MT systems handle idioms (Donthi et al., 2025; Castaldo and Monti, 2024). Building on this line of work, we investigate if paraphrasing helps improve translation across different VMWE types. By replacing the VMWE span with a literal counterpart, we get a semantically similar sentence without the VMWE. This also quantifies how much of the observed degradation is due to the VMWE phrase itself, rather than the sentence's semantic structure.

We provide Llama-3.3-70B with the sentence and the VMWE phrase, and use a chain-of-thought prompting strategy to guide the model in replacing the verbal expression with a literal paraphrase, while preserving the overall semantic structure of the sentence. For example, the sentence containing an idiom "*He finally decided to* ***spill the beans***" is paraphrased to "*He finally decided to* ***reveal the secret***." We then translate both original and paraphrased sentence to the target language. Given original sentence s_o, paraphrased sentence s_p, original sentence translation t_o, and paraphrased sentence translation t_p, we compute three QE scores as shown in Figure 9: Ori = $QE(s_o, t_o)$, Mix = $QE(s_o, t_p)$, and Para = $QE(s_p, t_p)$. We report two derived scores using MetricX-24:

i) δ_{mix} = Ori - Mix. This score shows if paraphrasing the source before translation improves translation quality. A positive value means the paraphrasing improves the translation, negative value means the paraphrasing hurts the translation.

(ii) δ_{para} = Ori - Para. This score shows how much degradation in performance is attributed to the

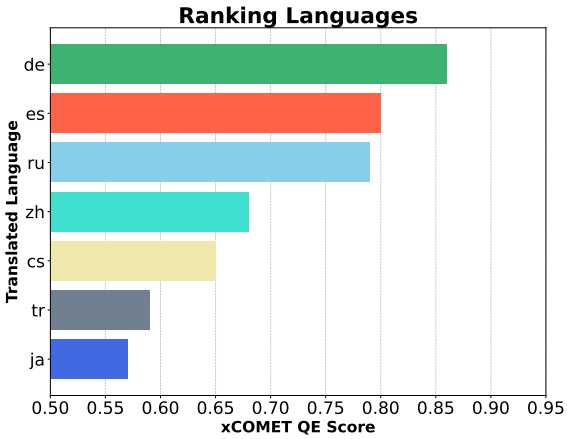

Figure 8: Ranking of languages based on the xCOMET QE scores.

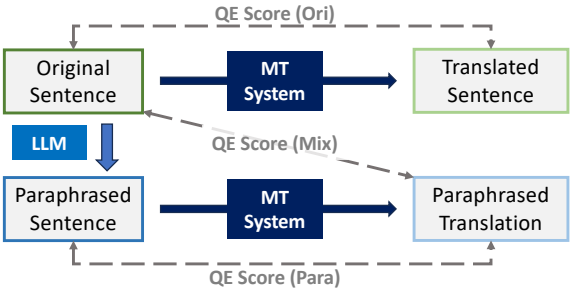

Figure 9: Paraphrasing structure. **Ori**: QE score between the original sentence and its direct translation. **Para**: QE score between paraphrased sentence and its direct translation. **Mix**: QE score between the original sentence and the translation of the paraphrased sentence.

Lang	VID		VPC		LVC	
	δ_{mix}	δ_{para}	δ_{mix}	δ_{para}	δ_{mix}	δ_{para}
Chinese	+0.70	+1.73	+0.24	+0.44	−0.04	+0.15
German	+0.54	+1.20	+0.24	+0.39	−0.17	+0.10
Russian	+0.63	+1.48	+0.48	+0.84	−0.30	+0.08
Czech	+1.06	+2.02	+0.51	+0.86	−0.29	+0.20
Spanish	+0.68	+1.49	+0.38	+0.68	−0.27	+0.10
Japanese	+0.12	+0.61	+0.12	+0.30	−0.17	+0.06
Turkish	+0.37	+1.46	+0.33	+0.77	−0.53	−0.07

Table 10: Paraphrasing results for MetricX-24 taking the average of all eight MT systems. δ_{mix}: Difference between original and mix translation (Ori − Mix). δ_{para}: Difference between original and paraphrased translation (Ori − Para). Positive δ value indicates paraphrasing improved the translation quality.

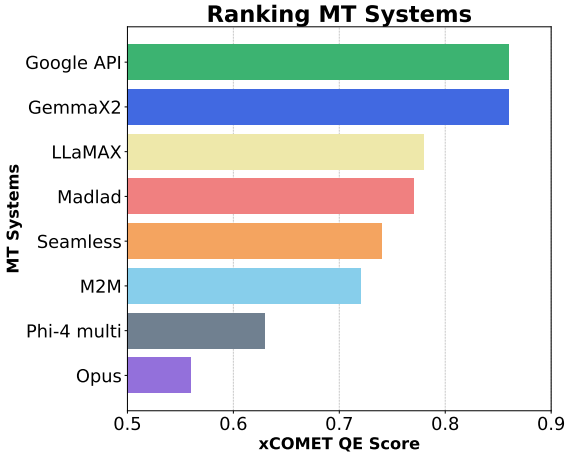

Figure 10: Ranking of MT systems based on the xCOMET QE scores.

VMWE candidate, as we compare two semantically similar sentences with and without the VMWE candidate. Higher values mean the paraphrased sentence translates better than the VMWE-containing original, estimating how much the VMWE degrades translation (values near zero imply minimal VMWE effect).

Table 10 reports the MetricX-24 results, full results for paraphrasing are given in Appendix N.1. The scores are calculated as the average of all eight MT systems. The values for δ_{para} are positive (except Turkish LVC), indicating that the paraphrased non-VMWE sentence translates a lot better than the original VMWE counterpart. The values are highest in VID, followed by VPC, while LVC shows the least improvement. This trend is consistent across all languages. The results suggest that VIDs, which are the most non-compositional, suffer the most during translation. LVCs, which are the most compositional, are relatively unaffected.

The δ_{mix} values indicates if paraphrasing improves translation performance. Evidently, VID shows consistent improvement across all languages, demonstrating that paraphrasing is an effective strategy for this VMWE category. VPC also modestly benefits. However, LVC performance consistently declines. Overall, these findings reveal a key trade-off: *paraphrasing enhances translation of highly non-compositional VMWEs but reduces domain-specific robustness, especially for compositional VMWE categories*. MT systems can leverage this strategy for VID and VPC cases to develop more robust, VMWE-aware translation systems.

N.1 Full Paraphrasing Results

Table 11 and 12 shows the full results of our paraphrasing experiment with MetricX and xCOMET as the evaluation model for the language pairs English to—Czech (*en–cs*), German (*en–de*), Russian (*en–ru*) and Chinese (*en–zh*). We don't report *en-cs* scores for GemmaX2 due to the high translation errors. The observations align with table 10. VID

	MT System	en-zh			en-de			en-ru			en-cs		
		Ori	δ_{mix}	δ_{para}	Ori	δ_{mix}	δ_{para}	Ori	δ_{mix}	δ_{para}	Ori	δ_{mix}	δ_{para}
VID	Madlad	7.25	+1.77	+2.21	3.00	+0.24	+0.92	6.49	+0.87	+1.76	7.02	+0.24	+1.23
	Seamless	6.98	+0.50	+0.94	3.72	+0.27	+1.01	6.72	+0.41	+1.32	8.65	+0.29	+1.32
	M2M	7.07	+1.29	+1.81	5.53	+1.01	+1.72	9.56	+1.66	+2.59	11.63	+1.63	+2.68
	Opus	7.95	+1.87	+2.32	6.87	+2.39	+3.10	9.50	+1.54	+2.45	13.41	+3.21	+4.24
	LLaMAX	4.76	+0.24	+0.72	4.49	+0.27	+0.90	6.96	+0.38	+1.21	10.59	+0.58	+1.48
	Phi-4-multi	4.63	+0.18	+0.58	4.22	+0.47	+0.97	7.34	+0.48	+1.22	14.39	+1.02	+1.68
	GemmaX2	3.31	−0.04	+0.41	2.41	−0.13	+0.54	4.69	−0.05	+0.73	-	-	-
	Google API	2.89	−0.15	+0.37	1.92	−0.19	+0.45	3.69	−0.25	+0.61	7.88	+0.49	+1.57
VPC	Madlad	6.46	+0.78	+0.92	2.44	+0.06	+0.35	5.72	+0.59	+0.97	6.71	+0.31	+0.73
	Seamless	6.22	+0.20	+0.39	3.15	+0.15	+0.44	6.55	+0.59	+0.96	8.39	+0.65	+1.12
	M2M	5.99	+0.45	+0.71	4.78	+0.70	+1.01	8.61	+1.15	+1.53	10.29	+1.16	+1.65
	Opus	6.48	+0.65	+0.90	4.82	+0.74	+1.07	8.33	+0.87	+1.27	10.90	+1.10	+1.58
	LLaMAX	4.40	+0.07	+0.28	3.87	+0.22	+0.43	6.64	+0.39	+0.76	9.64	+0.19	+0.49
	Phi-4-multi	4.12	−0.08	+0.07	3.68	+0.11	+0.33	6.82	+0.23	+0.59	13.78	+0.28	+0.50
	GemmaX2	2.86	−0.12	+0.07	2.18	+0.07	+0.32	4.48	+0.18	+0.43	-	-	-
	Google API	2.71	−0.02	+0.20	1.69	−0.07	+0.16	3.40	−0.11	+0.21	7.18	+0.42	+0.86
LVC	Madlad	5.56	+0.37	+0.59	2.09	−0.19	+0.09	4.59	−0.17	+0.21	5.78	−0.39	+0.12
	Seamless	6.00	−0.09	+0.13	2.70	−0.17	+0.14	5.41	−0.27	+0.12	7.00	−0.36	+0.16
	M2M	5.24	−0.12	+0.14	3.69	−0.32	+0.04	6.80	−0.37	+0.11	8.54	−0.37	+0.17
	Opus	5.71	−0.05	+0.20	3.95	+0.06	+0.39	6.90	−0.46	−0.01	9.49	−0.13	+0.42
	LLaMAX	4.26	−0.17	+0.04	3.67	−0.14	+0.15	6.00	−0.25	+0.14	8.82	−0.32	+0.16
	Phi-4-multi	4.28	−0.08	+0.02	3.14	−0.17	+0.05	6.06	−0.32	+0.02	12.47	−0.08	+0.29
	GemmaX2	2.94	−0.07	+0.10	1.82	−0.26	−0.00	3.88	−0.29	+0.06	-	-	-
	Google API	2.65	−0.16	+0.05	1.39	−0.23	+0.01	3.05	−0.34	+0.01	6.09	−0.41	+0.13

Table 11: Paraphrasing result. δ_{mix}: Difference between original and mix translation (Ori − Mix). δ_{para}: Difference between original and paraphrased translation (Ori − Para). Positive δ value indicates paraphrasing improved the translation quality.

improves the most with the paraphrasing, followed by VPC and LVC. xCOMET seems to be more strict regarding the VID phrase paraphrasing, having mixed δ_{mix} values for all three categories, as compared to MetricX-24 which had mostly positive values in VID and VPC. Some MT systems gain substantially from paraphrasing, like Madlad (*en-zh*) in VID gaining 22.22 and 37.31 in δ_{mix} and δ_{para}.

Table 13 presents the paraphrasing results for MetricX-24 for the language pairs English to—Spanish (*en-es*), Japanese (*en-ja*) and Turkish (*en-tr*). Table 14 presents the results for xCOMET. Japanese has a very poor translation scores across both evaluation models for Opus. The trend for these three languages align with our primary observation, with MetricX-24 having generally favorable δ_{mix} scores, while xCOMET is more strict, having mixed results. δ_{para} improves the scores across both the evaluations, with the VID, VPC and LVC gaining the most in that order.

O LLM Prompts for VMWE Extraction

Table 15 shows the prompts we used for GPT-4o, Llama-3.3-70B-Instruct, Phi-4-multi and DeepSeek-R1-Distill-Llama-70B classifying VMWE candidate sentences from the WMT dataset. The VMWE candidate sentences are collected as discussed in Appendix H

P LLM Prompts for VMWE Candidate Paraphrasing

Table 16 shows the prompts we used for paraphrasing the VMWE candidate. We use Llama 3.3 70B for this experiment, with temperature 0.9 and top-p value of 0.9. We use one or two few-shot examples in each category.

“‘latex

	MT System	en-zh			en-de			en-ru			en-cs		
		Ori	δ_{mix}	δ_{para}	Ori	δ_{mix}	δ_{para}	Ori	δ_{mix}	δ_{para}	Ori	δ_{mix}	δ_{para}
VID	Madlad	35.55	+22.25	+37.37	89.15	−2.24	+5.89	77.14	+1.43	+9.63	68.15	−3.55	+18.62
	Seamless	53.31	−2.76	+16.72	87.07	−2.36	+6.01	76.37	−0.21	+8.51	62.18	−4.01	+18.94
	M2M	56.49	+0.08	+17.05	81.16	−0.57	+8.81	67.25	+3.48	+12.47	51.55	+1.15	+23.40
	Opus	46.00	+2.91	+23.56	57.91	+22.93	+31.96	62.63	+6.73	+15.79	39.38	+10.90	+33.96
	LLaMAX	66.10	−3.83	+13.40	81.70	+0.49	+9.51	73.78	+0.99	+9.40	51.33	−1.31	+21.46
	Phi-4-multi	60.87	−2.88	+15.86	78.44	−2.04	+11.19	68.48	−0.06	+11.84	25.48	−0.31	+13.60
	GemmaX2	73.97	−4.15	+11.32	90.26	−2.33	+5.40	83.03	−1.08	+6.12	-	-	-
	Google API	71.32	−1.42	+15.61	91.43	−2.52	+5.09	84.19	−0.45	+7.58	62.09	+0.63	+22.57
VPC	Madlad	61.56	+4.29	+8.79	93.01	−0.16	+2.02	82.53	+1.81	+4.11	78.93	+0.00	+6.94
	Seamless	64.49	+0.51	+6.36	91.15	−0.29	+1.97	79.87	+1.43	+3.89	72.61	+0.69	+8.14
	M2M	66.83	+1.89	+6.67	85.66	+1.33	+4.10	73.92	+2.23	+4.83	65.80	+2.05	+9.60
	Opus	60.99	+2.63	+8.59	86.40	+1.52	+4.16	73.72	+1.80	+4.66	63.03	+2.22	+9.63
	LLaMAX	75.44	−0.53	+4.00	89.52	−0.17	+2.32	79.70	+0.53	+2.92	66.27	−0.96	+5.58
	Phi-4-multi	73.64	−1.52	+3.84	86.85	−0.34	+2.99	76.96	+0.31	+3.45	34.41	−0.72	+3.89
	GemmaX2	82.62	−0.80	+3.31	93.81	−0.11	+2.03	86.33	+0.03	+2.09	-	-	-
	Google API	84.26	−1.38	+3.19	95.26	−0.47	+1.37	89.91	−0.39	+1.63	76.37	+0.67	+7.32
LVC	Madlad	70.56	+0.92	+3.51	94.62	−0.72	+0.56	86.65	−0.67	+0.70	85.63	−2.86	+1.39
	Seamless	68.32	−0.84	+2.05	92.96	−0.71	+0.51	84.29	−0.84	+0.75	80.17	−1.85	+2.41
	M2M	72.47	−1.49	+1.29	89.76	−0.59	+0.72	78.63	−0.59	+1.09	74.46	−1.83	+2.50
	Opus	68.38	−1.76	+1.56	89.21	−0.06	+1.27	77.78	−1.04	+0.73	70.05	−1.17	+3.53
	LLaMAX	78.04	−1.88	+0.93	90.56	−0.69	+0.64	81.30	−0.70	+0.75	72.22	−2.04	+2.34
	Phi-4-multi	75.57	−1.37	+1.21	89.81	−1.23	+0.51	79.38	−1.12	+0.53	38.74	−1.14	+1.62
	GemmaX2	84.44	−1.28	+1.10	95.39	−0.85	+0.11	89.20	−0.75	+0.26	-	-	-
	Google API	86.02	−1.83	+0.53	96.71	−0.92	+0.23	91.76	−0.77	+0.56	83.70	−2.54	+1.75

Table 12: Paraphrasing Result for xCOMET. All values are multiplied by 100 for better readability. δ_{mix}: Difference between mixed translation and original translation (Mix - Ori). δ_{para}: Difference between paraphrased translation and original translation (Para - Ori). Positive δ value indicates paraphrasing improved the translation quality.

	MT System	en-es			en-ja			en-tr		
		Ori	δ_{mix}	δ_{para}	Ori	δ_{mix}	δ_{para}	Ori	δ_{mix}	δ_{para}
VID	Madlad	4.86	+0.44	+1.29	6.39	−0.27	+0.34	7.67	+0.05	+1.24
	Seamless	5.35	+0.54	+1.39	8.17	+0.15	+0.87	8.53	+0.28	+1.52
	M2M	7.76	+1.46	+2.29	8.02	+0.92	+1.52	11.93	+1.22	+2.42
	Opus	7.22	+1.73	+2.57	21.11	+0.10	−0.01	-	-	-
	LLaMAX	5.76	+0.43	+1.18	-	-	-	-	-	-
	Phi-4-multi	5.61	+0.55	+1.23	6.78	+0.14	+0.61	14.23	+0.75	+1.40
	Google API	3.38	−0.34	+0.51	4.39	−0.29	+0.33	5.69	−0.45	+0.72
VPC	Madlad	4.19	+0.28	+0.63	6.40	+0.02	+0.25	7.11	+0.21	+0.66
	Seamless	4.78	+0.41	+0.75	7.86	+0.32	+0.56	7.67	+0.18	+0.67
	M2M	6.79	+1.09	+1.47	7.28	+0.50	+0.76	10.71	+0.78	+1.28
	Opus	5.91	+0.97	+1.29	21.00	−0.07	−0.10	-	-	-
	LLaMAX	5.08	+0.19	+0.51	-	-	-	-	-	-
	Phi-4-multi	4.93	+0.18	+0.49	6.54	+0.10	+0.21	13.63	+0.48	+0.78
	Google API	3.17	−0.05	+0.29	4.20	−0.11	+0.15	5.50	+0.04	+0.50
LVC	Madlad	3.39	−0.31	+0.07	6.16	−0.26	+0.01	6.19	−0.55	−0.06
	Seamless	3.87	−0.31	+0.09	7.51	−0.19	+0.10	6.90	−0.48	+0.08
	M2M	5.12	−0.29	+0.13	6.63	−0.12	+0.18	8.95	−0.54	−0.03
	Opus	4.60	−0.21	+0.21	22.27	−0.09	+0.04	-	-	-
	LLaMAX	4.33	−0.32	+0.01	-	-	-	-	-	-
	Phi-4-multi	4.26	−0.15	+0.19	6.12	−0.16	+0.05	11.61	−0.52	−0.21
	Google API	2.69	−0.35	+0.02	4.10	−0.23	+0.03	4.71	−0.60	−0.13

Table 13: Paraphrasing Results for MetricX-24. Positive δ value indicates paraphrasing improved the translation quality.

	MT System	en-es			en-ja			en-tr		
		Ori	δ_{mix}	δ_{para}	Ori	δ_{mix}	δ_{para}	Ori	δ_{mix}	δ_{para}
VID	Madlad	78.74	−1.71	+11.82	42.66	+6.06	+32.80	60.63	−2.28	+20.38
	Seamless	78.33	−2.77	+11.00	39.92	−2.65	+26.36	59.56	−3.31	+19.39
	M2M	69.26	+0.87	+15.34	52.41	−0.61	+21.95	47.39	−0.30	+22.54
	Opus	53.99	+19.28	+32.56	18.32	+0.11	−2.04	-	-	-
	LLaMAX	74.27	+0.08	+13.54	-	-	-	-	-	-
	Phi-4-multi	64.07	−2.55	+18.69	52.38	−4.00	+19.11	27.28	−0.47	+12.74
	Google API	84.61	−4.68	+8.59	76.05	−7.29	+14.92	71.86	−7.09	+15.64
VPC	Madlad	86.40	+0.83	+4.28	68.79	−1.45	+6.54	73.70	−0.34	+6.78
	Seamless	84.50	+1.18	+4.83	56.25	−0.32	+7.69	73.24	−0.60	+6.15
	M2M	77.33	+3.48	+7.47	66.84	+0.37	+7.16	62.00	+0.72	+7.88
	Opus	80.56	+2.78	+6.69	15.97	−0.19	−0.07	-	-	-
	LLaMAX	83.98	+0.57	+4.13	-	-	-	-	-	-
	Phi-4-multi	77.47	+0.63	+5.95	65.91	−1.11	+5.12	34.47	+0.03	+4.82
	Google API	89.79	−0.19	+3.21	86.36	−1.77	+4.42	82.34	−1.45	+4.66
LVC	Madlad	90.52	−1.40	+0.75	76.39	−3.34	+1.26	80.70	−2.88	+0.73
	Seamless	89.32	−1.57	+0.51	63.97	−2.79	+1.42	78.54	−2.52	+1.05
	M2M	84.43	−1.28	+0.81	73.13	−1.77	+1.84	69.81	−2.59	+0.61
	Opus	85.87	−1.04	+1.05	16.52	−0.00	−0.01	-	-	-
	LLaMAX	87.94	−1.36	+0.45	-	-	-	-	-	-
	Phi-4-multi	83.51	−2.02	+0.85	70.53	−2.33	+1.01	42.98	−2.69	+0.05
	Google API	93.04	−1.65	+0.27	89.96	−2.49	+0.64	88.17	−3.23	+0.21

Table 14: Paraphrasing Result for xCOMET. Positive δ value indicates paraphrasing improved the translation quality.

Table 15: Prompts for VMWE extraction from WMT dataset.

VMWE	Prompt
LVC	Linguistic Analysis Task: Multi-choice Classification of Verb-Noun Constructions. Context Sentence: sentence Target Components:- VERB: verb_lemma - NOUN: noun_lemma LVC-specific decision tree: Apply test LVC.0 - [N-ABS: Is the noun abstract?] No -> It is not an LVC, exit Unsure -> Apply test LVC.1 - [N-PRED: Is the noun predicative?] No -> It is not an LVC, exit Yes/Unsure -> Apply test LVC.2 - [V-SUBJ-N-ARG: Is the subject of the verb a semantic argument of the noun?] No -> It is not an LVC, exit Yes/Unsure -> Apply test LVC.3 - [V-LIGHT: The verb only adds meaning expressed as morphological features?] No -> It is not an LVC, exit Yes -> Apply test LVC.4 - [V-REDUC: Can a verbless NP-reduction refer to the same event/state?] No -> It is not an LVC, exit Yes -> It is an LVC Classification Options: A: the noun is not abstract. B: the noun is not predicative (i.e. lacks semantic arguments). C: All tests pass – the construction qualifies as an LVC. D: the verb's subject is not a semantic argument of the noun. E: the verb adds more meaning than mere morphological features. F: a verbless NP cannot be formed referring to the same event/state. Instructions: Provide your classification choice (A, B, C, D, E, or F) with brief reasoning in 3 sentences or less, ending with 'Final Answer: [Choice]' Please provide your classification with reasoning as instructed.

(Continued on next page)

(Continued from previous page)

VMWE	Prompt
VPC	Determine if the verb-particle combination is a VPC (phrasal verb) based on the following analysis. Context: sentence Combination: verb_lemma particle VPC Decision Tree: 1. Is the second element a particle (e.g., 'up', 'off')? - No → Not VPC (A) - Yes → Continue 2. Remove the particle from the combination. Does the remaining verb convey the same meaning as the full verb-particle phrase? - Yes → Not VPC (B) - No → Continue 3. Does the inclusion of the particle create a non-compositional meaning that is significantly different from the verb's original meaning? - No → Not VPC (C) - Yes → VPC (D) A: Not a particle B: Meaning remains similar without the particle C: Particle does not significantly alter the meaning "D: Valid VPC (Particle significantly alters meaning) Answer with reasoning and 'Final Answer: [answer] "Is this a valid VPC? Provide analysis.
VID	Verbal Idiom (VID) Classification Task Context Sentence: sentence Candidate Phrase: candidate VID Decision Tree: 1. [CRAN] Contains cranberry word? Yes → VID No → Next test 2. [LEX] Regular replacement changes meaning? Yes → VID No → Next test 3. [MORPH] Morphological changes affect meaning? Yes → VID No → Next test 4. [MORPHSYNT] Morphosyntactic changes affect meaning? Yes → VID No → Next test 5. [SYNT] Syntactic changes affect meaning? Yes → VID No → Not VID Examples: - VID: 'kick the bucket', 'let the cat out of the bag' - Non-VID: 'take a walk', 'make a decision' Instructions: 1. Analyze each test sequentially 2. Provide brief reasoning for each test 3. Conclude with 'Final Answer: [Yes/No]' Is this candidate a Verbal Idiom (VID)? Apply the decision tree.

Table 16: Prompts for VMWE candidate paraphrasing.

VMWE	Prompt
LVC	You are an expert in linguistics. Given a sentence containing a multi-word expression (VMWE), a Light Verb Construct (LVC). Your task is to rephrase the sentence to remove the VMWE while keeping the meaning intact. Think in the following steps. Step 1 - Look at the phrase in the sentence that will be provided to you that constitutes the VMWE in the sentence. Step 2 - Replace that phrase with an alternative wording that removes the VMWE while maintaining the sentence's meaning. Step 3 - Provide your output in the following format: Rephrased Sentence: [Sentence without VMWE] —Few-shot 1– Sentence: She gave a smile before walking away. ‖ Phrase: gave a smile Step 1 - Provided VMWE phrase: The given phrase "gave a" is a light-verb construction (LVC), where "gave" (a light verb) combines with "smile" (a noun) to describe the act of smiling. Step 2 - Replace the VMWE: A more lexicalized way to express this would be "smiled." Step 3 - Output Format: Rephrased Sentence: She smiled before walking away. —Few-shot 2– He took a deep breath before speaking. ‖ Phrase: took a deep breath Step 1 - Provided VMWE phrase: The given phrase "took a deep breath" is a light verb construction (LVC), where "take" (a light verb) combines with "breath" (a noun) to describe the act of inhaling. Step 2 - Replace the VMWE: A more lexicalized way to express this would be "breathed deeply. Step 3 - Output Format: Rephrased Sentence: He breathed deeply before speaking. Sentence: *sentence* ‖ Phrase: *candidate*
VPC	You are an expert in linguistics. Given a sentence containing a multi-word expression (VMWE), a Verb Particle Construct (VPC). Your task is to rephrase the sentence to remove the VMWE while keeping the meaning intact. Think in the following steps. Step 1 - Look at the phrase in the sentence that will be provided to you that constitutes the VMWE in the sentence. Step 2 - Replace that phrase with an alternative wording that removes the VMWE while maintaining the sentence's meaning. Step 3 - Provide your output in the following format: Rephrased Sentence: [Sentence without VMWE] —Few-shot 1– Sentence: His eyes welled up during the song. ‖ Phrase: welled up Step 1 - Provided VMWE phrase: The given phrase "welled up" is a verb-particle construction (VPC), where "well" (verb) combines with "up" (particle) to indicate the accumulation or rising of liquid, often referring to tears forming in the eyes. Step 2 - Replace the VMWE: A more literal way to express this would be "filled with tears." Step 3 - Output Format: Rephrased Sentence: His eyes filled with tears during the song. Sentence: *sentence* ‖ Phrase: *candidate*

(Continued on next page)

(Continued from previous page)

VMWE	Prompt
VID	You are an expert in linguistics. Given a sentence containing a multi-word expression (VMWE), a Verbal Idiom (VID). Your task is to rephrase the sentence to remove the VMWE while keeping the meaning intact. Think in the following steps. Step 1 - Look at the phrase in the sentence that will be provided to you that constitutes the VMWE in the sentence. Step 2 - Replace that phrase with an alternative wording that removes the VMWE while maintaining the sentence's meaning. Step 3 - Provide your output in the following format: Rephrased Sentence: [Sentence without VMWE] —Few-shot 1– Sentence: I think we still might take them up on it. ‖ Phrase: take on [pron] Step 1 - Provided VMWE phrase: The given phrase "take them up on" is a verbal idiom (VID), where "take up on" is an idiomatic expression meaning to accept an offer or invitation. Step 2 - Replace the VMWE: A more literal way to express this would be "accept their offer." Step 3 - Output Format: Rephrased Sentence: I think we still might accept their offer. —Few-shot 2– Sentence: I did not pay much attention to it back then. ‖ Phrase: pay to [pron] Step 1 - Provided VMWE phrase: The given phrase "pay attention to" is a verbal idiom (VID), where "pay" (verb) combines with "attention" to mean focusing or noticing something. Step 2 - Replace the VMWE: A more literal way to express this would be "notice." Step 3 - Output Format: Rephrased Sentence: I did not notice it much back then. Sentence: *sentence* ‖ Phrase: *candidate*