Incorporating Multiword Expressions in Galician Neural Machine Translation: Compositionality, Efficiency, and Performance

Daniel Solla, Paula Pinto-Ferro, Laura Castro
Pablo Gamallo and **Marcos Garcia**
CiTIUS - Centro Singular de Investigación en Tecnoloxías da Información
Universidade de Santiago de Compostela
{pablo.gamallo,marcos.garcia.gonzalez}@usc.gal

Abstract

This paper explores the behavior of neural machine translation models on two newly introduced datasets containing noun-adjective MWEs with different degrees of semantic ambiguity and compositionality. We compare general-domain machine translation systems with fine-tuned models exposed to small subsets of the target MWEs. By assessing the effects of the learning steps and corpus size, we found that carefully designed fine-tuned may improve MWE handling while mitigating catastrophic forgetting. However, our error analysis reveals that models still struggle in several scenarios, particularly when translating MWEs with idiomatic meanings. Both the datasets and the experiments focus on translation involving Galician, English, and Spanish.

1 Introduction

In the last decade, Neural Machine Translation (NMT) evolved from recurrent networks (Kalchbrenner and Blunsom, 2013; Bahdanau et al., 2015) to Transformer-based sequence-to-sequence (*seq2seq*) architectures (Vaswani et al., 2017). More recently, the use of Large Language Models (LLMs) as translation engines has gained popularity, especially due to their impressive performance in high-resource languages (Zhu et al., 2024). However, the performance of *seq2seq* models in low-resource languages remains competitive with current LLMs (Robinson et al., 2023; Gibert et al., 2025), while also providing other advantages such as faster inference and lower computational energy costs.

In either approach, incorporating new lexicon, contexts, or textual domains in a Machine Translation (MT) system can be challenging. On the one hand, training a model from scratch and exploring different hyperparameter configurations incurs substantial computational costs. On the other hand, fine-tuning (FT) an existing model with new datasets risks catastrophic forgetting, where gains on the new data can lead to reduced performance on previously learned contexts or domains (McCloskey and Cohen, 1989; Gu and Feng, 2020).

One particular challenge that NMT systems often struggle with is the accurate translation of Multiword Expressions (MWEs). In addition to being pervasive across all languages (Ramisch, 2023), MWEs present ambiguities at multiple levels. First, from the perspective of semantic compositionality, MWEs exhibit varying degrees of idiomaticity (e.g., the compositional *apple tree*, or *red herring* meaning a 'misleading clue' idiomatically). Second, at the level of individual components, some words within MWEs have high degrees of polysemy (e.g., *green bank* where 'bank' may refer to a financial institution or to a river bank, while 'green' may also have multiple senses). Finally, like single words, MWEs can convey different meanings depending on the context (e.g., *glass ceiling* as an 'invisible barrier' or as a physical structure). Among other factors, these make MWEs particularly challenging to model, not only for NMT but also for neural causal language modeling (Dankers et al., 2022; Liu et al., 2025; He et al., 2025).

This paper investigates continual learning approaches for integrating MWEs into NMT models, and evaluates their performance on such expressions. We focus on the translation from Galician (GL) to two languages: English (EN) —as an example of a distant language— and Spanish (ES) —a related language that exerts influence on Galician. Specifically, we present *i)* two new parallel datasets (Galician-English and Galician-Spanish) composed of sentences with MWEs with different types of ambiguity;[1] *ii)* an extensive set of experiments assessing the influence of learning steps and

[1]The dataset is available at https://github.com/marcospln/parallel_noun-adj_gl-en-es/. A subset of manually created sentences is kept private to prevent data contamination in future research.

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 75–85

size of the training corpus; *iii)* quantitative analyses comparing base and fine-tuned models; and *iv)* a qualitative error analysis showing the effects of the compositionality and frequency levels of the MWEs on their translations.

Our findings show that *i)* the degree of overlap in MWE composition seems to be a function of linguistic proximity (i.e., Galician-Spanish MWEs are more similar than Galician-English equivalents), and this crucially affects translation performance; *ii)* targeted fine-tuning can be an effective and efficient strategy for enhancing MWE translation, although only in some cases; *iii)* NMT still struggles to translate MWEs, primarily due to idiomaticity, with factors such as frequency and linguistic distance also contributing to the difficulty.

2 Related work

Recent studies in low-resource NMT demonstrate that combining multilingual pre-trained models with synthetic corpora can achieve strong translation performance even when high-quality parallel data is scarce (Sant et al., 2024). Fine-tuning self-supervised multilingual models such as mBART on small parallel corpora has proven effective for adapting to new language pairs (Thillainathan et al., 2021). A key challenge in continual training is catastrophic forgetting (McCloskey and Cohen, 1989). Several works analyze forgetting at both module and parameter levels, showing that excessive parameter drift during domain adaptation can degrade general-domain performance (Gu and Feng, 2020). Complementary research highlights that the extent of forgetting is related to the properties of the adaptation data, such as the introduction of new target vocabulary (Saunders and DeNeefe, 2024). These findings suggest that careful, targeted fine-tuning is less likely to harm overall translation quality than broad domain shifts or large, uncontrolled adaptation datasets.

Instruction-based fine-tuning, originally developed for LLMs, has also been adapted to traditional encoder-decoder NMT systems. Such methods allow models to learn multiple translation customization tasks jointly through compact fine-tuning stages, demonstrating that specialized behaviors can be efficiently acquired without full retraining (Raunak et al., 2024). According to recent reviews, multilingual pre-training on generic language tasks allows models to internalize shared structures across languages. This shared knowledge helps compensate for the lack of parallel data during fine-tuning for translation (Ataman et al., 2025).

In the context of MWEs and idiomaticity, early work demonstrated the benefits of detecting and specially handling phrasal verbs, which significantly improved translation consistency and fluency (Kordoni and Simova, 2014). In the same line, MWE-aware NMT approaches using annotation and data augmentation with external linguistic resources have shown substantial improvements (Zaninello and Birch, 2020). More recent research on Transformer-based NMT demonstrates that these systems exhibit an excessive bias toward compositionality, leading to systematic difficulties in modeling non-compositional expressions (Dankers et al., 2022; Liu et al., 2025)

Several datasets have been released aimed at evaluating the performance of NMT and other models on MWEs with different degrees of idiomaticity, such as the English-based MAGPIE (Haagsma et al., 2020), MultiMWE, including Chinese, English, and German (Han et al., 2020b), or the also multilingual AlphaMWE —with the same languages and Polish— (Han et al., 2020a), recently expanded to other varieties including Italian and Arabic (Han et al., 2025).

Regarding NMT for Galician, the *Proxecto Nós* (*Nós Project*) recently released state-of-the-art models[2], and its research reported benefits from architectural adaptations such as smaller BPE vocabularies, which consistently improve performance across data scales (Outeirinho et al., 2024). New resources, including the CorpusNÓS (de Dios-Flores et al., 2024) and parallel datasets[3], provide large-scale training material for both NMT and LLM-based translation models enabling improved translation quality in low-resource settings.

To the best of our knowledge, there is currently no parallel MWE dataset for Galician that provides manually curated translations together with annotations for idiomaticity class, sense, and frequency. Moreover, no prior work has specifically investigated continual learning strategies for NMT with a focus on MWEs including Galician. Our work directly addresses these gaps.

[2] https://nos.gal/gl/proxecto-nos

[3] https://github.com/proxectonos/corpora#traducion-automatica

3 New parallel corpus of noun-adjective MWEs

This section introduces two new parallel corpora (Galician-Spanish and Galician-English) composed of sentences containing fine-grained annotation of noun-adjective MWEs.

Source data: The source data is the dataset presented by Castro et al. (2025), which comprises 240 noun-adjective MWEs in Galician conveying 322 different senses. Each of these senses is contextualized in up to 6 sentences, totaling 1,858 examples (average of 5.77 per sense). The initial identification of the 240 MWEs was performed using *i)* a dependency-based approach, i.e., extracting contiguous and non-contiguous noun adjective pairs linked by a dependency relation, and *ii)* a frequency-based criterion, selecting expressions from two ranges: high and low frequency.

MWEs' properties: At the token-level, each contextualized MWE is classified according to its compositionality class in the given context as *idiomatic* (e.g. *obra morta*, 'freeboard', literally 'death job'), *compositional* (e.g. *centro comercial*, 'shopping center'), or *partial* (e.g. *campo magnético*, 'magnetic field'). At the type-level, MWEs that may have different compositionality scales depending on the context are marked as *potentially idiomatic expressions* (e.g. *montaña rusa*, literally 'Russian mountain' which can refer idiomatically to a roller coaster or to an actual mountain in Russia).

Translations: Each of the 1,858 sentences was manually translated into Spanish and English by a professional translator, and then reviewed by a second one, ensuring the quality of the parallel resource. The translators were asked to perform an adequate translation taking into account both the meaning of the MWE and of the whole sentence, so that the original MWEs (in Galician) were not always translated by a MWE in the target languages (e.g. *fondo mariño*, literally 'sea bottom', translated as 'seabed'). During this process, the translators also identified those elements in the new sentences conveying the meaning of the original MWE, allowing for further qualitative analyses. The final resources are two parallel corpus composed of bilingual pairs of 1,858 sentences. Table 1 includes some examples of the original sentences and their English translations.

Semantic phenomena: The dataset contains three main types of linguistic phenomena that may challenge language modeling in general and NMT in particular: *i)* compositionality class, including idiomatic, partially idiomatic, and compositional expressions (see examples above); *ii)* ambiguity of the components: a component of the MWE, namely the head, may have different meanings: e.g., *número enteiro*, where *número* ('number') may have the following three senses and corresponding translations: mathematical ('integer'), graphical (e.g., 'whole number'), journalistic ('full issue', as in the complete volumes of a journal); *iii)* ambiguity of the whole MWE, e.g., *auga doce* (literally 'sweet water') which may refer to 'fresh water' (vs. 'salty water'), or to 'sweetened water' (i.e., with sugar). It is worth mentioning that the former ambiguities (at the word and at the MWE level) may occur in the same compositionality class or across different scales.

Note that the degree of semantic divergence is considerably smaller between Galician and Spanish than between Galician and English. For many MWEs, a literal, word-for-word translation from Galician into Spanish is often accurate and preserves both meaning and structure, whereas this is not the case for Galician-English. This contrast further motivates the use of GL-EN MWEs as a challenging evaluation setting, since the model must resolve non-literal correspondences that are not predictable from morphology or word-level semantics alone.

4 Materials and methods

Models: All experiments are based on the NMT models developed within the *Proxecto Nós*, publicly available through HuggingFace.[4] These models are implemented using the OpenNMT framework (Klein et al., 2017) and follow a Transformer-based encoder-decoder architecture. They represent state-of-the-art systems for Galician-centric MT (Buján et al., 2025).

Fine-tuning data: We use the new MWE datasets (§ 3) to both fine-tune and evaluate the performance of the NMT models. As mentioned, each parallel corpus is composed of up to 6 sentences for each of the 322 MWE senses. We first select two sets of 322 sentences, using one for fine-tuning and the other for evaluation. We then increase the

[4] https://huggingface.co/collections/proxectonos/mt

MWE	Sentence (GL)	Sentence (EN)
fonte principal	...a principal fonte da cidade...	...the main fountain of the city...
fonte principal	...eran a fonte principal de recrutamento das súas tropas.	...and they were also the main source of recruitment for their troops.
zona húmida	As zonas máis húmidas son as rexións occidental e central...	The western and central regions are the most humid areas...
zona húmida	...das zonas húmidas europeas foron totalmente destruídos.	...of European wetlands were totally destroyed.
centro comercial	É o principal centro comercial e industrial do estado...	It is the main commercial and industrial hub in the state...
centro comercial	O centro comercial tiña no seu proxecto inicial...	The initial design for the shopping mall included...

Table 1: Examples of the original MWEs and sentences in Galician and their English translations.

training data with more sets to analyze the impact of the amount of learning data.[5]

Evaluation sets: We assess the impact of fine-tuning in both generic parallel corpus and in a subset of the MWE dataset for each pair of languages. As standard datasets, we use parallel sentences from Flores (Goyal et al., 2022), Tatoeba (Tiedemann, 2020), and from the *Nós Project*, which contain generic datasets and a detailed test-suite focused on particular linguistic phenomena.[6] Table 2 shows the size of each of the evaluation sets. These test sets allow us to measure both general MT performance and improvements on targeted MWEs, providing a balanced evaluation of fine-tuning effects and potential catastrophic forgetting.

Dataset	GL–EN	GL–ES
Flores	1,012	1,012
Tatoeba	1,018	3,131
Nós_MT_Gold_1	1,777	1,998
Nós_MT_Gold_2	1,777	1,998
Nós_MT_Test-suite	364	334
MWE dataset	322	322

Table 2: Size (in number of sentences) of the test sets.

Evaluation metrics: For general translation quality, we use the standard BLEU and TER metrics.[7] The results on the general MT corpora are reported using the micro-average across the five datasets (Table 2). Besides, during experimentation, model selection is guided by *Score_mwe*, a metric proposed by Zaninello and Birch (2020). This metric is specifically designed to evaluate how accurately MT systems render MWEs, without relying on explicit phrase alignment. It operates at the sentence level by comparing each word of the reference translation of a source-side MWE with the closest matching word in the system hypothesis, using a character-based Levenshtein distance.

5 Experiments

All experiments were conducted using the OpenNMT framework.[8] Fine-tuning was performed on a single NVIDIA Tesla V100S GPU (32GB) on an HPC cluster. All models share the same architecture and base training configuration (detailed in Appendix A). Fine-tuning experiments use the hyperparameter settings of the original models, and differ only in the number of FT steps and the size of the MWE corpus.

As the fine-tuning datasets were designed using Galician MWEs as source, our analyses focus on NMT systems from Galician, although we also refer to the other translation directions when necessary.

Experiment 1. Learning steps: We conducted a series of experiments varying the number of fine-tuning steps, using values from 100 to 2,000 in increments of 100. For each translation direction we used one set of 322 sentences for fine-tuning, and the other for evaluating.[9]

[5] Some of the final sets contain slightly fewer than 322 sentences, particularly for low-frequency senses.

[6] https://github.com/proxectonos/corpora

[7] We also computed ChrF, BERTScore, and METEOR, but we mostly rely on BLEU to discuss the results as in general all of these metrics correlate in our experiments.

[8] https://opennmt.net

[9] To ensure that the results do not depend on the sets used, we also reverse their role for training and testing, with almost

Experiment 2. Corpus size: To assess the impact of the number of examples used for continual learning, we incrementally fine-tuned the model with additional sets of ≈322 sentence pairs (from 1 to 5 sets), until reaching 1,536 instances, i.e., the full dataset except for the held-out set. In each case, we fine-tuned the models during 600 and 1,200 steps.

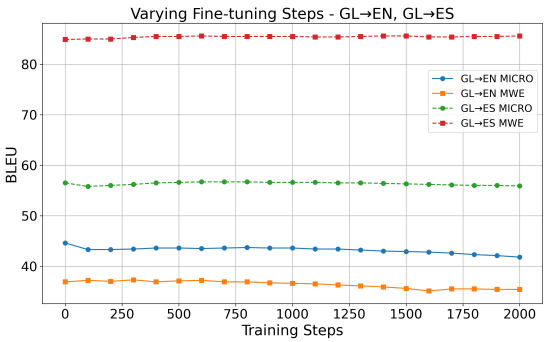
6 Results and discussion

Experiment 1. Number of Fine-Tuning Steps: Figure 1 displays the BLEU learning curves of GL-EN and GL-ES from 0 (original model) to 2,000 fine-tuning steps. Figure 1a shows the micro-average in the general corpora and in the MWE test sets, while Figure 1b plots the results of the *Score_mwe* metric. In GL→EN, as fine-tuning progresses, average scores steadily decline in BLEU, indicating gradual drops in overall translation quality. Meanwhile, MWE evaluations show small gains in BLEU and *Score_mwe* early on, but the improvements do not hold and drop rapidly, showing that targeted MWE updates offer limited benefits and cannot compensate for the declining general performance. In GL→ES, average scores remain stable throughout fine-tuning, with BLEU around 56 and consistent general translation quality. MWE evaluations show strong and sustained improvements, reaching *Score_mwe* above 0.98. In this respect, it is worth recalling that, due to the linguistic similarity between the two languages, MWEs in Galician can be translated literally into Spanish in most cases (see § 3).

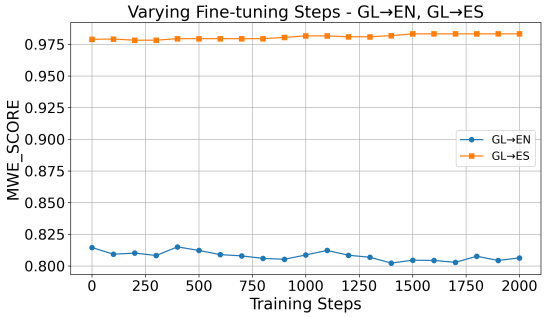
In general, the number of fine-tuning steps influences both overall translation quality and MWE learning. Early steps primarily stabilize general performance, while moderate fine-tuning (600–1000 steps, depending on the translation direction) tends to maximize MWE gains without significantly destabilizing the model. However, excessive fine-tuning steps can lead to overfitting or gradual drops in general BLEU scores, showing the need for careful step selection to balance MWE learning with overall translation quality.

In sum, the results of this set of experiments suggest that continual learning of MWEs, especially for a distant linguistic variety and with limited data (just 322 sentence pairs) is a hard task that requires careful analysis of the trade-off between learning new expressions and overall performance of the

identical results.



(a) BLEU scores GL→EN, GL→ES



(b) *Score_mwe* GL→EN, GL→ES

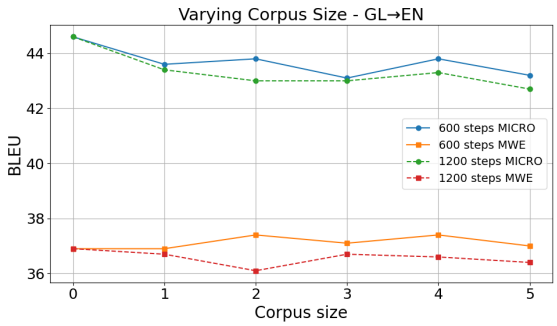
Figure 1: Effect of varying fine-tuning steps on BLEU and *Score_mwe* across different translation directions.

NMT systems. In the next experiment, we explore whether more training data enables improvements in the performance of the models.

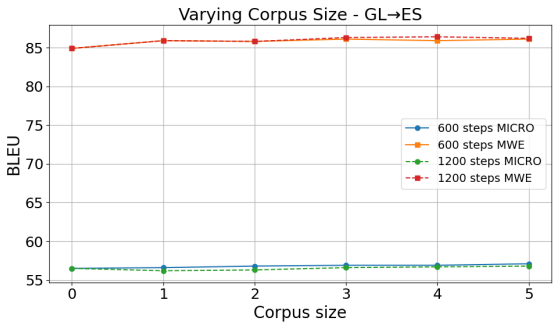
Experiment 2. Corpus Size: In GL-EN, increasing the MWE learning data involves a gradual loss in overall translation quality, although with modest gains in the performance on the MWE dataset (left column of Figure 2). In GL-ES, more training data also generally improves the quality of MWE translations while maintaining competitive results in the general domain, although the BLEU in the MWE dataset was high in the original model (Figure 2, right column).

Overall, the impact of increasing the MWE fine-tuning corpus size is highly direction-dependent. While larger data can substantially improve MWE translation, these gains are not always aligned with improvements in general translation quality. Taken together, these results show that scaling MWE fine-tuning data is more effective when combined with moderate fine-tuning steps and when the translation direction exhibits inherent robustness to domain-specific updates.

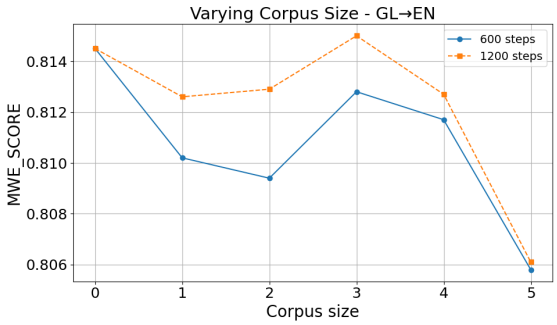
Best model selection: The best-performing models for each translation direction were selected



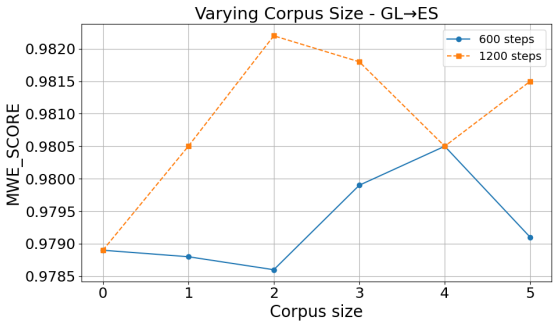
(a) BLEU scores



(c) BLEU scores



(b) *Score_mwe*



(d) *Score_mwe*

(A) GL→EN translation vs. MWE corpus size.

(B) GL→ES translation vs. MWE corpus size.

Figure 2: Translation performance as a function of MWE corpus size for GL→EN and GL→ES.

based on a balanced evaluation of general-domain translation quality and MWE-specific performance. In particular, model selection aimed to maximize improvements in MWE translation accuracy while preserving performance on general test sets, thereby minimizing the risk of catastrophic forgetting. This selection criterion ensures that improvements in MWE handling are not obtained at the expense of overall translation quality.

For each of the four translation directions, we compare the original baseline model with its fine-tuned counterpart. General translation quality is assessed using standard MT metrics (BLEU and TER) computed over multiple general-domain test sets and aggregated using micro-averaging. In parallel, MWE-specific performance is evaluated on a dedicated MWE test set using the same metrics, together with the value of the *Score_mwe*. Table 3 summarizes the results for the models (original and fine-tuned versions), reporting micro-average BLEU and TER scores on the general-domain test sets alongside the corresponding scores on the MWE test set. Although they are not the focus of this analysis, we include the results of the models from English and Spanish to Galician.

Model	**General**		**MWE**	
	BLEU	**TER**	**BLEU**	**TER**
GL→EN B	44.6	40.4	36.9	46.5
GL→EN FT	43.6	40.6	36.9	**46.3**
EN→GL B	38.6	45.9	37.9	47.7
EN→GL FT	**38.9**	**45.1**	**40.2**	**45.5**
GL→ES B	56.5	33.2	84.9	8.8
GL→ES FT	**56.6**	**32.7**	**86.3**	**7.9**
ES→GL B	52.3	36.3	83.7	9.2
ES→GL FT	51.6	**36.1**	83.4	**9.0**

Table 3: Summary of best models (Base and Fined-Tuned) for all translation directions. Results are general-domain and MWE-specific micro-averaged BLEU and TER. Numbers in bold are FT models with better results.

7 Qualitative Analysis

To complement the quantitative results, we use the best-performing models to carry out a series of qualitative analyses of the MWE translations from Galician to EN and ES. This analysis also allows us to gain a finer-grained understanding of how fine-tuning affects the translation of noun–adjective MWEs with different semantic properties.

MWE (GL)	Base model output	Fine-tuned model output
zona húmida	wet zone	wetland
liña férrea	iron line	railway line
montaña rusa	mountain of Russian	rollercoaster
réxime franquista	Franco regime	Francoist regime
banda estadounidense	band	American band

Table 4: Examples of MWE translations improved after fine-tuning (GL→EN)

To do so, a language expert manually classified each MWE translation in the test sets as *Correct*, *Variant*, and *Incorrect*.[10] An instance was labeled as *Correct* when the MWE was translated with the same expression as the reference. *Variant* labels were assigned to meaning-preserving alternatives that differed lexically or stylistically from the reference. *Incorrect* cases were subsequently classified as *i)* inadequate literal translations, *ii)* wrong sense disambiguation, *iii)* partial or full omission of the MWE, and *iv)* other errors (e.g., spelling errors, untranslated source-language words, or otherwise unintelligible outputs).

MWE translation accuracy: The first analysis examined the influence of the reference sentences on the quantitative results by computing accuracy under two evaluation criteria: (i) only *Correct* instances are counted as correct, and (ii) both *Correct* and *Variant* instances are treated as adequate translations: For GL-ES, the results were very similar (accuracies around 0.97 in every case), while for GL-EN the results increased from 0.62 (original and fine-tuned models in the first scenario) to 0.77 (both models in the second evaluation). These results reinforce the need for qualitative evaluation of MT systems.

The manual review of all translation allowed us to observe MWE translation differences between the original and the fine-tuned models. In this regard, in some GL-EN cases, FT enabled the model to correctly translate several MWEs that were previously mistranslated by the base model. These include cases involving non-compositional meaning or strong sense shifts, such as *zona húmida* (*wet zone→wetland*) or the idiomatic *montaña rusa* (*Russian mountain→roller coaster*). Additional improvements were observed in cases where the fine-tuned model selected a more idiomatic or semantically precise variant (e.g. *Franco regime→Francoist regime*; *harsh blow→hard blow*). Table 4 highlights cases where fine-tuning successfully corrected MWE translations that were previously incorrect in the base model. At the same time, a small number of MWEs that were correctly translated by the base model became incorrect after fine-tuning. These regressions typically involved increased literalness or minor lexical degradation, such as *main fountain* being rendered as *main source*, or reduced lexical realization (e.g. *Olympic gold medals → Olympic golds*). Although these cases are relatively few, they illustrate the delicate balance between specialization and generalization in continual learning. Table 5 presents representative examples of observed error types from the GL→EN experiments.

Despite the high performance of the original GL-ES model, fine-tuning still leads to targeted improvements. Several MWEs previously mistranslated by the base model are correctly handled after adaptation (e.g., *yacimiento→yacimiento arqueológico*, *piedra filosofalda→piedra filosofal*, *pescado azul→pez azul*). Improvements also occur in stylistic variants (e.g., *pariente cercano→pariente próximo*). Some representative GL-ES error examples can be seen in Table 6.

Error types: Table 7 shows the distribution of the translation error types of the MWEs. In every case, wrong sense disambiguation and wrong literal translations account for the majority of errors, confirming the difficulty of resolving meaning for non-compositional MWEs. While in GL-ES there are no significant differences between the base and fine-tuned models, in GL-EN, the fine-tuned models seem to perform better semantic disambiguation, but also produce more (inadequate) literal translations.

Frequency effects: We take advantage of the MWE frequency classification included in the original dataset to observe if it has any effect in the quality of the translation. The results (Table 8) indicate that in most cases high-frequency MWEs

[10] In a first step, we included the category *Doubt* reserved for borderline cases, which were solved before the final analysis.

Error type	MWE (GL)	System output (EN)	Reference (EN)
Literal translation	xénero musical	musical gender	music genre
Wrong sense disamb.	montaña rusa	Russian mountain	roller coaster
Omission	terra firme	land	dry land
Others	fosa común	mass mass	mass grave

Table 5: Representative error types in GL→EN MWE translation

Error type	MWE (GL)	System output (ES)	Reference (ES)
Literal translation	letra grosa	letra gruesa	letra negrita
Wrong sense disambiguation	peixe azul	pescado azul	pez azul
Spelling error	cambio climático	cambio climatico	cambio climático
Untranslated words	xogador novo	jugadoras novas	jugadoras jóvenes
Others	pedra filosofal	piedra filosofalda	piedra filosofal

Table 6: Representative error types in GL→ES MWE translation

Error type	GL-EN		GL-ES	
	Base	FT	Base	FT
Literal trans.	0.41	0.44	0.38	0.38
Wrong sense dis.	0.53	0.48	0.25	0.24
Omission	0.05	0.07	0.12	0.00
Others	0.01	0.01	0.25	0.38

Table 7: Percentage of MWE translation error types.

are more easily translated than low frequency ones. While in GL–ES the translation accuracy of frequent MWEs is, on average, less than 1% higher than that of low-frequency MWEs, in GN–EN the gap is much more pronounced, reaching almost 10% on average.

Pair	Frequency	Base	FT
GL-EN	High	81.53%	82.16%
	Low	72.73%	71.51%
	Overall	77.02%	76.71%
GL-ES	High	98.09%	97.45%
	Low	96.97%	97.57%
	Overall	97.51%	97.51%

Table 8: Percentage accuracy of MWE translation by frequency.

Compositionality effects: However, because MWEs can display different contextual senses independently of their frequency, we additionally examine translation performance across token-level compositionality classes. The results in Table 9 show that, namely in GL-EN (as in the previous analyses), idiomatic expressions remain challenging for NMT, both for generic systems and for FT models trained with examples of the target MWEs, yielding the lowest average accuracy (54%). By contrast, performance is substantially higher for partial MWEs (77%) and fully compositional MWEs (82.5%), confirming a clear correlation between degree of compositionality and translation accuracy.

Pair	Comp. class	Base	FT
GL-EN	Idiomatic	56.25%	52.08%
	Partial	76.47%	77.65%
	Compositional	82.54%	82.54%
GL-ES	Idiomatic	93.75%	95.83%
	Partial	96.47%	95.29%
	Compositional	98.94%	98.94%

Table 9: Percentage accuracy of MWE translation by compositionality class.

8 Conclusions and further work

This study explored the effects of targeted fine-tuning on MWE translation in Galician-English and Galician-Spanish, systematically analyzing the impact of both the number of learning steps and the size of MWE corpora. We release two new manually created datasets composed of pairs of 1,858 sentences with detailed annotation of the MWEs.

The results of systematic evaluations suggest that moderate FT (around 600–1000 steps) generally provides the best balance between general translation quality and MWE-specific improvements, but this depends on the language pair under evaluation. In this regard, for similar varieties where the semantics of MWEs may be less divergent (Galician and Spanish vs. Galician and English), the performance

of the original models is competitive.

Regarding the datasets, increasing the size of the MWE fine-tuning corpus does not always guarantee improvements. GL→EN models show limited sensitivity, and ES→GL only exhibits modest gains, indicating again that language-specific characteristics influence how effectively additional MWE data can be leveraged.

A qualitative analysis allowed us to observe that high-frequency MWEs are generally easier to translate, and that idiomatic ones are harder to translate than compositional MWEs, indicating that non-compositional meaning remains difficult to capture.

Building on the insights from this study, several directions for future research are suggested: First, to explore adaptive FT methods that adjust the number of steps or learning rate based on model performance on both general and MWE-specific validation sets. Second, to experiment with synthetic or automatically extracted MWE corpora to increase coverage of rare and idiomatic expressions, and assess their impact on model robustness and generalization. Finally, we plan to extend the proposed FT approach to types of MWEs (e.g., verb–object constructions) and semantic phenomena (e.g., various types of lexical ambiguity).

Limitations

This study presents a controlled analysis of targeted fine-tuning for MWE translation. However, several limitations should be acknowledged.

First, the size of the manually curated MWE datasets is relatively small. Although the corpora were carefully constructed and translated to ensure high linguistic quality, the limited number of sentences per MWE sense restricts the statistical power of the results and may limit their generalizability to broader domains or unseen MWE types.

Second, the experiments focus on a restricted set of MWE categories, noun-adjective pairs. Other types of MWEs are not covered, and as a result, the conclusions may not directly transfer to all forms of non-compositional language.

Third, the analysis is limited to four translation directions involving Galician, English, and Spanish, with a primary focus on translations from Galician. While these directions offer valuable insights into low-resource and asymmetric translation settings, the findings may not generalize to languages with different typological properties. Furthermore, the original MWEs were compiled in only one of the languages (Galician).

Finally, fine-tuning was performed under controlled experimental conditions using fixed architectures and hyperparameters. Alternative model architectures, parameter-efficient fine-tuning methods, or dynamic training strategies were not explored and could lead to different trade-offs between MWE performance and general translation quality.

Acknowledgments

This paper was funded by MCIU/AEI/10.13039/501100011033 (grants with references PID 2021-128811OA-I00, PID2024-161928OB-I00, CNS2024-154902, and AIA2025-163322-C62), by the Galician Government (ED431G 2023/04 and ED431B 2025/16), and by the *Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia* - Funded by EU — NextGenerationEU within the framework of the project *Desarrollo Modelos ALIA*.

References

Duygu Ataman, Alexandra Birch, Nizar Habash, Marcello Federico, Philipp Koehn, and Kyunghyun Cho. 2025. Machine Translation in the Era of Large Language Models:A Survey of Historical and Emerging Problems. *Information*, 16(9):723. Publisher: Multidisciplinary Digital Publishing Institute.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*.

Saúl Buján, Daniel Bardanca Outeiriño, Pablo Gamallo, Iria de Dios Flores, and José Ramón Pichel Campos. 2025. Machine translation for low-resource languages: Performance trade-offs between seq2seq and generative approaches. *Procesamiento del Lenguaje Natural*, 75:297–315.

Laura Castro, Anna Temerko, and Marcos Garcia. 2025. Compositionality and Ambiguity in Multiword Expressions: A Dataset for the Evaluation of Language Models in Galician. In *Progress in Artificial Intelligence*, pages 228–240, Cham. Springer Nature Switzerland.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. Can transformer be too compositional? analysing idiom processing in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.

Iria de Dios-Flores, Silvia Paniagua Suárez, Cristina Carbajal Pérez, Daniel Bardanca Outeiriño, Marcos Garcia, and Pablo Gamallo. 2024. CorpusNÓS: A massive Galician corpus for training large language models. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 593–599, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Ona de Gibert, Dayyán O'Brien, Dušan Variš, and Jörg Tiedemann. 2025. Mind the gap: Diverse NMT models for resource-constrained environments. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 209–216, Tallinn, Estonia. University of Tartu Library.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Shuhao Gu and Yang Feng. 2020. Investigating catastrophic forgetting during continual training for neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4315–4326, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

Lifeng Han, Gareth Jones, and Alan Smeaton. 2020a. AlphaMWE: Construction of multilingual parallel corpora with MWE annotations. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 44–57, online. Association for Computational Linguistics.

Lifeng Han, Gareth Jones, and Alan Smeaton. 2020b. MultiMWE: Building a multi-lingual multi-word expression (MWE) parallel corpora. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2970–2979, Marseille, France. European Language Resources Association.

Lifeng Han, Najet Hadj Mohamed, Malak Rassem, Gareth Jones, Alan Smeaton, and Goran Nenadic. 2025. Towards a resource for multilingual lexicons: an mt assisted and human-in-the-loop multilingual parallel corpus with multi-word expression annotation. *Language Resources and Evaluation*. Forthcoming.

Wei He, Tiago Kramer Vieira, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2025. Investigating idiomaticity in word representations. *Computational Linguistics*, 51:505–555.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, Washington, USA. Association for Computational Linguistics.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Valia Kordoni and Iliana Simova. 2014. Multiword Expressions in Machine Translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1208–1211, Reykjavik, Iceland. European Language Resources Association (ELRA).

Linfeng Liu, Saptarshi Ghosh, and Tianyu Jiang. 2025. Evaluating the impact of verbal multiword expressions on machine translation. *Preprint*, arXiv:2508.17458.

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation - Advances in Research and Theory*, 24(C):109–165.

Daniel Bardanca Outeirinho, Pablo Gamallo Otero, Iria de Dios-Flores, and José Ramom Pichel Campos. 2024. Exploring the effects of vocabulary size in neural machine translation: Galician as a target language. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 600–604, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Carlos Ramisch. 2023. Multiword expressions in computational linguistics. Habilitation à diriger des recherches. Aix Marseille Université (AMU).

Vikas Raunak, Roman Grundkiewicz, and Marcin Junczys-Dowmunt. 2024. On instruction-finetuning neural machine translation models. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1155–1166, Miami, Florida, USA. Association for Computational Linguistics.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Aleix Sant, Daniel Bardanca, José Ramom Pichel Campos, Francesca De Luca Fornaciari, Carlos Escolano, Javier Garcia Gilabert, Pablo Gamallo, Audrey Mash, Xixian Liao, and Maite Melero. 2024. Training and

Fine-Tuning NMT Models for Low-Resource Languages Using Apertium-Based Synthetic Corpora. In *Proceedings of the Ninth Conference on Machine Translation*, pages 925–933, Miami, Florida, USA. Association for Computational Linguistics.

Danielle Saunders and Steve DeNeefe. 2024. Domain adapted machine translation: What does catastrophic forgetting forget and why? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12660–12671, Miami, Florida, USA. Association for Computational Linguistics.

Sarubi Thillainathan, Surangika Ranathunga, and Sanath Jayasena. 2021. Fine-Tuning Self-Supervised Multilingual Sequence-To-Sequence Models for Extremely Low-Resource NMT. In *2021 Moratuwa Engineering Research Conference (MERCon)*, pages 432–437. ISSN: 2691-364X.

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

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

Andrea Zaninello and Alexandra Birch. 2020. Multiword Expression aware Neural Machine Translation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3816–3825, Marseille, France. European Language Resources Association.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

Appendix

A Training and Fine-Tuning Hyperparameters

Table 10 summarizes the architecture and training hyperparameters used for both base training and continual fine-tuning of the Seq2Seq Transformer models. All models were trained using the OpenNMT framework, using the same configuration across all translation directions.

Parameter	**Value**
Model architecture	
Encoder layers	12
Decoder layers	12
Attention heads	16
Hidden size	512
Feed-forward size	2048
Dropout	0.1
Label smoothing	0.1
Position encoding	Enabled
Training and fine-tuning	
Optimizer	Adam
β_2	0.998
Learning rate	2
Warmup steps	8,000
Batch size	4,096 tokens
Gradient accumulation	4 steps
Maximum sequence length	150 tokens
Training precision	FP32

Table 10: Model architecture and training hyperparameters