Minding the "gApp" in the neural machine translation of discontinuous idioms (ES>EN/ZH)[1]

Cuidado con el "gApp" en la traducción automática neuronal de locuciones discontinuas (ES>EN/ZH)

Carlos Manuel Hidalgo-Ternero
University of Malaga, IUITLM (Spain)
cmhidalgo@uma.es
Xiaoqing Zhou-Lian
Rey Juan Carlos University (Spain)
xiaoqing.zhou@urjc.es

Abstract: *This study examines the effectiveness of gApp, an innovative text preprocessing system developed to automatically convert discontinuous idioms into their continuous forms in order to enhance current neural machine translation (NMT) systems. Through comprehensive testing with Google Translate and DeepL, we evaluated 500 instances (250 discontinuous and 250 continuous forms) of common Spanish Verb+ Prepositional Phrase (VPP) idioms like* estar hasta los cojones, estar hasta los huevos, estar hasta las narices, estar hasta el gorro, *and* estar hasta la coronilla, *comparing gApp's automatic conversion with the manual conversion (the gold standard). To our knowledge, this is the first study evaluating gApp's effectiveness in translating VPP idioms using NMT in the ES>EN and ES>ZH directions. In this context, the promising outcomes obtained for this idiom category offer valuable insights into potential improvements for idiom-aware NMT systems.*

Keywords: Neural Machine Translation (NMT), Verb + Prepositional Phrase (VPP) idioms, text preprocessing system, discontinuity, Spanish>English/Chinese.

Resumen: *Este estudio examina la eficacia de gApp, un innovador sistema de preprocesamiento de texto desarrollado para convertir automáticamente locuciones discontinuas hacia sus formas continuas con el fin de mejorar los actuales sistemas de traducción automática neuronal (TAN), en concreto, Google Translate y DeepL. Así, evaluamos 500 instancias (250 formas discontinuas y 250 formas continuas) de locuciones frecuentes en español con la estructura Verbo + Sintagma Preposicional (VPP, por sus siglas en inglés) como* estar hasta los cojones, estar hasta los huevos, estar hasta las narices, estar hasta el gorro *y* estar hasta la coronilla, *comparando la conversión automática de gApp con la conversión manual (el patrón oro). Hasta donde sabemos, este es el primer estudio que evalúa la eficacia de gApp en la traducción de locuciones VPP mediante TAN en las direcciones ES>EN y ES>ZH. Los prometedores resultados obtenidos ofrecen información valiosa sobre las posibles mejoras de los sistemas TAN para las locuciones.*

Palabras clave: Traducción Automática Neuronal (TAN), locuciones con estructura Verbo + Sintagma Preposicional (VPP), sistema de preprocesamiento de texto, discontinuidad, español>inglés/chino.

[1] This research was carried out within the framework of several research projects (ref. PID2020-112818GB-I00, ProyExcel_00540, HUM106-G-FEDER, TED2021-129789B-I00, and JA.A1.3-06) at the University of Malaga and at the Research Institute of Multilingual Language Technologies.

1. Introduction

While neural machine translation (NMT) has made remarkable progress in recent years, the accurate translation of idiomatic expressions remains a persistent challenge. Idioms often exhibit syntactic irregularities, non-compositional meanings, and discontinuous structures where intervening elements separate idiom components (e.g., "*keep* all those interesting things in mind"), which can be particularly problematic for their automatic translation and identification (Constant *et al*., 2017; Rohanian *et al*., 2019).

In order to mitigate the persistent challenges that discontinuous idioms pose even to advanced NMT systems (cf. Colson, 2019; Zaninello & Birch, 2020), we have developed gApp[2], a text-preprocessing system designed to automatically identify discontinuous idioms and convert them into their continuous forms with the aim of enhancing NMT performance. To test gApp's effectiveness, 6 previous experiments have been carried out, in which the challenge posed by idiom discontinuity for NMT has been quantified: it meant an average decline in NMT performance by 11.5% when compared to the continuous forms of the idioms. In Figure 1, we can observe to what extent gApp was able to address this specific issue and improve NMT performance for discontinuous idioms throughout those 6 previous experiments (experiment 1 in Hidalgo-Ternero & Corpas Pastor, 2020; 2 in Hidalgo-Ternero & Corpas Pastor, 2024; 3 in Hidalgo-Ternero, 2021; 4 in Hidalgo-Ternero & Corpas Pastor, 2025; 5 in Hidalgo-Ternero & Zhou-Lian, 2022; 6 in Hidalgo-Ternero, 2024).

	Typology of MWEs	**Number of cases**	**NMT system**	**Direc-tion**	**NMT accuracy before gApp**	**NMT accuracy after gApp**	**Improve-ment with gApp**	**Gold Standard**
1	Noun-Verb construc-tions (VNCs)	560	DL	ES>EN	80.7%	90.7%	10%	+3.2%
			GT		60.7%	75.4%	14.6%	+2.1%
2	VNCs	400	DL	ES>EN	49%	62.5%	13.5%	+0.5%
				ES>DE	43.5%	52.5%	9%	+0.5%
3	VNCs	400	DL	FR>EN	40%	58%	18%	=
				FR>ES	41.5%	58%	16.5%	=
4	Verb + Prepositional Phrase construc-tions (VPPs)	300	MMT	ES>EN	50%	60%	10%	=
				ES>DE	23.3%	33.3%	10%	=
				ES>FR	49.3%	60%	10.7%	=
				ES>IT	56.7%	60.7%	4%	=
				ES>PT	56%	58.7%	2.7%	=
			DL	ES>EN	70.7%	81.3%	10.7%	+0.7%
				ES>DE	59.3%	66.7%	7.3%	+0.7%

[2] gApp is accessible at the following link: http://lexytrad.es/gapp/app.php. The application has been registered with Safe Creative under the identifier https://www.safecreative.org/work/2011165898461-gapp.

				ES>FR	69.3%	74%	4.7%	+0.7%
				ES>IT	76%	80%	4%	+0.7%
				ES>PT	68%	74%	6%	+0.7%
			GT	ES>EN	66%	75.3%	9.3%	=
				ES>DE	35.3%	43.3%	8%	=
				ES>FR	65.3%	73.3%	8%	=
				ES>IT	78.7%	79.3%	0.7%	=
				ES>PT	72.7%	79.3%	6.7%	=
5	VPPs	400	GT	ES>EN	21.5%	25%	3.5%	=
				ES>ZH	11%	14%	3%	=
			DL	ES>EN	57%	54.5%	-2.5%	=
				ES>ZH	42.5%	39.5%	-3%	=
6	VPPs	400	VIP	ES>EN	45.5%	67%	21.5%	+0.5%
			DL	ES>EN	77%	85.5%	8.5%	=
			GT	ES>EN	64%	77.5%	13.5%	+1%
x̄		**2460**			**53%**	**63.7%**	**10.7%**	**+0.8%**

Fig. 1: gApp's results through experiments 1-6.

Across these six previous experiments, gApp has been evaluated for different types of multiword expressions (MWEs), specifically verb-noun constructions (VNC) and verb+prepositional phrase (VPP) constructions. It has also been tested with various NMT systems, including DeepL (hereinafter "DL"), Google Translate (hereinafter "GT"), ModernMT (MMT), and VIP, as well as multiple translation directions (ES/FR > ES/PT/EN/DE/FR/IT/ZH). Analysing a total of 2,460 cases, the average (x̄) results indicate that NMT accuracy improved from 53% before gApp to 63.7% after its application—an overall increase of 10.7%. Notably, gApp's results closely align with those achieved through manual conversion, the gold standard, which outperformed gApp by just 0.8% (i.e., it attained 11.5% improvement). Against this backdrop, one of the primary objectives of the present study is to examine the extent to which the results obtained here align with those from our six previous experiments, with the aim of validating our hypothesis: that gApp can enhance NMT performance by automatically converting source-text discontinuous idioms into their continuous forms, also in the case of VPP MWEs in the ES>EN and ES>ZH translation directions.

To our knowledge, this is the first study to evaluate gApp's effectiveness in translating VPP idioms via neural machine translation in the ES>EN/ZH language pairs. To achieve this, we will analyse the performance of GT and DL across 500 test cases—250 using the discontinuous form of the idioms and 250 using their continuous form. These results will be compared against both gApp's conversion and the manual conversion, with the latter serving as the gold standard. The idioms under analysis include Spanish VPP somatisms—expressions involving lexemes related to body-parts (Mellado Blanco, 2004)—such as *estar hasta los cojones, estar hasta los huevos, estar hasta las narices, estar hasta el gorro,* and *estar hasta la coronilla,* all variants through paradigmatic relation with

the literal meaning of 'to be up to the balls/the noses/the hat/the crown of my head' and the figurative meaning of 'being completely fed up or sick and tired'.

The remainder of this paper is structured as follows: Section 2 reviews related work, while Section 3 provides an overview of gApp. Section 4 outlines the research methodology, and Section 5 first assesses gApp's precision and recall before evaluating to what extent it can enhance DL and GT's performance when handling idiom discontinuity in Spanish-to-English and Spanish-to-Chinese translation directions. Section 6 discusses the findings, and Section 7 concludes with insights on how gApp could contribute to the advancement of idiom-aware NMT systems as well as with future avenues of research.

2. Related work

The pursuit of optimising how NMT systems handle MWEs has fostered extensive research and debate. In this context, a growing body of literature has already contributed significant progress, particularly in the field of MWE detection (Riedl & Biedman, 2016; Buljan & Šnajder, 2017; Klyueva *et al*., 2017; Piasecki & Kanclerz, 2022, among others). More specifically, substantial advancements have been made in the automatic identification of discontinuous MWEs (Schneider *et al*., 2014; Berk *et al*., 2019; Foufi *et al*., 2019; Rohanian *et al*., 2019; Lion-Bouton *et al*., 2023) and in improving how MWEs are processed by NMT systems (Huang *et al*., 2018; Zaninello & Birch, 2020; Joshi & Katytyan, 2023).

Within this landscape, analogously to gApp, several other state-of-the-art MWE processing systems also employ a Lexicon Lookup Method (Ramisch & Villavicencio, 2018), i.e., these systems can perform a token-based MWE identification relying on predefined lexicons of patterns to automatically detect MWEs in running text. Among these tools, some specifically target discontinuous expressions, employing analogous strategies to gApp. These include setting a gap-length parameter to define the maximum number of tokens allowed within an MWE (Ramisch, 2015), using surface realization schemas (SRS) that encode information about word order, continuity constraints, and inflectional variations (Alegria *et al*., 2004), and requiring that both preposed and postposed lexical items in a discontinuous MWE trigger the identification process simultaneously (Foufi *et al*., 2019), among other techniques.

However, the question of how to preprocess discontinuous idioms to improve NMT remains underexplored. In this context, the text preprocessing tool gApp represents a significant advancement, since, to the best of our knowledge, it is the only system on the market that is able to both detect discontinuous MWEs in running text and automatically convert them into their continuous form. This capability greatly enhances the performance of leading NMT systems, such as GT and DL, as will be shown in the following sections.

3. Overview of gApp

gApp was developed using Python 3.7, along with the Spacy library, which specialises in a wide range of advanced natural language processing tasks. These include non-destructive tokenization, part-of-speech tagging, dependency parsing, lemmatisation, and rule-based matching, among others (Honnibal & Montani, 2018). Specifically, the pretrained machine-learning model for Spanish *es_core_news_sm* was employed—a multi-task convolutional neural network (CNN) trained on WikiNER (Nothman *et al*., 2017) and UD Spanish AnCora (Martínez Alonso & Zeman, 2016).

Against this backdrop, gApp was developed to perform two core tasks in the preprocessing of MWEs: first, the detection of discontinuous idioms, and second, their conversion into continuous forms to improve NMT performance.

3.1. Automatic detection of discontinuous idioms

The system gApp relies on a token-based approach for MWE identification. To achieve this, following a Lexicon Lookup Method (Ramisch & Villavicencio, 2018), it utilises a predefined lexicon of semi-fixed idioms—expressions that allow for internal morphosyntactic variation from their canonical forms. In this sense, such idioms may appear in discontinuous forms within texts, with other elements inserted between their constituent parts.

Before implementing the detection patterns, it was essential to determine which types of n-grams could appear within the discontinuous form of the idioms. To achieve this, we analysed two large web-crawled Spanish corpora —*esTenTen18* and the *Timestamped JSI Web Corpus 2014-2021 Spanish*— both available through Sketch Engine. The *esTenTen18* corpus consists of over 17 billion words from a diverse range of Spanish-language sources, encompassing both European and Latin American varieties. It includes a heterogeneous mix of text types and diasystematic variations, as well as user-generated content. The *Timestamped JSI Web Corpus 2014-2021 Spanish* contains over 16.4 billion words extracted from news articles via RSS feeds (Kilgarriff *et al*., 2004).

Building on the corpus-based research methodology proposed by Hidalgo-Ternero & Corpas Pastor (2020), we used Sketch Engine's Corpus Query Language (CQL) schemas to extract both the discontinuous forms of the somatisms under study—hereafter referred to as *relevant results*—and other concordances with similar patterns but unrelated to the somatisms (*irrelevant results*). This process helps define the necessary constraints for gApp to optimise its precision and recall. The CQL schema used in this study is presented in Figure 2.

Sequence	CQL schema
Estar [1-3 tokens] hasta el/la/los/las cojones/huevos/narices/gorro/coronilla	[lemma="estar"][]{1,3}[word="hasta"][lemma="el"] [word="cojones|huevos|narices|gorro|coronilla"]

Fig. 2: CQL schema to retrieve the discontinuous form of the somatisms under study.

To maximise the detection of relevant results while filtering out as many irrelevant ones as possible, several rule-based matching patterns were incorporated into the lexicon. These patterns consist of a list of dictionaries, each containing detailed descriptions of both the fixed tokens within the idiom and the elements that may appear within the sequence. For example, in the case of the somatism *estar hasta los cojones*, corpus analysis revealed that the verb *estar* undergoes inflection. Consequently, the algorithm classifies it as a lemma to ensure accurate detection. In contrast, the remaining tokens (*hasta los cojones*) are set to be recognised in their canonical form, as they do not undergo any morphological changes.

Regarding the intervening elements within the idiom (the gap), the corpus data show that this phrase can be split by various adverbial, nominal, and prepositional phrases, such as *ya* ('already'), *un poco* ('a bit'), *de ti* ('of you'), or even different subjects. To account for this, gApp restricts the first token within the gap to some specific grammatical categories: determiners, adverbs, prepositions, pronouns, and proper nouns. Additionally, for tokens within the gap that may or may not appear, the *optional* attribute is assigned. Finally, to exclude irrelevant results, the algorithm filters out cases where the last token before *hasta los cojones* belongs to one of the following grammatical categories: verbs, conjunctions, subordinate conjunctions, or prepositions. Corpus analysis showed that when these categories directly preceded the trigram *hasta los cojones*, the results were unrelated to discontinuous instances of *estar hasta los cojones* and needed to be filtered out to enhance gApp's precision.

3.2. Automatic conversion of discontinuous idioms into their continuous forms

After the detection phase, the system proceeds to a second stage, which automatically converts discontinuous idioms into their continuous forms. This process is executed using a *for-loop* that first checks whether the algorithm detects any predefined pattern within the somatism lexicon.

If a match is found, the system identifies the first dictionary entry in the match (*pos_ini*, or "initial position") and the last one (*pos_fin*, or "final position"), treating the gap between them as optional elements within the sequence. The first token within the gap (*gap1*) is assigned the position *pos_ini + 1*, while the last token (*gap3*) is set to *pos_fin-3*.

Once the gap is defined, the algorithm reconstructs the text by concatenating:

1. The portion of text from the beginning up to and including *pos_ini*.
2. The segment named *pos_fin*.
3. The content from *gap1* to *gap3*.
4. The portion following *pos_fin* up to the end of the text.

This process results in the final output with the idiom in its continuous form. If none of the predefined patterns in the lexicon is detected, gApp makes no

modifications to the text. The process is repeated iteratively until all discontinuous idioms in the text have been converted.

4. Methodology

This section outlines the research methodology used to evaluate the extent to which gApp can enhance the performance of GT and DL in the ES>EN and ES>ZH translation directions. Since the API version of DL provides the option of selecting either the "Classic language model" (described by the company as NMT) or the "Next-gen language model" (a translation-specific LLM), and given that the objective of this study was to evaluate how gApp can improve Neural Machine Translation, we opted for DL's "Classic language model."

Similarly to Hidalgo-Ternero & Corpas Pastor (2020), the concordances containing the discontinuous somatisms under study were extracted from the Spanish corpora described in Section 3.1: esTenTen18 and the Timestamped JSI web corpus 2014-2021 Spanish. To analyse the different translation outcomes provided by GT and DL for the source-text (ST) somatisms in English and Chinese, we used the Sketch Engine corpora enTenTen20 (36.5 billion words) and the Timestamped JSI web corpus 2014-2021 English (60.4 billion words) for English, as well as zhTenTen17 Simplified (13.5 billion words) for Chinese.

Despite the challenges posed by user-generated content—including frequent ST errors, noise, and out-of-vocabulary tokens—even for the most advanced NMT systems (Belinkov & Bisk, 2018; Lohar *et al*., 2019), a heterogeneous sample was selected. This sample incorporated diverse language varieties, text sources and types (including user-generated content) to mitigate sampling bias, which could arise from exclusively analysing NMT canonical training data for the somatisms under study.

A total of 500 cases were examined, consisting of 250 discontinuous and 250 continuous forms (i.e., after conversion) of the somatisms *estar hasta los cojones*, *estar hasta los huevos*, *estar hasta las narices*, *estar hasta el gorro*, and *estar hasta la coronilla*, with varying unigrams, bigrams, or trigrams inserted between their components. These MWEs were chosen because they exhibit a range of linguistic properties, such as idiomaticity, non-compositionality, cultural specificity, syntactic flexibility (with the potential for both continuous and discontinuous form), as well as a high frequency in the selected corpora and a high degree of variability, in order to examine the performance of NMT systems under the challenge of idiom variation.

For each somatism, 50 irrelevant results were compiled to initially calculate both precision and recall, taking into account all idiom constituents. Below, we provide examples of irrelevant results for *estar hasta los cojones* and *estar hasta las narices*, where the irrelevant sequence is highlighted in bold.

(1) Aquel toro no se me iba a escapar por nada del mundo y **estaba** dispuesto a cortarle **hasta los cojones**, si fuese preciso.

(2) Se necesita ayuda de Estados Unidos en materia de inteligencia policial, porque el país vecino **está** involucrado **hasta las narices** en el problema de la inseguridad.

Examples 1 and 2 show instances of the patterns "estar (1-3 tokens) hasta los cojones" and "estar (1-3 tokens) hasta las narices", which nevertheless are not related to the discontinuous forms of the analysed idioms *estar hasta los cojones* and *estar hasta las narices*, respectively. In Example 1, *hasta los cojones* is used literally: the speaker (a matador) intends to emphasise his determination to even cut off the bull's testicles because "there is no way that bull is going to get away from me." In this sentence, *estar* and *hasta los cojones* belong to different clauses. *Estar* functions as the copula for the predicate adjective *dispuesto* ('ready'), while *hasta los cojones* serves as the direct object of the verb *cortar* ('to cut'), forming part of the subordinate clause introduced by the main clause *estaba dispuesto a…* In Example 2, *estar hasta las narices* represents a different idiom from those analysed in this study. It is a variant of *estar hasta el cuello*, meaning 'completely' (*completamente*). In this context, "el país vecino está involucrado hasta las narices en el problema de la inseguridad" can therefore be translated as "the neighbouring country is up to its neck in the problem of insecurity". These examples are hence considered irrelevant results and must therefore be filtered out (i.e., not converted) by gApp.

After quantifying both precision and recall, the results of GT and DL's performance for the various concordances were categorised into three main groups: before conversion, after automatic conversion with gApp, and after manual conversion (the gold standard). The same analysis was conducted for both translation directions: ES>EN and ES>ZH.

The NMT outputs for these different scenarios were then manually evaluated by two professional translators specialising in Spanish-to-Chinese/English translation. Evaluator 1 is a native speaker of Spanish, and Evaluator 2 is a native speaker of Chinese, however, they both have over five years of experience in those translation directions (Spanish>Chinese/English and Chinese>Spanish/English). Evaluator 1 has a Bachelor's degree in Translation and Interpreting, and Evaluator 2 has a Master's degree in Spanish Philology, with a specialisation in Translation and Interpreting (Spanish<>Chinese).

Manual evaluation was chosen due to the challenges that automatic metrics face in accurately assessing idiom translation, especially since they regard the full translation and, hence, the specific rendering of idioms into the target text is often overlooked (Baziotis *et al*., 2023). Since both evaluators worked independently on the data set, in order to ensure consistency and a higher inter-annotator agreement (IAA), we provided annotation guidelines. In this context, using a binary evaluation scale (1 = good, 0 = bad), the evaluators were asked to focus exclusively on the accurate translation of those idioms in the target text. They were specifically instructed to assess whether discontinuous idioms were

properly handled—considering both the preservation of idiomatic meaning and the appropriate treatment of any intervening elements. For continuous idioms, a score of 1 was assigned if the system correctly identified and translated the Spanish idiom into an accepted English and/or Chinese equivalent (e.g., ES: "Él está hasta los cojones de su jefe." → EN: "He is fed up with his boss." / ZH: "他受够了他的老板。"). If the output was literal or inappropriate (e.g., ES: "Él está hasta los cojones de su jefe." → EN: "He is up to the balls of his boss." / ZH: "他受够了老板的蛋蛋。"), it received a score of 0. For discontinuous idioms, a score of 1 was given only if the translation successfully captured both the idiomatic meaning and appropriately handled the intervening element (e.g., ES: "Estoy de ti hasta las narices." → EN: "I'm sick and tired of you." / ZH: "我受够你了。"). However, if the idiom was translated accurately but the intervening element was omitted (e.g., ES: "Estoy de ti hasta las narices." → "I'm sick and tired." / ZH: "我受够了。"), the output was considered incorrect and scored 0.

To ensure the reliability of the human evaluation results, IAA was calculated using Cohen's kappa, which yielded a score of 1. This indicates perfect agreement between the two evaluators on the scores assigned to the dataset examples. While such a high level of agreement is rare in MT evaluation frameworks like MQM or Adequacy/Fluency ratings (Rossi & Carré, 2022), it is more attainable when using a binary evaluation scale.

Regarding the five idioms under study, they predominantly appear in their continuous forms within the esTenTen18 corpus, with frequencies of 85.2% for the continuous form of *estar hasta los cojones*, 86% for *estar hasta los huevos*, 85.3% for *estar hasta las narices*, 85% for *estar hasta el gorro*, and 76.4% for *estar hasta la coronilla*. This means that continuous occurrences are approximately five times more frequent than discontinuous ones. Based on this, our study aims to test the hypothesis that gApp can enhance NMT performance by converting idioms into their canonical continuous forms, specifically focusing on Verb + Prepositional Phrase (VPP) idioms in the ES>EN/ZH translation directions.

5. Results

This section presents the results in two primary phases: first, we detail gApp's precision and recall for each analysed idiom, and second, we assess how effectively gApp improves GT and DL's handling of idiom discontinuity in the ES>EN and ES>ZH translation directions.

Concerning gApp's performance, for the idiom *estar hasta los cojones*, 49 instances were automatically converted, all of which were true positives, resulting in a precision of 100% (49/49) and a recall of 98% (49/50). For *estar hasta los huevos*, the system made 52 conversions—48 true positives and 4 false positives—yielding a precision of 92.3% (48/52) and a recall of 96% (48/50). The idioms *estar hasta las narices* and *estar hasta la coronilla* were each successfully converted 50 times, all true positives, achieving 100% precision and recall. For

estar hasta el gorro, gApp executed 53 conversions, 50 of which were true positives and 3 false positives, resulting in a precision of 94.3% (50/53) and a recall of 100% (50/50). These findings are summarised in Figure 3.

	Precision	Recall	F1
estar hasta los cojones	100%	98%	99%
estar hasta los huevos	92.3%	96%	94.1%
estar hasta las narices	100%	100%	100%
estar hasta el gorro	94.3%	100%	97,1%
estar hasta la coronilla	100%	100%	100%
Final average	97.3%	98.8%	98%

Fig. 3: gApp's precision and recall.

Regarding the improvement in GT's output before and after gApp's application, Figure 4 reveals varying impacts by idiom and translation direction.

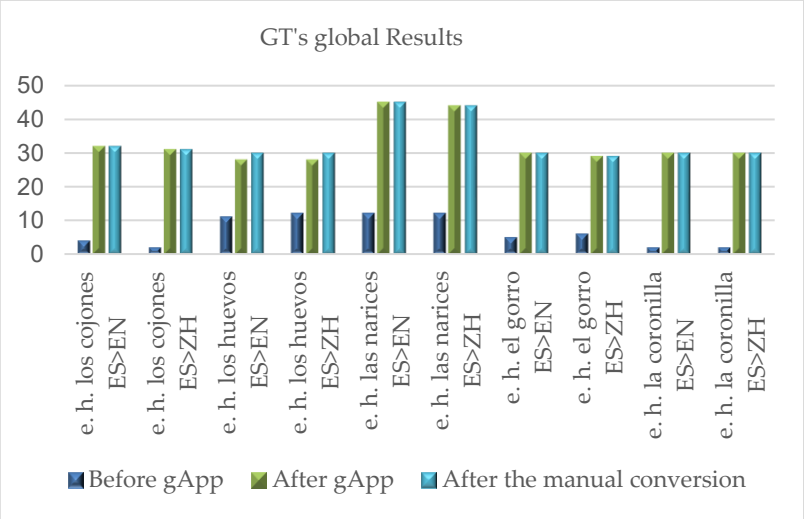

Fig. 4: GT's performance for the idioms.

As Figure 4 illustrates, except for *estar hasta los huevos*, gApp's performance matched that of manual conversion. For *estar hasta los huevos*, gApp produced a 56% improvement, compared to 60% with manual conversion in both translation directions. Although two cases were not converted, gApp still improved GT's accuracy by 34% (17 cases) in ES>EN and 32% (16 cases) in ES>ZH. Manual conversion led to improvements of 38% and 36%, respectively.

For *estar hasta los cojones*, transforming it into a continuous form enhanced ES>EN translation by 56% and ES>ZH by 58%. *Estar hasta las narices* saw substantial improvements of 66% (ES>EN) and 64% (ES>ZH). For *estar hasta el gorro*, both gApp and manual methods improved translations by 50% and 46% for ES>EN and ES>ZH, respectively. *Estar hasta la coronilla* also yielded a 56% enhancement in both directions.

Figure 5 illustrates the gApp conversion for *estar hasta los cojones*, while Figure 6 outlines GT's performance before and after conversion (the complete sequence is in bold and the idiom is underlined).

	KWIC extracts
ST [ES] Disc. form, pre-gApp	Nos lo dice la intuición, a falta de una comparación sistemática, atendiendo a las proclamas del coro –"Erradicar la pobreza, repartir la riqueza", " **Estamos de ladrones hasta los cojones** "– y la introducción de sesgo feminista de Praxágora.[3]
ST [ES] Cont. form, post-gApp	Nos lo dice la intuición, a falta de una comparación sistemática, atendiendo a las proclamas del coro –"Erradicar la pobreza, repartir la riqueza", " **Estamos hasta los cojones de ladrones** "– y la introducción de sesgo feminista de Praxágora.

Fig. 5: ST KWIC extracts with estar hasta los cojones before and after gApp.

	GT's outcomes
TT [EN] Disc. form pre-gApp	Intuition tells us, in the absence of a systematic comparison, taking into account the proclamations of the choir –"Eradicate poverty, distribute wealth", "**We are thieves to the balls**" – and the introduction of a feminist bias by Praxagora.
TT [EN] Cont. form post-gApp	Intuition tells us, in the absence of a systematic comparison, taking into account the proclamations of the choir –"Eradicate poverty, distribute wealth", "**We are fed up with thieves**" – and the introduction of a feminist bias by Praxagora.
TT [ZH] Disc. form pre-gApp	直觉告诉我们，在没有系统比较的情况下，考虑到合唱团的宣言——"消除贫困，分配财富"，"**我们是窃贼**"——以及普拉克萨戈拉引入的女权主义偏见。
TT [ZH] Cont. form pre-gApp	[…] 直觉告诉我们，在没有系统比较的情况下，考虑到合唱团的宣言——"消除贫困，分配财富"，"**我们受够了小偷**"——以及普拉克萨戈拉引入的女权主义偏见。

Fig. 6: GT's outcomes for ST estar hasta los cojones *before and after gApp.*

The examples in Figure 6 reveal clear differences in the outcomes before and after the automatic conversion of the ST somatism in both ES>EN and ES>ZH translation directions. The pre-conversion outputs *estar de ladrones hasta los cojones* rendered "to be thieves to the balls" (EN) and "是窃贼" *shì qièzéi* (ZH), which means 'to be a thief.' The English translation is literal, mapping "balls" to *cojones*, whereas the Chinese output omits the idiom, resulting in a loss of figurative meaning. After gApp's intervention, GT accurately interpreted and conveyed the idiom in both target languages.

For DL, Figure 7 illustrates translation quality before and after using gApp. The system's automatic conversions consistently aligned with manual conversions for all the idioms. For *estar hasta los cojones*, accuracy increased by 24% (12 cases) in both ES>EN (from 32 to 44) and ES>ZH (from 27 to 39). *Estar hasta los huevos* saw improvements of 16% (8 cases) in ES>EN and 30% (15 cases) in ES>ZH. *Estar hasta las narices* improved by 10% (5 cases) for ES>EN and 12% (6

[3] Due to space limitations, we only present here the KWIC extracts for each text that contain the idiom in both its continuous and discontinuous forms, along with their respective translations.

cases) for ES>ZH. *Estar hasta el gorro* improved by 20% (10 cases) in ES>EN and 14% (7 cases) in ES>ZH. *Estar hasta la coronilla* improved by 50% (25 cases) in ES>EN and 34% (17 cases) in ES>ZH.

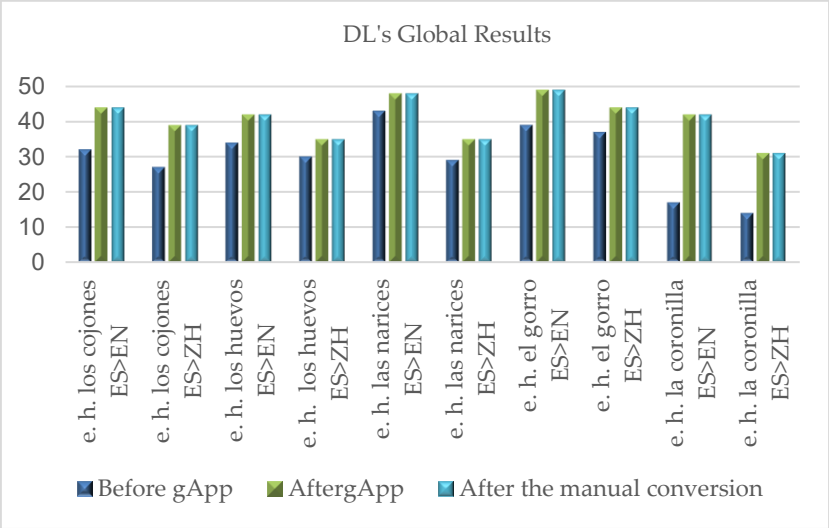

Fig. 7: DL's results for the idioms under study.

Figure 8 provides an overview of the global results. In ES>EN translations using GT, gApp led to a 52.4% improvement (131-case gain), close to the 53.2% improvement from manual conversion (133-case gain). In ES>ZH, gApp increased accuracy by 51.2% (128 cases), while manual conversion yielded 52% improvement (130 cases). For DL, switching to continuous forms enhanced accuracy by 24% (60 cases) in ES>EN and by 18.8% (47 cases) in ES>ZH.

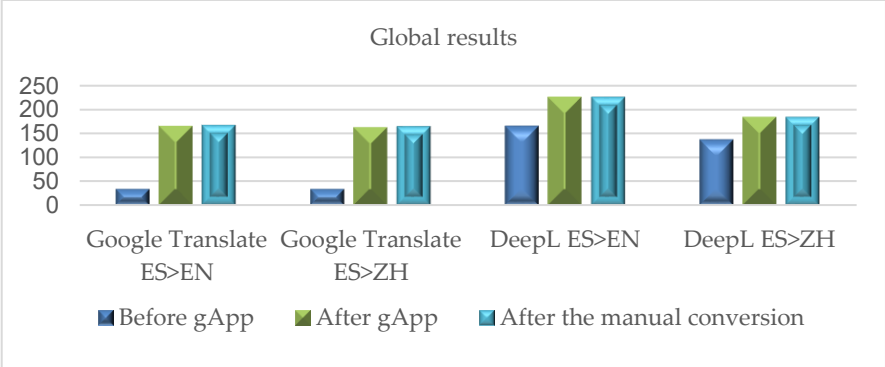

Fig. 8: Global results.

6. Analysis of results

The global results from this study (hereafter called "Paper 7" for contrastive purposes) indicate that gApp's automatic conversions closely matched manual conversions, thanks to gApp's high precision (97.3%) and recall (98.8%). In this sense, only 2.7% of converted cases were irrelevant results, and just 1.2% of relevant results were missed. These metrics slightly outperform global results from previous studies (Papers 1–6), which still showed an average F1 score above 96% (Figure 9).

	Paper 1	Paper 2	Paper 3	Paper 4	Paper 5	Paper 6	Paper 7	Global
Precision	94.8%	95.1%	96.1%	94.9%	95.2%	98.3%	97.3%	96%
Recall	96.8%	97.5%	98.5%	99.3%	96%	92%	98.8%	97%
F1	95.8%	96.3%	97.3%	97.1%	95.6%	95%	98%	96.5%

Fig. 9: gApp's precision and recall from Papers 1-7.

Additionally, global results show that converting discontinuous idioms into their continuous forms significantly improved translation outcomes. In this regard, GT benefited more from gApp than DL. Specifically, gApp raised GT's accuracy by 52.4% (ES>EN) and 51.2% (ES>ZH), compared to manual conversion's 53.2% and 52%. For DL, gains were 24% (ES>EN) and 18.8% (ES>ZH). The smaller gains for DL can mainly be attributed to its higher baseline performance. In ES>EN, DL improved from 66% to 90% post-gApp, while GT rose from 13.6% to 66%. In ES>ZH, DL improved from 54.8% to 73.6%, while GT increased from 13.6% to 64.8%. Despite this contrast, gApp consistently improved NMT outcomes, averaging a 36.6% boost.

As shown in Figure 10, the average accuracy of the NMT systems prior to gApp's intervention in Paper 7 was 37%, which is 16% lower than the weighted average across the six previous experiments (53%). However, the post-gApp improvement in Paper 7 was 26% higher than the average improvement observed in Papers 1–6 (36.6% vs. 10.7%). Moreover, while the difference between gApp and manual conversion was already minimal in Papers 1–6 (a divergence of 0.8%), this gap was reduced even further in Paper 7 to just 0.4%. This indicates that the system has become even more effective in approximating manual conversion.

	Directionality	NMTs	Pre-gApp NMTs' accuracy	Post-gApp NMTs' accuracy	Improvement with gApp	Gold Standard
7	ES>EN	GT	13.6%	66%	52.4%	+0.8%
		DL	66%	90%	24%	=
	ES>ZH	GT	13.6%	64.8%	51.2%	+0.8%
		DL	54.8%	73.6%	18.8%	=
	Total (experiment 7)		37%	73.6%	36.6%	+0.4%
Total (experiments 1-6)			53%	63.7%	10.7%	+0.8%
Total (experiments 1-7)			50.3%	65.4%	15.1%	+0.7%

Fig. 10: Overview of gApp's results for experiments 1-7.

These promising results may be attributed to the inclusion of a different idiom typology (VPP constructions) and the introduction of a new translation direction (ES>ZH). In light of the findings from Paper 7, additional experiments in this new translation direction are necessary to determine to what extent the

conversion of discontinuous idioms can consistently lead to significant improvements in NMT quality from Spanish into Chinese.

Another revealing discovery can be found when comparing both translation directions (ES>EN vs. ES>ZH). In this regard, we can observe that ES>ZH generally delivers a worse performance than ES>EN both with DL (after gApp, 90% in ES>EN vs. 73.6% in ES>ZH) and with GT (66% vs. 64.8%). This gap likely stems from ES>ZH pivoting through English (i.e., ES>EN>ZH), given the abundance of English-centric training data (see Hidalgo-Ternero & Corpas Pastor, 2024, and Hidalgo-Ternero & Zhou-Lian, 2022). Let us observe, for instance, the different performance of GT when translating *estar hasta los cojones* into English vs. into Chinese in Figure 11.

	KWIC extracts
ST [ES]	En definitiva, se gastaron un pastón en empeorar algo que no necesitaba cambios y **están hasta los cojones** todos los médicos del programita…
	GT's outcomes
TT [EN]	In short, they spent a lot of money to make something worse that didn't need changes and all the doctors in the little program **are fed up with the balls…**
TT [ZH]	简而言之，他们花了很多钱使不需要改变的事情变得更糟，并且小程序中的所有医生都**受够了球…**

Fig. 11: Instance of GT's mistranslation in English and Chinese for estar hasta los cojones.

Figure 11 illustrates that, in the ES>ZH translation direction, *estar hasta los cojones* was translated as *受够了球 shòu gòu le qiú* ('be fed up with ball'). In fact, *受够了shòu gòu le* is the appropriate translation for this idiom, but, by adding the object *球 qiú* ('ball'), the translation becomes inadequate. The English target text (*are fed up with the balls*) can explain this mistranslation into Chinese. In this regard, the Chinese word 球 qiú ('ball') is completely unrelated (lexically and semantically) to the ST Spanish idiom *estar hasta los cojones*; hence, the final translation into Chinese can only be explained with this pivoting through English:

> *estar hasta los cojones* > *be fed up with the balls* [*balls* (as body part) are lexically related to *cojones* in Spanish] > *受够了球shòu gòu le qiú* [*球 qiú* is lexically related to *balls* (as in the 'round object') in English]

These errors caused by pivoting through English highlight the necessity of more NMT bilingual training data in non-English language pairs, to reduce English-centric NMT outputs.

7. Conclusion

The results of this study confirm our hypothesis: gApp can significantly improve NMT output for discontinuous idioms by transforming them into their continuous forms, especially for VPP idioms in ES>EN and ES>ZH. In this regard, the system enhanced average performance by 36.6%, nearly equalling manual conversion (37%).

These promising outcomes for flexible VPP somatisms in the ES>EN and the ES>ZH translation directions indicate potential for extending gApp's detection lexicon and conversion mechanism to assess whether it can also enhance NMT for other discontinuous MWE typologies, such as collocations or verb-particle constructions. Additionally, further research is needed to evaluate the scalability of this system to other language-specific preprocessing tools for the automatic conversion of discontinuous idioms in other syntactically-flexible languages, to enhance idiom-aware NMT systems.

Future work will also focus on robustness testing to assess the extent to which gApp can handle unexpected, invalid, or stressful conditions without failing. Examples of those conditions include inputting large corpora of texts, heavily noisy user-generated content, extremely long sentences, or running the system under high user load or a constrained network. This stress-testing will allow us to evaluate whether gApp crashes, distorts meaning or chokes on edge cases. Based on the results, necessary adjustments can be made in the system to prevent the propagation of errors downstream into machine translation systems as well as to ensure that gApp successfully flags problematic cases with useful error handling, rather than silently misprocessing them.

Finally, another future avenue of research will be to examine the extent to which gApp can also improve the performance of state-of-the-art Large Language Models (LLMs) in the machine translation of discontinuous MWEs. In contrast to traditional NMT systems, which are trained exclusively for translation, LLMs function as general-purpose models capable of performing translation via context-sensitive prompts (Wang *et al*., 2023), alongside a wide range of other tasks. Given their versatility, LLMs present promising potential for addressing linguistic challenges such as discontinuous MWEs. However, systematic evaluation remains necessary, particularly in comparison with state-of-the-art NMT systems, as research in this area is still scarce due to the relative novelty of LLMs.

Bibliography

ALEGRIA, I., ANSA, O., ARTOLA, X., EZEIZA, N., GOJENOLA, K. & URIZAR, R. (2004). Representation and treatment of multiword expressions in Basque. In *Proceedings of the Second ACL Workshop on Multiword Expressions: Integrating Processing* (pp. 48–55). Association for Computational Linguistics.

BAZIOTIS, C., PRASHANT, M. & HASLER, E. (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*, (pp. 3682–3700). Association for Computational Linguistics.

BELINKOV, Y. & BISK, Y. (2018). Synthetic and natural noise both break neural machine translation. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*. https://arxiv.org/abs/1711.02173

BERK, G., ERDEN, B. & GÜNGÖR, T. (2019). Representing overlaps in sequence labeling tasks with a novel tagging scheme: bigappy-unicrossy. In *International Conference on Computational Linguistics and Intelligent Text Processing* (pp. 622-635). Springer Nature Switzerland.

BULJAN, M., & ŠNAJDER, J. (2017). Combining linguistic features for the detection of Croatian multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)* (pp. 194-199). Association for Computational Linguistics.

COLSON, J.-P. (2019). Multi-word Units in Machine Translation: why the Tip of the Iceberg Remains Problematic–and a Tentative Corpus-driven Solution. In G. Corpas Pastor, R. Mitkov, M. Kuilovskaya & M. A. Losey León (Eds.), *Proceedings of the Third International Conference EUROPHRAS 2019* (pp. 145–156). Tradulex.

CONSTANT, M., ERYIĞIT, G., MONTI, J., VAN DER PLAS, L., RAMISCH, C., ROSNER, M. & TODIRASCU, A. (2017). Multiword expression processing: A survey. *Computational Linguistics*, *43*(4), 1–92. https://doi.org/10.1162/COLI_a_00302

FOUFI, V., NERIMA, L. & WEHRLI, E. (2019). Multilingual parsing and MWE detection. In Y. Parmentier & J. Waszczuk (Eds.), *Representation and parsing of multiword expressions: Current trends*, (pp. 217–237). Language Science Press. https://langsci-press.org/catalog/view/202/2028/1554-1

HIDALGO-TERNERO, C. M. (2021). El algoritmo ReGap para la mejora de la traducción automática neuronal de expresiones pluriverbales discontinuas (FR>EN/ES). In G. Corpas Pastor, M. Rosario Bautista Zambrana & C. M. Hidalgo-Ternero (Eds.), *Sistemas fraseológicos en contraste: enfoques computacionales y de corpus* (pp. 253-270). Comares.

HIDALGO-TERNERO, C. M. (2024). ¿DeepL, Google Translate o VIP? Qué sistema ofrece un mejor rendimiento en la traducción de locuciones continuas y discontinuas. In G. Corpas Pastor & F. J. Veredas Navarro (Eds.),

Tecnologías lingüísticas multilingües: desarrollos actuales y transición digital (pp. 61-78). Comares.

HIDALGO-TERNERO, C. M. &, CORPAS PASTOR, G. (2020). Bridging the 'gApp': improving neural machine translation systems for multiword expression detection. *Yearbook of Phraseology, 11*, 61-80. https://doi.org/10.1515/phras-2020-0005

HIDALGO-TERNERO, C. M. & CORPAS PASTOR, G. (2024). ReGap: a text preprocessing algorithm to enhance MWE-aware neural machine translation systems. In J. Monti, G. Corpas Pastor, R. Mitkov & C. M. Hidalgo Ternero (Eds.), *Recent Advances in MWU in Machine Translation and Translation technology* (pp. 18-39). John Benjamins Publishing Company.

HIDALGO-TERNERO, C. M. & CORPAS PASTOR, G. (2025/forthcoming). Qué se traerá gApp entre manos… O cómo mejorar la traducción automática neuronal de variantes somáticas (ES>EN/DE/FR/IT/PT). In M. Seghiri & M. Pérez Carrasco (Eds.), *Nuevos enfoques en traducción científica, técnica y agroalimentaria.* Peter Lang. ISBN: 978-3-631-92973-5.

HIDALGO-TERNERO, C. M. & ZHOU-LIAN, X. (2022). Reassessing gApp: does MWE discontinuity always pose a challenge to Neural Machine Translation? In G. Corpas Pastor & R. Mitkov (Eds.), *Computational and Corpus-Based Phraseology* (pp. 116–132). Springer.

HONNIBAL, M. & MONTANI, I. (2018). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks, and incremental parsing. *Zenodo.* https://doi.org/10.5281/zenodo.1212303

HUANG, P. S., WANG, C., HUANG, S., ZHOU, D. & DENG, L. (2018). Towards neural phrase-based machine translation. *arXiv*. https://arxiv.org/abs/1706.05565

JOSHI, N. & KATYAYAN, P. (2023). Implications of multi-word expressions on English to Bharti braille machine translation. In *6th International Conference on Information Systems and Computer Networks* (pp. 1-5). IEEE.

KILGARRIFF, A., PAVEL SMRZ, P. R. & TUGWELl, D. (2004). The Sketch Engine. In *Proceedings of the 11th EURALEX International Congress* (pp. 105-116).

KLYUEVA, N., DOUCET, A. & STRAKA, M. (2017). Neural Networks for Multi-Word Expression Detection. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)* (pp. 60–65). https://doi.org/10.18653/v1/W17-1707

LION-BOUTON, A., SAVARY, A. & ANTOINE, J. Y. (2023). A MWE lexicon formalism optimised for observational adequacy. In *Proceedings of the 19th workshop on multiword expressions (MWE 2023)* (pp. 121-130).

LOHAR, P., POPOVIĆ, M., ALFI, H. & WAY, A. (2019). A systematic comparison between SMT and NMT on translating user-generated content. In *20th*

International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2019).

MARTÍNEZ ALONSO, H. & ZEMAN, D. (2016). Universal Dependencies for the AnCora treebanks. *Procesamiento del Lenguaje Natural*, *[S.l.], 57*: 91-98. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5341

MELLADO BLANCO, C. (2004). *Fraseologismos somáticos del alemán*. Peter Lang.

NOTHMAN, J., RINGLAND, N., RADFORD, W., MURPHY, T. & Curran, J. R. (2017). Learning multilingual named entity recognition from Wikipedia. *figshare*. Dataset. https://doi.org/10.6084/m9.figshare.5462500.v1

PIASECKI, M. & KANCLERZ, K. (2022). Non-contextual vs contextual word embeddings in multiword expressions detection. In *International Conference on Computational Collective Intelligence* (pp. 193-206). Cham: Springer International Publishing.

RAMISCH, C. (2015). *Multiword Expressions Acquisition: A Generic and Open Framework*. *Theory and Applications of Natural Language Processing series XIV*. Springer.

RAMISCH, C. & VILLAVICENCIO, A. (2018). Computational treatment of multiword expressions. In R. Mitkov (Ed.), *Oxford Handbook on Computational Linguistics* (2ª ed). https://doi.org/10.1093/oxfordhb/9780199573691.013.56

RIEDL, M. & BIEMANN, C. (2016). Impact of MWE resources on multiword recognition. In *Proc. of the ACL 2016 Workshop on MWEs* (pp. 107–111).

ROHANIAN, O., TASLIMIPOOR, S., KOUCHAKI, S., AN HA, L. & MITKOV, R. (2019). Bridging the Gap: Attending to Discontinuity in Identification of Multiword Expressions. In J. Burstein, C. Doran & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 1* (pp. 2692–2698). Association for Computational Linguistics.

ROSSI, C. & CARRÉ A. (2022). How to choose a suitable neural machine translation solution: Evaluation of MT quality. In D. Kenny (Ed.), *Machine translation for everyone: Empowering users in the age of artificial intelligence*, (pp. 51–79). Berlin: Language Science Press. https://doi.org/10.5281/zenodo.6759978

SCHNEIDER, N., DANCHIK, E., DYER, C. & SMITH, N. A. (2014). Discriminative lexical semantic segmentation with gaps: Running the MWE gamut. *TACL, 2*: 193–206.

WANG, L., LYU, C., JI, T., ZHANG, Z., YU, D., SHI, S. & TU, Z. (2023). Document-Level Machine Translation with Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp.

16646–16661), Singapore. Association for Computational Linguistics. https://doi.org/10.48550/arXiv.2304.02210

ZANINELLO, A. & BIRCH, A. (2020). Multiword expression aware neural machine translation. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)* (pp. 3816–3825). European Language Resources Association.

Received: 15/04/2025
Accepted: 10/09/2025