Closing the "gApp" in Google Translate and ChatGPT's machine translation of discontinuous multiword expressions (CA→EN/ES)

Carlos Manuel Hidalgo-Ternero
University of Malaga, IUITLM (Spain)
cmhidalgo@uma.es

Relevant UniDive WGs: WG1, WG2, WG3, WG4

1 Introduction

Although neural machine translation (NMT) has established itself as a dominant approach in machine translation (MT) over the last decade, certain weaknesses remain, particularly with multiword expressions (MWEs). MWEs such as the Catalan *fer la guitza* (literally, 'to kick someone'; figuratively, 'to give someone a hard time') are especially problematic because they can appear either in continuous form, where all components occur together (e.g., "Li **feien la guitza** de nou." meaning 'They were giving him/her a hard time again.') or in discontinuous form, where the elements are separated by intervening words (e.g., "Li **feien** de nou **la guitza**."). Despite the considerable advances in NMT, discontinuous MWEs continue to represent a major challenge (Constant et al., 2017; Rohanian et al., 2019; Ramisch and Villavicencio, 2018).

The emergence of large language models (LLMs) has brought about a new direction in MT research (Hendy et al., 2023). In contrast to conventional NMT architectures designed specifically for translation tasks, LLMs are general-purpose systems capable of performing translation through context-aware prompting (Wang et al., 2023), alongside many other NLP applications. Given their potential to handle linguistically complex phenomena such as MWEs, a systematic comparison between LLMs and state-of-the-art NMT systems is essential. Nevertheless, empirical evidence is still relatively scarce due to the recent development of these models. In this regard, a recent study conducted by Hidalgo-Ternero and Briva-Iglesias (2026) indicates that LLMs surpass NMT systems in translating both continuous and discontinuous MWEs for high-resource language pairs (ES→EN). Their effectiveness in medium- and low-resource language scenarios, however, remains largely underexplored.

In this context, to address the difficulties that discontinuous MWEs still represent for even the most robust MT systems, we developed gApp[1] (Hidalgo-Ternero and Corpas Pastor, 2020), a text preprocessing tool designed to automatically identify discontinuous MWEs and convert them into their continuous variants to improve downstream MT performance. To demonstrate gApp's effectiveness, eight different experiments have previously been conducted. They comprise a total of 5,610 cases (1,870 discontinuous cases + 1,870 continuous ones after gApp + 1,870 continuous ones after the manual conversion), including different MWE typologies (verb-noun constructions and verb + prepositional phrase constructions), NMT systems (DeepL, Google Translate, and ModernMT), and translation directions (ES/FR/IT→EN/DE/FR/ES/IT/PT/ZH). The global average results of Experiments 1–8 are presented in Table 1, which shows NMT accuracy before gApp (column 1), NMT accuracy after gApp (column 2), the improvement obtained with gApp (column 3), and the results of the gold standard, i.e., the manual conversion (column 4). Columns 3 and 4 are expressed in percentage points (pp).

Before gApp	After gApp	Improvement with gApp	Gold Standard
49.2%	62.6%	+13.4 pp	+14 pp

Table 1: Global results from Experiments 1-8

In this context, we have now adapted gApp's detection and conversion mechanisms to another syntactically-flexible Romance language (Catalan), in order to test our main hypothesis: that gApp can also enhance MT performance from a medium-resource language (Catalan) into two high-resource languages (English and Spanish).

2 Overview of gApp

gApp is a Python-based preprocessing tool built with the spaCy NLP library (Honnibal et al., 2020). To adapt it for Catalan, the pre-trained

[1] gApp is available at the following link: https://lexytrad.es/gapp/app.php

Catalan pipeline `ca_core_news_sm` was used, i.e., a multi-task CNN-based model trained on the WikiNER dataset (Nothman et al., 2013) and the Universal Dependencies Catalan AnCora treebank (Taulé et al., 2008).

gApp performs two main tasks: (i) automatic detection of discontinuous MWEs and (ii) their conversion into continuous forms prior to MT. Discontinuous MWEs are identified using a token-based lexicon lookup method (Ramisch and Villavicencio, 2018). The system matches running text against a curated lexicon of semi-fixed MWEs whose components may be separated by intervening material. Detection relies on rule-based morphosyntactic patterns that model allowable variation, including lemmatised verbal components, fixed lexical elements, and constrained gaps where specific grammatical categories may occur. These patterns were defined through corpus-driven analysis of authentic Catalan data, enabling the system to distinguish true discontinuous MWE instances from superficially similar sequences. For this analysis, we employed two corpora searchable through the corpus query system Sketch Engine (Kilgarriff et al., 2014): caTenTen14 (210 million tokens) and Timestamped JSI web corpus 2014-2021 Catalan (515 million tokens). Once identified, discontinuous MWEs are automatically normalised into their canonical continuous forms, making them more accessible to downstream MT systems.

3 Methodology

For the compilation of our dataset, the corpora caTenTen14 and Timestamped JSI Web Corpus 2014–2021 Catalan were also employed, as they comprise a heterogeneous sample in terms of text sources, types, and language varieties. The dataset includes 2,400 texts with 40 Catalan verbal idioms (VIDs), which will be translated using Google Translate (GT) and ChatGPT (GPT-5.2). For GPT-5.2, a basic zero-shot generic prompt will be employed: "*Please translate the following text into [target language].*" We will compare three scenarios: discontinuous form vs. continuous form after gApp vs. continuous form after manual conversion. The translations will be manually evaluated by three professional translators using a binary adequacy scale (1 = good, 0 = bad), focusing exclusively on the accurate rendering of the MWEs in the target text.

4 Preliminary results

Preliminary results (so far, 600 texts with MWEs have been tested) show that GPT-5.2 significantly outperforms GT in all scenarios, particularly with discontinuous VIDs. The most common error made by these MT systems was producing literal translations where they were inappropriate (e.g., translating "fer la guitza" literally as "to kick someone" instead of its correct figurative meaning, "to give someone a hard time"). These preliminary results also indicate that gApp can, on average, improve GT's performance by 14.7 pp (vs. 15.4 pp with the gold standard) and GPT-5.2's by 4.1 pp (vs. 4.3 pp with the gold standard). The smaller gains for GPT-5.2 may be due to its already high baseline performance with discontinuous MWEs (82%, vs. 43% in the case of GT), leaving little room for improvement after gApp. Global MT performance is, on average, 3.2 pp higher for CA→EN than for CA→ES (70.5% vs. 67.3%).

5 Conclusion

In conclusion, this project will contribute a publicly available, manually annotated dataset for benchmarking MWE translation. It also preliminarily highlights the potential of gApp and LLMs in medium-resource MT contexts. Future research will also evaluate gApp's performance in the case of low-resource dialects of Catalan, such as Valencian and Balearic. In this context, the construction of gApp's lexicon currently requires extensive manual effort and depends on language-specific heuristics, which do not generalise well across different languages or MWE typologies. Consequently, future research will also explore whether LLMs can help to expand gApp's lexicon to other MWEs and languages, or whether it is more efficient to develop the preprocessing prompt within the LLM. These are, therefore, some of the challenges in the relentless race to close the "gApp" in MT quality.

Acknowledgements

This research was carried out within the framework of several research projects (ref. PPRO-IUI-2023-06, PID2024-160929OB-I00/ MICIU/AEI/10.13039/501100011033/, HUM106-G-FEDER, and DGP_PIDI_2024_01159) at the University of Malaga and at the Research Institute of Multilingual Language Technologies (IUITLM) (Spain).

References

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

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation. ArXiv:2302.09210 [cs].

Carlos Manuel Hidalgo-Ternero and Vicent Briva-Iglesias. 2026. ChatGPT vs. DeepL vs. Google Translate: a human evaluation of multiword expressions' machine translation quality. In Diana González Pastor and Celia Rico Pérez, editors, *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*, volume 18 of *MonTI*.

Carlos Manuel Hidalgo-Ternero and Gloria Corpas Pastor. 2020. Bridging the "gApp": improving neural machine translation systems for multiword expression detection. *Yearbook of Phraseology*, 11(1):61–80.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. Software documentation, Explosion AI.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine. *Lexicography*, 1(1):7–36.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. 2013. Learning multilingual named entity recognition from Wikipedia. *Artificial Intelligence*, 194:151–175.

Carlos Ramisch and Aline Villavicencio. 2018. Computational Treatment of Multiword Expressions. In Ruslan Mitkov, editor, *The Oxford Handbook of Computational Linguistics*, page 0. Oxford University Press.

Omid Rohanian, Shiva Taslimipoor, Samaneh Kouchaki, Le An Ha, and Ruslan Mitkov. 2019. Bridging the Gap: Attending to Discontinuity in Identification of Multiword Expressions. ArXiv:1902.10667 [cs].

Mariona Taulé, M. Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-Level Machine Translation with Large Language Models. ArXiv:2304.02210 [cs].