Recibido / Received: 26/05/2025
Aceptado / Accepted: 08/03/2026

Para enlazar con este artículo / To link to this article:
https://dx.doi.org/10.6035/MonTI.2026.18.04

Para citar este artículo / To cite this article:
HIDALGO-TERNERO, Carlos Manuel & Vicent BRIVA-IGLESIAS. (2026) "ChatGPT vs. DeepL vs. Google Translate: a human evaluation of multiword expressions' machine translation quality." In: GONZÁLEZ PASTOR, Diana & Celia RICO PÉREZ (eds.) 2026. *La Inteligencia Artificial al servicio de la labor de la traducción: investigaciones en torno a los nuevos avances tecnológicos / Artificial Intelligence at the service of translation: research on new technological achievements*. *MonTI* 18, pp. 120-145.

CHATGPT VS. DEEPL VS. GOOGLE TRANSLATE: A HUMAN EVALUATION OF MULTIWORD EXPRESSIONS' MACHINE TRANSLATION QUALITY

CARLOS MANUEL HIDALGO-TERNERO
cmhidalgo@uma.es
University of Malaga, IUITLM (Spain)
https://orcid.org/0000-0002-8338-2627

VICENT BRIVA-IGLESIAS
vicent.brivaiglesias@dcu.ie
Dublin City University, CTTS, SALIS, ADAPT Centre (Ireland)
https://orcid.org/0000-0001-8525-2677

Abstract

Multiword expressions (MWEs) remain a persistent challenge in neural machine translation (NMT), particularly when they appear in discontinuous forms. In this context, the present study evaluates the ability of general-purpose large language models (LLMs) to address these limitations by systematically comparing the performance of Google Translate, DeepL, and GPT-4o in translating Spanish-to-English MWEs. A dataset of 600 examples—balanced between continuous and discontinuous MWEs—was machine translated and manually evaluated by two expert linguists. The results indicate that GPT-4o statistically significantly outperforms NMT systems in both forms of the MWEs, while DeepL and Google Translate exhibit substantial declines in performance for discontinuous MWEs. The results from this study show that general-purpose LLMs handle syntactic flexibility and idiomaticity more effectively than traditional NMT approaches. The study also contributes an open-source dataset and invites further research on LLM applications in MT.

Keywords: ChatGPT. DeepL. Google Translate. Multiword Expressions. Machine Translation.

Resumen

Las expresiones pluriverbales (EP) siguen suponiendo un desafío para la traducción automática neuronal (TAN), especialmente en sus formas discontinuas. En este contexto, el presente estudio evalúa la capacidad de los modelos de lenguaje de gran tamaño (LLM, por sus siglas en inglés) de propósito general para abordar estas limitaciones, mediante una comparación sistemática del rendimiento de Google Translate, DeepL y GPT-4o en la traducción de EP del español al inglés. Para ello, se tradujo automáticamente y se evaluó manualmente, por lingüistas expertos, un conjunto de datos (de código abierto) compuesto por 600 ejemplos, equilibrado entre EP continuas y discontinuas. Los resultados indican que GPT-4o supera estadísticamente a los sistemas TAN para ambos tipos de EP, así como revelan que los LLM tratan la flexibilidad sintáctica y la idiomaticidad con mayor eficacia que la TAN, abriendo la puerta a nuevas investigaciones sobre las implicaciones de los LLM en la traducción automática.

Palabras clave: ChatGPT. DeepL. Google Translate. Expresiones Pluriverbales. Traducción Automática.

1. Introduction

Despite the advancements of neural machine translation (NMT) in the last decade and its establishment as the state-of-the-art machine translation (MT) paradigm, errors still persist and, in numerous instances, NMT output requires human edits before dissemination (Briva-Iglesias 2023; Kenny 2022). For instance, multiword expressions (MWEs) remain one of the most persistent challenges for NMT systems. MWEs, such as the Spanish *llevarse el gato al agua* (literal translation, "to carry the cat to the water"; figurative meaning, "to win the day"), introduce unique difficulties, as they can appear both in their continuous (the constituents of the MWE occur together) or discontinuous form (the constituents of the MWE are split by other elements not belonging to the MWE, as in "take your brother's advice into account"). The non-adjacent occurrence of MWEs' constituents can hinder their automatic detection and translation by MT systems. Regardless of significant advances in NMT, the translation of discontinuous MWEs

continues to be one of their Achilles' heels (Constant *et al.* 2017; Ramisch & Villavicencio 2022; Rohanian *et al.* 2019).

The emergence of large language models (LLMs) has introduced a new paradigm for MT (Hendy *et al.* 2023). Unlike traditional NMT systems trained exclusively for translation tasks, LLMs operate as a general-purpose language model capable of direct translation through context-aware prompts (Wang *et al.* 2023), as well as many other tasks, such as summarisation, plain language editing, coding, etc. (Brown *et al.* 2020; Briva-Iglesias & Peñuelas Gil 2025). General-purpose LLMs' potential for handling linguistic complexities such as MWEs requires a systematic evaluation, particularly in comparison to state-of-the-art NMT systems, as there is a paucity of literature in this regard due to the relative novelty of such systems.

In this context, this study aims to evaluate the performance of two state-of-the-art NMT systems (Google Translate and DeepL, in its classic NMT mode) against a state-of-the-art general-purpose LLM (GPT-4o) in translating Spanish-to-English MWEs. By focusing on both continuous and discontinuous forms of MWEs, we seek to determine the extent to which general-purpose LLMs can address the limitations of traditional NMT systems. To guide this investigation, we pose the following research questions:

- RQ1. How do Google Translate, DeepL, and GPT-4o handle MWEs in their continuous and discontinuous forms from Spanish into English?
- RQ2. Do LLMs outperform traditional NMT systems in translating MWEs?

The remainder of this paper is structured as follows. Section 2 reviews previous work on MWEs and their challenges in MT. Section 3 describes the methodology employed to evaluate the translation performance of the three systems under study. Section 4 presents and discusses the results of the evaluation. Finally, Section 5 concludes the paper with insights into the implications of this study, limitations, and directions for future research.

2. Previous and Related Work

MWEs are an inherent feature of natural language, characterised by their varying degrees of non-compositionality, syntactic irregularity, and fixedness. These linguistic properties make MWEs particularly challenging for computational processing, including MT (Constant *et al.* 2017; Corpas Pastor 2013; Ramisch & Villavicencio 2022; Rohanian *et al.* 2019). In NMT, the problem is exacerbated by the nature of the models, which rely on large-scale parallel corpora (Pérez-Ortiz, Forcada & Sánchez-Martínez 2022). Idioms often occur less frequently in such datasets, leading to insufficient representation in NMT training. Furthermore, discontinuous MWEs introduce structural variations that are difficult for NMT models to capture effectively, resulting in literal translations that fail to convey the intended meaning (Zaninello & Birch 2020), as we will be able to observe in Section 4 (see Tables 1, 2, and 3). Prior approaches to mitigate these issues have included pre-processing techniques, such as aligning MWEs in source and target languages or converting discontinuous MWEs into their continuous forms (e.g., Hidalgo-Ternero & Corpas Pastor 2020). Some tools for pre-processing MWEs before sending them to a NMT system have been developed, such as gApp, which can automatically convert discontinuous MWEs into their continuous forms (see Hidalgo-Ternero & Corpas Pastor 2020; Hidalgo-Ternero & Zhou-Lian 2022), and Paidiom (Hidalgo-Ternero & Lima-Florido 2023), which, besides this conversion into the continuous form, can also translemmatise MWEs, i.e., to directly convert MWEs into their target-text equivalents to improve NMT. These tools have demonstrated an average improvement in MT accuracy for MWEs, more specifically an increase of 14.1 percentage points with gApp and 83 percentage points with Paidiom. However, they require extensive manual effort and often rely on language-specific heuristics that do not generalise well across languages, MWE typologies, or domains.

Despite the recent surge in popularity of LLMs, substantial research has already been conducted on their application to MT. Several studies have demonstrated that LLMs can function as highly accurate MT systems, achieving results comparable to those of NMT systems (Hendy *et al.* 2023; Jiao *et al.* 2023). The analysis of LLMs for MT has been diverse, covering a

wide range of applications and challenges. Researchers have explored how LLMs can be used as adaptive MT systems, capable of fine-tuning their outputs based on specific user needs or domain requirements (Moslem *et al.* 2023). Other studies have focused on the ability of LLMs to maintain contextual coherence across long documents, a task that has traditionally been challenging for NMT systems (Castilho *et al.* 2023; Karpinska & Iyyer 2023; Wang *et al.* 2023). Additionally, LLMs have been evaluated for their performance in translating specialised texts, such as legal documents, across both major and minor language pairs, demonstrating their versatility and robustness in handling domain-specific terminology and complex linguistic structures (Briva-Iglesias, Dogru & Camargo 2024). However, to the best of our knowledge, no study has yet investigated how general-purpose LLMs handle one of the most persistent challenges in MT: the translation of discontinuous MWEs. This raises a critical question: How will LLMs behave when faced with the translation of MWEs, particularly in their discontinuous forms? This study seeks to fill this gap by systematically evaluating the performance of general-purpose LLMs in comparison to traditional NMT systems when translating MWEs, with the aim of exploring whether LLMs can overcome one of the main limitations of NMT.

It is also worth stressing that evaluating the translation of MWEs remains a significant challenge in MT evaluation. While automatic metrics such as BLEU and METEOR are widely used to assess overall MT quality (Kocmi *et al.* 2021), they are often inadequate for capturing the complexities of MWEs, especially idioms:

> Global metrics, such as BLEU (Papineni *et al.* 2001), consider the full translation, and thus, the effects of idiom translation are overshadowed. Previous efforts on targeted evaluation isolate the idiom translation using word alignments (Fadaee *et al.* 2018) or word edit distance (Zaninello & Birch 2020). These approaches measure the accuracy of idiom translation but do not account for literal translation errors. Shao *et al.* (2018) proposed a method for estimating the frequency of such errors, but it requires the creation of language-specific handcrafted lists (i.e., blocklists) with words that correspond to literal translation errors. (Baziotis, Mathur & Hasler 2023)

Given these limitations, manual evaluation by expert human annotators remains the most reliable method for exclusively assessing the phenomenon of MWE translation. Human evaluators can consider factors such as idiomaticity, cultural appropriateness, and the preservation of meaning across languages, which automatic metrics cannot capture to the same extent. However, manual evaluation is resource-intensive and time-consuming. Creating representative datasets that include MWEs in their natural context adds another layer of complexity and requires a careful selection of examples, especially for discontinuous forms, which are often underrepresented in parallel corpora (Zhou-Lian, Corpas Pastor & Hidalgo-Ternero 2024).

3. Methodology

This section outlines the research methodology employed to evaluate the performance of Google Translate, DeepL, and GPT-4o in translating Spanish-to-English MWEs, with a focus on both continuous and discontinuous forms. To answer our Research Questions, we designed a systematic evaluation framework that combines corpus-based analysis with human evaluation. The methodology is structured into three main components: (1) the selection and preparation of the MWEs under study, (2) the description of the MT systems evaluated, and (3) the evaluation process itself.

3.1 The MWEs under study

For this study, we selected a set of six Spanish MWEs that are representative of the challenges posed by idiomatic expressions, particularly those involving non-compositionality and syntactic flexibility. The selected MWEs include:

1. *Haber gato encerrado*: literally, "there is a locked-up cat"; figuratively, "there's something fishy going on."
2. *Ser cuatro gatos*: literally, "to be four cats"; figuratively, "to be just a bunch of people."
3. *Llevarse el gato al agua*: literally, "to carry the cat to the water"; figuratively, "to win the day."

4. *Pagar el pato*: literally, "to pay the duck"; figuratively, "to take the rap."
5. *Dormir la mona*: literally, "to sleep the female monkey"; figuratively, "to sleep something off."
6. *Ganar/costar/pagar... cuatro perras*: literally, "to earn/cost/pay... four female dogs"; figuratively, "to earn/cost/pay... peanuts."

These MWEs were chosen because they exhibit a range of linguistic properties, including idiomaticity, cultural specificity, and the potential for both continuous and discontinuous forms. Replicating the collection methodology of Hidalgo-Ternero & Corpas Pastor (2020), the concordances containing the discontinuous MWEs under study were retrieved from the Sketch Engine corpora esTenTen23 (over 33 billion tokens) and Timestamped JSI web corpus 2014-2021 Spanish (over 18 billion tokens), ensuring a representative sample of real-world usage (Kilgarriff *et al.* 2014). Despite the challenges still posed by ubiquitous source-text error, noise and out-of-vocabulary tokens in user-generated content (UGC) for even the most robust NMT systems (Belinkov & Bisk 2018; Lohar *et al.* 2019), a heterogeneous sample in terms of language varieties, text sources and types (including UGC) was selected for the analysis to alleviate sampling bias, which could otherwise originate from exclusively examining canonical NMT training data for these MWEs.

The final analysed dataset comprises 100 examples for each idiom (600 examples in total), including 50 continuous and 50 discontinuous examples for every idiom (in total, 300 continuous and 300 discontinuous examples). These examples did not only consist of the idiom by itself, but also the preceding and following sentences, so that the MT systems can have the context for translation (some examples are provided in Section 4, Tables 1, 2, and 3).[1]

In this context, the present study is designed as a controlled, exploratory investigation into the translation of discontinuous idioms, aiming to prioritise internal validity and methodological transparency over broad

1. The complete dataset, including the source texts, the different MT proposals and the human evaluation is shared open source in Zenodo to allow other researchers to further investigate the topic or replicate this study: <https://zenodo.org/records/18351134>.

generalisation. To this end, the analysis is restricted to a limited number of idiom types while systematically varying their contextual realisations, allowing specific variables—such as syntactic discontinuity and contextual sensitivity—to be examined in isolation. The focus on verbal idioms is motivated by their syntactic flexibility and their interaction with tense, aspect, and argument structure, which make them particularly challenging for machine translation systems and less comparable to nominal or adjectival idioms (Hidalgo-Ternero 2020). Within this category, the study concentrates on zoologisms (MWEs containing lexemes that refer to animal names, so-called "zoonyms") due to their high degree of conventionalisation, cross-linguistic availability, figurative autonomy and cohesion, and the existence of well-established idiomatic equivalents, which facilitate expert evaluation and reduce uncertainty regarding idiomatic status (see Kekić 2008; Luque Nadal 2012; Škvárová & Šlechta 2015, among others).

3.2 The MT systems

Three state-of-the-art MT systems were evaluated in this study. On the one hand, Google Translate (GT) was chosen, since it is one of the most widely-used NMT systems globally, especially in the English-Spanish language pair (Rivera-Trigueros 2022). DeepL was also selected as one of the best-performing NMT systems to date (Kamaluddin *et al.* 2024). Since the API version of DeepL allows for selecting between the "Classic language model" (NMT, according to the company) and the "Next-gen language model" (a translation-specific LLM, according to the company) and we wanted to assess its NMT version for the sake of this study (i.e., contrasting general-purpose LLMs vs. traditional NMT), we selected the "Classic language model" of DeepL. The translations of both GT and DeepL were generated through their commercial APIs, providing each of the examples at a time to ensure that the systems only considered the context for each of the examples. Then, MT results were pasted into a spreadsheet to create the dataset.

As per the general-purpose LLM, we generated the MT output through the gpt-4o-2024-11-20 model of OpenAI. This was the most recent model available at the time of writing (January 2025) and was ranked as the third

best closed model in the ChatBot Arena LLM Leaderboard (Chiang *et al.* 2024) with an Arena score of 1365 in late January 2025. The prompt used in the API call was "Please translate the following text into English: [TEXT]," as per the results of Jiao *et al.* (2023). Even if using the same prompt may generate different results when using a specific LLM, we include the prompt here for replicability. Since the main goal of the paper was to evaluate a significant number of translations (600 examples) via human evaluation, we preferred to scale the number of examples with only one prompt.

These systems were selected to represent both traditional NMT approaches (GT and DeepL) and the emerging paradigm of general-purpose LLMs (GPT-4o). The goal was to compare their performance in handling the complexities of MWEs, particularly in their discontinuous forms.

3.3 The Evaluation and Analysis Process

Two professional Spanish-to-English translators with +5 years of experience were recruited to manually assess the translations. The evaluators used a binary scale (1 = good, 0 = bad) to rate the quality of each translation, focusing exclusively on the accurate rendering of the MWEs in the target text. The evaluators were instructed to pay particular attention to the translation of discontinuous MWEs, assessing whether the intervening elements were appropriately handled and whether the idiomatic meaning was preserved.

For MWEs in their continuous form, if the system identified and adequately translated the Spanish MWE into an accepted and culturally valid English form, the score would be 1 (e.g., "Él gana cuatro perras." > "He earns peanuts."); otherwise, it would be 0 (e.g., "Él gana cuatro perras." > "He earns four dogs.").

For MWEs in their discontinuous form, only the MT outputs that identified the Spanish MWE and appropriately translated it into an accepted and culturally valid English form, while respecting the intervening element, were scored 1 (e.g., "Ellos se llevaron ayer el gato al agua." > "They came out on top yesterday."). If the MT systems identified and translated adequately an MWE, but the intervening element was omitted, the translation was considered as incorrect (score 0) (e.g., "Ellos se llevaron ayer el

gato al agua." > "They came out on top.," omitting the intervening element "ayer"/"yesterday").

To ensure that the human evaluation results were reliable, inter-annotator agreement was measured with Cohen's kappa, giving a result of 1 (Artstein 2017). This result demonstrates that both reviewers entirely agreed with the scores for the different examples in the dataset, and even if this agreement is unusually high for MT evaluation methods such as MQM and/or Adequacy/Fluency ratings (Rossi & Carré 2022), achieving such a high agreement is easier with a binary evaluation scale.

Then, the results of the human evaluation were analysed using a mixed-methods approach. First, a 3x2 repeated measures ANOVA was conducted to determine whether there were statistically significant differences in the performance of the three MT systems (three levels: DeepL, GT and GPT-4o) and the MWE form (two levels: continuous and discontinuous) (Mellinger and Hanson 2016). Second, two interesting examples were analysed through a qualitative perspective to illustrate the difficulties that the different systems were facing regarding MWEs, especially in their discontinuous forms (see Tables 1, 2, and 3 in Section 4).

4. Results and discussion

This section presents the results of the analysed MT systems when translating Spanish-to-English MWEs, focusing on both continuous (easier to detect and translate) and discontinuous (more difficult to detect and translate) forms. Consequently, the analysis is conducted independently for each of these MWE forms. The results are based on human evaluation, with scores ranging from 0 to 50, where higher scores indicate better translation accuracy. The findings are summarised in two box plot charts that showcase the average score that the MT systems achieved throughout the six MWEs analysed.

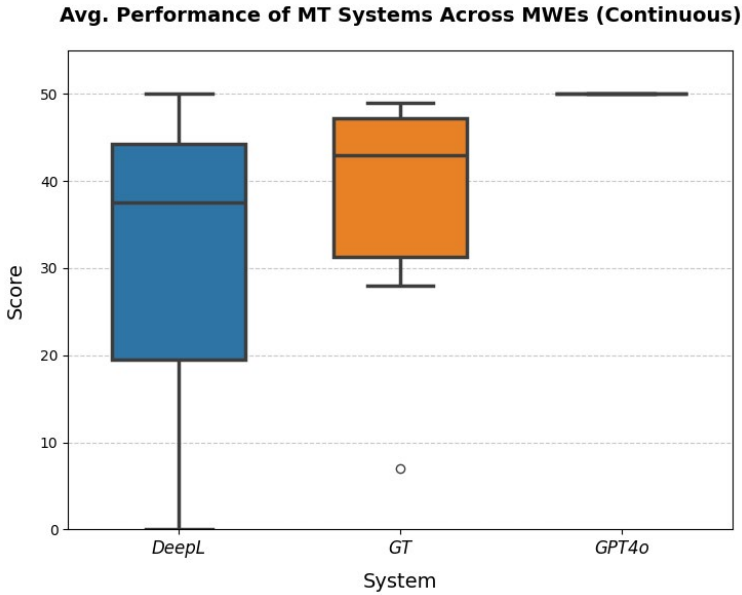

Figure 1: Average performance of MT systems across MWEs (continuous)

The box plot for continuous MWEs (see Figure 1) illustrates the performance of the three MT systems across the selected MWEs. GPT-4o consistently achieved the highest scores, with a perfect score of 50 for the six MWEs under study. This indicates that GPT-4o excels in properly detecting and translating the idiomatic meaning of these continuous MWEs. DeepL and GT showed more variability in their performance. DeepL scored moderately well for MWEs like "llevarse el gato al agua" (45 out of 50) and "pagar el pato" (50 out of 50) but struggled with "ganar/costar/pagar… cuatro perras," scoring 0. GT performed better than DeepL for most MWEs, with scores ranging from 28 ("ser cuatro gatos") to 49 ("pagar el pato"). However, GT also struggled with "ganar/costar/pagar… cuatro perras," only obtaining a score of 7 out of 50.

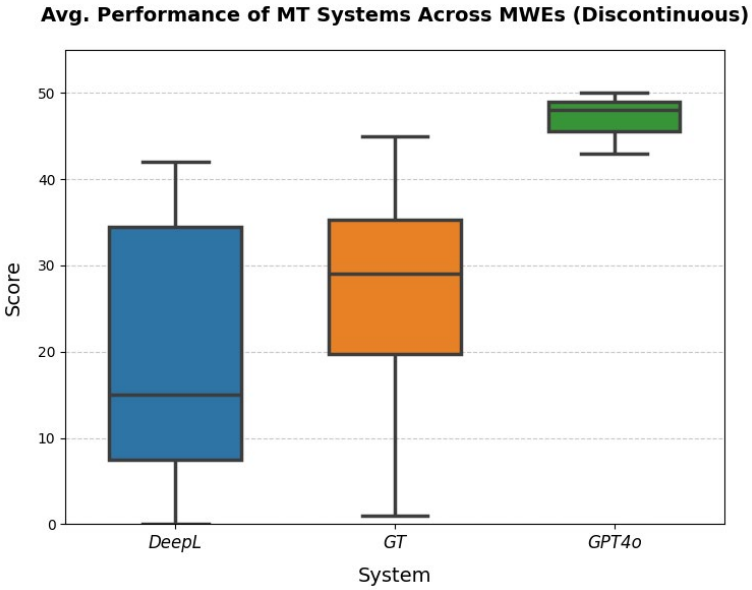

Figure 2: Average performance of MT systems across MWEs (discontinuous)

The box plot for discontinuous MWEs (see Figure 2) reveals a similar trend, with GPT-4o outperforming both DeepL and GT. GPT-4o achieved near-perfect scores for all MWEs, with the lowest score being 43 for "pagar el pato." This suggests that GPT-4o is highly effective in handling the additional complexity introduced by discontinuous forms, maintaining the idiomatic meaning even when the MWE constituents are separated by intervening elements. This supports previous findings on MT conducted by LLMs, which discussed their strengths in respecting contextual coherence and understanding, if compared with NMT (Castilho *et al.* 2023; Karpinska & Iyyer 2023; Wang *et al.* 2023; Briva-Iglesias, Dogru & Camargo 2024, Briva-Iglesias 2025). DeepL and GT showed significant drops in performance for discontinuous MWEs compared to their performance on continuous forms. DeepL's scores ranged from 0 (*ganar/costar/pagar… cuatro perras*) to 42 (*llevarse el gato al agua*), while GT's scores ranged from

1 (*ganar/costar/pagar… cuatro perras*) to 45 (*pagar el pato*). Both systems struggled particularly with *ganar/costar/pagar… cuatro perras*, indicating that this MWE, appearing in various forms, poses a significant challenge for traditional NMT systems, especially in its discontinuous form.

The ANOVA revealed a statistically significant main effect of MT systems on translation accuracy ($F(2,36) = 45.72$, $p < .001$). This indicates that the performance of the three MT systems differed statistically significantly. Post-hoc comparisons using Tukey's HSD test showed that GPT-4o ($M = 48.67$, $SD = 2.05$) outperformed both GT ($M = 34.33$, $SD = 14.12$) and DeepL ($M = 24.17$, $SD = 18.12$), with all pairwise comparisons reaching statistical significance ($p < .001$). GT also performed statistically significantly better than DeepL ($p = .012$).

There was also a statistically significant main effect of MWE form on translation accuracy ($F(1,36) = 32.18$, $p < .001$). This suggests that the form of the MWE (continuous vs. discontinuous) significantly influenced translation performance. On average, continuous MWEs ($M = 41.25$, $SD = 16.34$) were translated more accurately than discontinuous MWEs ($M = 29.58$, $SD = 18.72$).

The ANOVA also revealed a statistically significant interaction effect between MT system and MWE form ($F(2,36) = 8.94$, $p < .001$). This indicates that the impact of MWE form on translation accuracy also varied depending on the MT system. Specifically, GPT-4o showed minimal variation in performance between continuous ($M = 49.83$, $SD = 0.41$) and discontinuous ($M = 47.50$, $SD = 2.74$) forms, suggesting its robustness in handling both types of MWEs. GT and DeepL, however, exhibited a more pronounced decline in performance for discontinuous MWEs compared to continuous forms. For GT, continuous MWEs obtained, on average, better results ($M = 38.50$, $SD = 14.12$) than discontinuous MWEs ($M = 30.17$, $SD = 15.34$). Similarly, DeepL's performance was better at translating continuous MWEs ($M = 35.33$, $SD = 18.12$) than at translating discontinuous MWEs ($M = 13.00$, $SD = 15.67$).

In order to illustrate the different performance of the MT systems, let us observe the following examples in Tables 1, 2, and 3 with the source-text MWEs *ser cuatro gatos*, *llevarse el gato al agua*, and *costar cuatro perras*,

respectively, translated by DeepL, GT and GPT-4o (the whole sequence with the MWEs and their translation is underlined, for illustration purposes).

Source Text	En realidad, el éxito de la movilización de anoche -- acabemos con el Borbón antes de irnos a ver el fútbol-- fue equiparable al de la huelga general de la semana pasada, o sea, que fue un fiasco. Eran también cuatro gatos, pero muy molestos, los que estuvieron en Colliure martirizando al personal con sus consignas de primero de primaria.
DeepL	In reality, the success of last night's mobilisation - let's get rid of the Bourbon before we go and watch the football - was comparable to that of last week's general strike, in other words, it was a fiasco. There were also four cats, but very annoying ones, who were in Collioure, martyring the staff with their first-grade slogans.
Google Translate	In reality, the success of last night's mobilization -- let's finish off the Bourbon before we go watch football -- was comparable to that of the general strike last week, that is, it was a fiasco. There were also a few cats, but very annoying, those who were in Colliure tormenting the staff with their first-grade slogans.
GPT-4o	The success of last night's mobilization—*"let's get rid of the Bourbon before heading out to watch the football game"*—was comparable to that of last week's general strike, meaning it was a failure. It was, again, just a handful of very annoying people in Colliure, tormenting the public with their elementary-school slogans.

Table 1: example with a discontinuous case of the MWE *ser cuatro gatos*.

In Table 1, we can observe distinctly different results among the MT systems regarding the translation of a discontinuous instance of the MWE *ser cuatro gatos*: *...Eran también cuatro gatos…* In DeepL, a fully literal translation was provided for that sequence: "…There were also four cats…" Nevertheless, the context (the "success" of two demonstrations being compared) and cooccurrences such as *ver un partido de fútbol* [watch a football game], *huelga* [strike], *fiasco* and *consignas* [chants] allow us to clearly understand that the authors are not referring to any feline whatsoever, but they are reporting that just a bunch of people attended those strikes, which were hence unsuccessful according to the authors. GT's proposal "…There were also a few cats…" seems to partially capture the idiomatic meaning of

"just a bunch of" evoked by "ser cuatro gatos"; however, it does not properly interpret the idiomatic meaning of "gatos" (figuratively, "people" here), providing a literal (and hence inadequate) translation for it. Only GPT-4o appears to properly identify the fully idiomatic meaning of that sequence and offer a translation accordingly: "…It was, again, just a handful of very annoying people…"

Source Text	Las chicas de la Usal se llevan la victoria en el derbi frente al Santa Marta (47-54) El equipo de baloncesto femenino de la Universidad de Salamanca se llevaba este sábado el gato al agua al superar al Santa Marta en el derbi disputado en tierras tormesinas y donde las universitarias se imponían por 47-54. De esta forma, el equipo de la Usal suma su tercer triunfo en cuatro partidos. A pesar de ello, el Santa Marta se mantiene segundo en la tabla con cuatro victorias y tan sólo dos derrotas en los seis encuentros que lleva disputados.
DeepL	The girls of the Usal take the victory in the derby against Santa Marta (47-54) The women's basketball team of the University of Salamanca took the cat to the water on Saturday by beating Santa Marta in the derby played in lands tormesinas and where the university women won by 47-54. In this way, the Usal team adds its third win in four games. Despite this, Santa Marta remains second in the table with four wins and only two defeats in the six games played.
Google Translate	The Usal girls take the victory in the derby against Santa Marta (47-54) The women's basketball team of the University of Salamanca took the cat out of the bag this Saturday by beating Santa Marta in the derby played in Tormes, where the university girls won by 47-54. In this way, the Usal team has its third victory in four games. Despite this, Santa Marta remains second in the table with four victories and only two defeats in the six games it has played.
GPT-4o	The women's basketball team from the University of Salamanca claimed victory in the derby against Santa Marta (47-54). The Usal team came out on top this Saturday, winning the derby played in Santa Marta's home territory. This marks the team's third victory in four games. Despite the loss, Santa Marta remains second in the standings, with four wins and just two losses in six games.

Table 2: example with a discontinuous case of the idiom *llevarse el gato al agua*.

In Table 2, we can once again observe markedly different outputs offered by the systems under study for the translation of a discontinuous case of the

MWE *llevarse el gato al agua*: …*se llevaba este sábado el gato al agua*… DeepL provides a literal (and hence inadequate) translation for the source-text idiom: "…took the cat to the water this Saturday…" In Google Translate's proposal "…took the cat out of the bag this Saturday…," it is possible to identify the idiom "take the cat out of the bag," which is an infrequent variant (only 7 concordances in enTenTen21, i.e., <0.01 per million tokens) of the idiom "let the cat out of the bag" (5128 concordances in enTenTen21, i.e., 0.08 per million tokens) with the meaning of "to give away a secret" (Merriam Webster – Several authors 2025a). This idiom hence conveys a completely unrelated meaning to the one of the source-text idiom, i.e., "to be victorious in a fight, competition, argument, etc." Finally, in GPT-4o's outcome "…came out on top this Saturday…," it is possible to detect the idiom "to come out on top" ("to win a competition," Merriam Webster – Several authors 2025b), having an analogous meaning to the source-text MWE *llevarse el gato al agua*.

Source Text	La principal diferencia es que tenías que tener bastante más pasta para hacerlo y tenías que pagar visados y presentar el puto pasaporte hasta para cruzar los Pirineos. Hoy los chavales se van de erasmus un año y les cuesta todo cuatro perras con matrículas gratis total, y si España no te gusta te vas fuera sin dar explicaciones a nadie y te buscas un trabajo mejor o te montas tu propia empresa o te haces autónomo con todos los derechos del país donde vayas.
DeepL	The main difference is that you had to have a lot more money to do it and you had to pay visas and present your fucking passport even to cross the Pyrenees. Today kids go on erasmus for a year and it costs them four quid with free tuition, and if you don't like Spain you go abroad without giving explanations to anyone and look for a better job or start your own company or become self-employed with all the rights of the country where you go.
Google Translate	The main difference is that you had to have a lot more money to do it and you had to pay for visas and present your fucking passport even to cross the Pyrenees. Today kids go on Erasmus for a year and everything costs them four cents with free tuition, and if you don't like Spain you go abroad without giving anyone any explanation and you look for a better job or you set up your own company or you become self-employed with all the rights of the country where you go.

GPT-4o	The main difference is that you used to need much more money to do it, and you had to pay for visas and present your damn passport even to cross the Pyrenees. Today, kids go on Erasmus for a year, and it costs them just a few pennies with completely free tuition. If you don't like Spain, you can leave without explaining to anyone, find a better job, start your own business, or become self-employed with all the rights of the country you go to.

Table 3: example with a discontinuous case of the idiom *costar cuatro perras*.

In Table 3, it can be observed that all three systems fail to deliver an appropriate translation for ...*cuesta todo cuatro perras*..., either because they are unable to provide an adequate equivalent for the Spanish idiom *costar cuatro perras* (in the case of DeepL and Google Translate) or because they omit the intervening element within the discontinuous idiom, leading to a significant change in the original meaning (in the case of DeepL and GPT-4o). The outputs produced by DeepL and Google Translate ("four quid" and "four cents") preserve the numeral "four," treating it as a compositional element rather than as part of a lexicalised idiom. In English, these expressions are interpreted literally as specific monetary amounts and do not function idiomatically with the meaning of "very cheap." This literal transfer introduces false precision and yields a pragmatically anomalous reading when applied to a complex experience such as a whole year abroad, which would not normally cost only "four quid" or "four cents." In the case of DeepL and GPT-4o, the omission of *todo* [everything] in the target text introduces a different meaning from that of the source text, since *todo* emphasises that the entire package (the year abroad, fees, logistics, etc.) costs almost nothing. This use of *todo* strengthens the evaluative force of the idiom and contributes to the speaker's slightly indignant, contrastive tone. Therefore, while GPT-4o's proposal ("it costs them just a few pennies") is closer to the figurative meaning of *costar cuatro perras*, it fails to include the intervening element *todo* (the translation should hence be "everything costs them just a few pennies"), resulting in a significant loss in the target text.

5. Conclusion

This study set out to evaluate the performance of GT, DeepL, and GPT-4o when translating Spanish-to-English MWEs, with a particular focus on both continuous and discontinuous forms. The findings reveal statistically significant differences in the ability of these systems to handle the complexities of MWEs, particularly those involving idiomaticity and syntactic flexibility.

GPT-4o consistently outperformed both GT and DeepL in translating the selected MWEs, achieving near-perfect scores for both continuous and discontinuous forms. These results demonstrate the robustness of LLMs in handling the non-compositional and syntactically flexible nature of these MWEs, supporting previous work that highlights the potential of LLMs for contextual-aware translations (Castilho *et al.* 2023; Karpinska & Iyyer 2023; Wang *et al.* 2023; Briva-Iglesias, Dogru & Camargo 2024, Briva-Iglesias 2025). While GT and DeepL showed moderate success in translating continuous MWEs, their performance declined significantly for discontinuous forms. This is particularly evident in the case of idioms such as *ganar/costar/pagar… cuatro perras*, for which both systems struggled to produce accurate translations. The results suggest that traditional NMT systems, despite their advancements, still face limitations in handling the syntactic and semantic complexities of MWEs.

In this context, the present study makes several important contributions to the field of MT. By systematically comparing the performance of traditional NMT systems and general-purpose LLMs in translating MWEs, this research provides new results into the capabilities and limitations of LLMs. The findings highlight LLMs' potential to overcome the challenges posed by idiomatic expressions and contextual understanding, contributing to the ongoing discussion on the role of LLMs in MT. In this regard, the dataset is shared open source to invite the broader research community to study the topic further. The use of human evaluation addresses the limitations of automatic metrics when assessing MWE translation quality (Baziotis, Mathur & Hasler 2023), which often fail to capture the complexities of idiomatic and non-compositional expressions. The study also demonstrates the value of combining corpus-based analysis with statistical

methods, such as ANOVA and Tukey's HSD tests, to provide a comprehensive evaluation of MT systems (Mellinger & Hanson 2016).

These findings can also have practical implications for professional translators and MT users. For instance, this study can open new avenues for future research evaluating whether the superior performance of GPT-4o regarding MWEs could also potentially reduce the need for extensive post-editing in translation workflows, particularly for texts rich in idiomatic expressions. Yet, while this study provides valuable results, it has some limitations that suggest other additional directions for future research. First, the study focused exclusively on Spanish-to-English translation. Future research could expand the analysis to other language pairs, particularly those involving low-resource languages, to assess the generalisability and/or transferability of the findings. Secondly, the MWEs analysed in this study were primarily general-purpose idioms. Future work could explore the translation of domain-specific MWEs, such as those found in legal texts, to evaluate the performance of MT systems in specialised contexts. Another limitation of this study is the use of GPT-4o, a proprietary LLM, as the representative of LLMs. At the time of evaluation, no open-source LLM was ranked high enough to be considered a state-of-the-art system. While GPT-4o's performance provides valuable information about the capabilities of LLMs, the reliance on a proprietary model limits the reproducibility and transparency of the research. For future work, open-source LLMs will be introduced to foster openness in research and enable broader access to the tools and methodologies used in this study. Finally, the dataset is limited to unambiguously idiomatic uses, excluding literal or ambiguous cases in order to isolate the systems' ability to produce idiomatic translations when such interpretations are pragmatically licensed by context. While semantic disambiguation constitutes an important challenge in machine translation, addressing it would require a distinct experimental design and is therefore left for future research.

In conclusion, this study demonstrates that general-purpose LLMs like GPT-4o represent a significant advancement in MT technology, particularly in their ability to handle the complexities of the analysed MWEs. While traditional NMT systems like GT and DeepL continue to face challenges in translating discontinuous MWEs, GPT-4o's consistent performance

suggests that LLMs can play a key role in improving the quality and contextual awareness of MT systems. These findings contribute to the growing body of research on the capabilities of general-purpose LLMs in MT and suggest some directions for future work in this area.

Acknowledgments

This research was carried out within the framework of several research projects (ref. PID2020-112818GB-I00, HUM106-G-FEDER, PPRO-IUI-2023-06, PID2024-160929OB-I00) at the University of Malaga and at the Research Institute of Multilingual Language Technologies (IUITLM) (Spain).

References

ARTSTEIN, Ron. (2017) "Inter-Annotator Agreement." In: Ide, Nancy & James Pustejovsky (eds.) 2017. *Handbook of Linguistic Annotation*. Dordrecht: Springer Netherlands, pp. 297–313. DOI: 10.1007/978-94-024-0881-2_11.

BAZIOTIS, Christos; Prashant Mathur & Eva Hasler. (2023) "Automatic Evaluation and Analysis of Idioms in Neural Machine Translation." In: Vlachos, Andreas & Isabelle Augenstein (eds.) 2023. *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 3682–3700. DOI: 10.18653/v1/2023.eacl-main.267.

BELINKOV, Yonatan & Yonatan Bisk. (2018) "Synthetic and Natural Noise Both Break Neural Machine Translation." arXiv. DOI: 10.48550/arXiv.1711.02173.

BRIVA-IGLESIAS, Vicent. (2023) "Translation Technologies Advancements: From Inception to the Automation Age." In: Bellés-Calvera, Lucía & María Pallarés-Renau (eds.) (2023) *La Família Humana: Perspectives Multidisciplinàries de La Investigació En Ciències Humanes i Socials*, pp. 137–52. Emergents 3. Castellón de la Plana: Publicacions de la Universitat Jaume I. Servei de Comunicació i Publicacions.

BRIVA-IGLESIAS, Vicent. (2025) "Are AI agents the new machine translation frontier? Challenges and opportunities of single- and multi-agent systems for multilingual digital communication." In: Pierrette Bouillon et al. (eds.). 2025. *Proceedings of Machine Translation Summit XX: Volume 1*. MTSummit 2025, Geneva, Switzerland: European Association for Machine

Translation, pp. 365–377. Electronic version: <https://aclanthology.org/2025.mtsummit-1.28/>.

Briva-Iglesias, Vicent; Gokhan Dogru & João Lucas Cavalheiro Camargo. (2024) "Large Language Models 'Ad Referendum': How Good Are They at Machine Translation in the Legal Domain?" *MonTI. Monografías de Traducción e Interpretación* 16, pp. 75–107. DOI: 10.6035/MonTI.2024.16.02.

Briva-Iglesias, Vicent & Isabel Peñuelas Gil. (2025) "Simplifying Healthcare Communication: Evaluating AI-Driven Plain Language Editing of Informed Consent Forms." Electronic version: <https://www.researchgate.net/profile/Vicent-Briva-Iglesias/publication/390916537_Simplifying_healthcare_communication_Evaluating_AI-driven_plain_language_editing_of_informed_consent_forms/links/68022f4fded43315572abc9a/Simplifying-healthcare-communication-Evaluating-AI-driven-plain-language-editing-of-informed-consent-forms.pdf>.

Brown, Tom B.; Benjamin Mann; Nick Ryder; Melanie Subbiah; Jared Kaplan; Prafulla Dhariwal; Arvind Neelakantan; Pranav Shyam; Girish Sastry; Amanda Askell; Sandhini Agarwal; Ariel Herbert-Voss; Gretchen Krueger; Tom Henighan; Rewon Child; Aditya Ramesh; Daniel M. Ziegler; Jeffrey Wu; Clemens Winter; Christopher Hesse; Mark Chen; Eric Sigler; Mateusz Litwin; Scott Gray; Benjamin Chess; Jack Clark; Christopher Berner; Sam McCandlish; Alec Radford; Ilya Sutskever & Dario Amodei. (2020) "Language Models Are Few-Shot Learners." arXiv. DOI: 10.48550/arXiv.2005.14165.

Castilho, Sheila; Clodagh Mallon; Rahel Meister & Shengya Yue. (2023) "Do Online Machine Translation Systems Care for Context? What about a GPT Model?" *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*. Tampere, Finland: European Association for Machine Translation, pp. 393–417. Electronic version: <https://events.tuni.fi/eamt23/>.

Chiang, Wei-Lin; Lianmin Zheng; Ying Sheng; Anastasios Nikolas Angelopoulos; Tianle Li; Dacheng Li; Hao Zhang; Banghua Zhu; Michael Jordan; Joseph E. Gonzalez & Ion Stoica. (2024) "Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference." arXiv. DOI: 10.48550/arXiv.2403.04132.

Constant, Mathieu; Gülşen Eryiğit; Johanna Monti; Lonneke van der Plas; Carlos Ramisch; Michael Rosner & Amalia Todirascu. (2017) "Survey:

Multiword Expression Processing: A Survey." *Computational Linguistics* 43:4, pp. 837–92. DOI: 10.1162/COLI_a_00302.

Corpas Pastor, Gloria. (2013) "Detección, descripción y contraste de las unidades fraseológicas mediante tecnologías lingüísticas." In: Olza Moreno, Inés & Elvira Manero Richard (eds.) 2013. *Fraseopragmática, 2013, ISBN 978-3-86596-448-9, pp. 335-374*, 335–74. Frank & Timme. Electronic version: <https://dialnet.unirioja.es/servlet/articulo?codigo=5546029>.

Fadaee, Marzieh; Arianna Bisazza & Christof Monz. (2018) "Examining the Tip of the Iceberg: A Data Set for Idiom Translation." In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC 2018). Miyazaki, Japan: European Language Resources Association (ELRA).

Hendy, Amr; Mohamed Abdelrehim; Amr Sharaf; Vikas Raunak; Mohamed Gabr; Hitokazu Matsushita; Young Jin Kim; Mohamed Afify & Hany Hassan Awadalla. (2023) "How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation." arXiv. DOI: 10.48550/arXiv.2302.09210.

Hidalgo-Ternero, Carlos Manuel. (2020) "Google Translate vs. DeepL: analysing neural machine translation performance under the challenge of phraseological variation." In: Mogorrón Huerta, Pedro (ed.) 2020. *Análisis multidisciplinar del fenómeno de la variación en traducción e interpretación / Multidisciplinary Analysis of the Phenomenon of Phraseological Variation in Translation and Interpreting. MonTI Special Issue* 6, pp. 154-177. DOI: 10.6035/MonTI.2020.ne6.5.

Hidalgo-Ternero, Carlos Manuel & Francisco Javier Lima-Florido. (2023) "How Can Paidiom Improve the Neural Machine Translation of Multiword Expressions?" *Translating and the Computer Conference (TC45)*. City of Luxembourg, Luxembourg: AsLing, The International Association for Advancement in Language Technology.

Hidalgo-Ternero, Carlos Manuel & Gloria Corpas Pastor. (2020) "Bridging the '*gApp*': Improving Neural Machine Translation Systems for Multiword Expression Detection." *Yearbook of Phraseology* 11:1, pp. 61–80. DOI: 10.1515/phras-2020-0005.

Hidalgo-Ternero, Carlos Manuel & Xiaoqing Zhou-Lian. (2022) "Reassessing gApp: Does MWE Discontinuity Always Pose a Challenge to Neural Machine Translation?" In: Corpas Pastor, Gloria & Ruslan Mitkov (eds.)

2022. *Computational and Corpus-Based Phraseology*. Cham: Springer International Publishing, pp. 116–32. DOI: 10.1007/978-3-031-15925-1_9.

JIAO, Wenxiang; Wenxuan Wang; Jen-tse Huang; Xing Wang & Zhaopeng Tu. (2023) "Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine." arXiv. DOI: 10.48550/arXiv.2301.08745.

KAMALUDDIN, Mohamad Ihsan; Moch Wildan Khoerul Rasyid; Fourus Huznatul Abqoriyyah & Andang Saehu. (2024) "Accuracy Analysis of DeepL: Breakthroughs in Machine Translation Technology." *Journal of English Education Forum (JEEF)* 4:2, pp. 122–26. DOI: 10.29303/jeef.v4i2.681.

KARPINSKA, Marzena & Mohit Iyyer. (2023) "Large Language Models Effectively Leverage Document-Level Context for Literary Translation, but Critical Errors Persist." arXiv. DOI: 10.48550/arXiv.2304.03245.

KEKIĆ, Katarina (2008) "El lenguaje figurado con zoonimos en serbio." *Language Design 10*, pp. 107–131.

KENNY, Dorothy. (2022) "Human and Machine Translation." In: Kenny, Dorothy (ed.) 2022. *Machine Translation for Everyone: Empowering Users in the Age of Artificial Intelligence*. Berlin: Language Science Press, pp. 23-49. DOI: 10.5281/zenodo.6759976.

KILGARRIFF, Adam; Vít Baisa; Jan Bušta; Miloš Jakubíček; Vojtěch Kovář; Jan Michelfeit; Pavel Rychlý & Vít Suchomel. (2014) "The Sketch Engine: Ten Years On." *Lexicography* 1:1, pp. 7–36. DOI: 10.1007/s40607-014-0009-9.

KOCMI, Tom; Christian Federmann; Roman Grundkiewicz; Marcin Junczys-Dowmunt; Hitokazu Matsushita & Arul Menezes. (2021) "To Ship or Not to Ship: An Extensive Evaluation of Automatic Metrics for Machine Translation." *Proceedings of the Sixth Conference on Machine Translation*. Association for Computational Linguistics, pp. 478-94. Electronic version: <https://aclanthology.org/2021.wmt-1.57>.

LOHAR, Pintu; Maja Popović; Haithem Alfi & Andy Way. (2019) "A Systematic Comparison between SMT and NMT on Translating User-Generated Content." *Proceedings of CICLing 2019, the 20th International Conference on Computational Linguistics and Intelligent Text Processing. Lecture Notes in Computer Science (LNCS)*. La Rochelle, France: Springer. Electronic version: <https://doras.dcu.ie/23869/>.

LUQUE NADAL, Lucía (2012) *Principios de culturología y fraseología españolas. Creatividad y variación en las unidades fraseológicas*. Berlin: Peter Lang. ISBN 978-3-631-60864-7.

MELLINGER, Christopher & Thomas Hanson. (2016) *Quantitative Research Methods in Translation and Interpreting Studies*. London: Routledge. DOI: 10.4324/9781315647845>.

MOSLEM, Yasmin; Rejwanul Haque; John D. Kelleher & Andy Way. (2023) "Adaptive Machine Translation with Large Language Models." arXiv. DOI: 10.48550/arXiv.2301.13294.

PAPINENI, Kishore; Salim Roukos; Todd Ward & Wei-Jing Zhu. (2001) "BLEU: A Method for Automatic Evaluation of Machine Translation." *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, 311. Philadelphia, Pennsylvania: Association for Computational Linguistics. DOI: 10.3115/1073083.1073135.

PÉREZ-ORTIZ, Juan Antonio; Mikel L. Forcada & Felipe Sánchez-Martínez. (2022) "How Neural Machine Translation Works." In: Kenny, Dorothy (ed.) 2022. *Machine Translation for Everyone: Empowering Users in the Age of Artificial Intelligence*. Translation and Multilingual Natural Language Processing 18. Berlin: Language Science Press, pp. 141-164.

RAMISCH, Carlos & Aline Villavicencio. (2022) "Computational Treatment of Multiword Expressions." In: Mitkov, Ruslan (ed.) 2022. *The Oxford Handbook of Computational Linguistics*. Oxford University Press. DOI: 10.1093/oxfordhb/9780199573691.013.56.

RIVERA-TRIGUEROS, Irene. (2022) "Machine Translation Systems and Quality Assessment: A Systematic Review." *Language Resources and Evaluation* 56:2, pp. 593–619. DOI: 10.1007/s10579-021-09537-5.

ROHANIAN, Omid; Shiva Taslimipoor; Samaneh Kouchaki; Le An Ha & Ruslan Mitkov. (2019) "Bridging the Gap: Attending to Discontinuity in Identification of Multiword Expressions." arXiv. DOI: 10.48550/arXiv.1902.10667.

ROSSI, Caroline & Alice Carré. (2022) "How to Choose a Suitable NMT Solution?: Evaluation of MT Quality." In: Kenny, Dorothy (ed.) 2022. *Machine translation for everyone: Empowering users in the age of artificial intelligence*. Berlin: Language Science Press, pp. 51–79. DOI: 10.5281/zenodo.6759978.

SEVERAL AUTHORS (Merrian Webster). (2025a) "Definition of LET THE CAT OUT OF THE BAG." Electronic version: <https://www.merriam-webster.com/dictionary/let+the+cat+out+of+the+bag>.

SEVERAL AUTHORS (Merrian Webster). (2025b) "Definition of COME OUT ON TOP." Electronic version: <https://www.merriam-webster.com/dictionary/come+out+on+top>.

SHAO, Yutong; Rico Sennrich; Bonnie Webber & Federico Fancellu. (2018) "Evaluating Machine Translation Performance on Chinese Idioms with a Blacklist Method." *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC 2018). Miyazaki, Japan: European Language Resources Association (ELRA).

ŠKVAROVA, Pavla & Petr Šlechta. (2015) "Estereotipos masculinos y femeninos en los zoologismos fraseológicos españoles y checos." *Eslavistica complutense 15*, pp. 65–88.

WANG, Longyue; Chenyang Lyu; Tianbo Ji; Zhirui Zhang; Dian Yu; Shuming Shi & Zhaopeng Tu. (2023) "Document-Level Machine Translation with Large Language Models." arXiv. DOI: 10.48550/arXiv.2304.02210.

ZANINELLO, Andrea & Alexandra Birch. (2020) "Multiword Expression Aware Neural Machine Translation." In: Calzolari, Nicoletta; Frédéric Béchet; Philippe Blache; Khalid Choukri; Christopher Cieri; Thierry Declerck; Sara Goggi; Hitoshi Isahara; Bente Maegaard; Joseph Mariani; Hélène Mazo; Asuncion Moreno; Jan Odijk & Stelios Piperidis (eds.) 2020. *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 3816–25. Electronic version: <https://aclanthology.org/2020.lrec-1.471/>.

ZHOU-LIAN, Xiaoqing; Gloria Corpas Pastor & Carlos Manuel Hidalgo-Ternero. (2024) "En Torno a La Traducción Automática Neuronal de Zoologismos." In: Sánchez Carnicer, Jaime & Lorena Arce Romeral (eds.) 2024. *Nuevos avances tecnológicos en la teoría y práctica de la traducción e interpretación*. Berlin: Peter Lang, pp. 111-138. Electronic version: <https://burjcdigital.urjc.es/items/c68196cb-1699-4cdd-bedb-2d215f344495>.

BIONOTES / NOTAS BIOGRÁFICAS

CARLOS MANUEL HIDALGO-TERNERO is an Assistant Professor in the Department of Translation and Interpreting at the University of Malaga (Spain) and a member of the LEXYTRAD research group (IUITLM). He has received the "Adam Kilgarriff" Award for the best PhD thesis in multilingual language technologies as well as the prize for best PhD thesis in the Linguistics, Literature and Translation programme at the University of Malaga. He specialises in machine translation, corpus linguistics, and computational phraseology.

VICENT BRIVA-IGLESIAS is an Assistant Professor in Translation Technology at Dublin City University (DCU). He specialises in human-centered machine translation and human-computer interaction. In addition to his role at DCU, Vicent is an Adjunct Professor at McGill University (Canada) and the Universitat Oberta de Catalunya (Spain), and frequently collaborates with the Barcelona Supercomputing Center as an external researcher of AI for healthcare.

CARLOS MANUEL HIDALGO-TERNERO es Profesor Ayudante Doctor en el Departamento de Traducción e Interpretación de la Universidad de Málaga (España) y miembro del grupo de investigación LEXYTRAD (IUITLM). Ha recibido el premio "Adam Kilgarriff" a la mejor tesis doctoral en tecnologías lingüísticas multilingües, así como el premio a la mejor tesis doctoral del Programa de Doctorado en Lingüística, Literatura y Traducción de la Universidad de Málaga. Sus líneas de investigación se centran en la traducción automática, la lingüística de corpus y la fraseología computacional.

VICENT BRIVA-IGLESIAS es Profesor Ayudante Doctor en Tecnología de la Traducción en la Dublin City University (DCU). Está especializado en traducción automática con un enfoque centrado en el ser humano, así como en la interacción humano-computadora. Además de su labor en la DCU, Vicent es profesor adjunto en la McGill University (Canadá) y en la Universitat Oberta de Catalunya (España), y colabora con frecuencia con el Barcelona Supercomputing Center como investigador externo en inteligencia artificial aplicada a la salud.