Improving Machine Translation of Idioms: A Spanish–Galician Parallel Dataset and Synthetic Augmentation Approach

Lúa Santamaría Montesinos Saúl Buján Daniel Bardanca Pablo Gamallo
Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS)
Universidade de Santiago de Compostela
{lua.santamaria.montesinos,saul.bujan,danielbardanca.outeirino,pablo.gamallo}@usc.gal

Abstract

Idiomatic expressions pose a significant challenge for neural machine translation, encompassing both traditional sequence-to-sequence models and large language models (LLMs). This paper presents a systematic approach to improve idiom translation between Spanish and Galician. First, we build a high-quality parallel dataset of idioms manually aligned across both languages. Then, we automatically extend this dataset into a large synthetic parallel corpus using LLMs, following a strategy that prioritizes the most frequent idioms observed in authentic corpora. This augmented dataset is used to retrain a seq2seq translation model. We evaluate the resulting system and compare it to both the original model without idiom data and to state-of-the-art LLM-based translators, such as SalamandraTA. Results show that the translation of idioms improves significantly after the training, alongside a slight boost in the model's overall performance.

1 Introduction

Idioms are one of the most persistent sources of errors in neural machine translation (NMT). While modern seq2seq architectures and LLMs achieve remarkable fluency and adequacy for literal language, they often fail to capture the non-compositional meaning of idiomatic expressions. For example, a literal translation of the Spanish idiom *"estar en las nubes"* ("to be in the clouds") would misrepresent its figurative meaning ("to be distracted") when translated word-for-word into Galician (a Portuguese variety), whose true translation would be *"estar nas verzas"* instead of the literal *"estar nas nubes"*.

This work addresses this limitation by focusing on idiom translation between two closely related Romance languages: Spanish and Galician. Despite their linguistic similarity, idiomatic expressions often diverge in subtle ways, making them a valuable test case for evaluating the semantic depth and pragmatic awareness of translation models.

Our main contributions are as follows:

1. We create a manually verified parallel dataset of Spanish–Galician idioms.

2. Using this dataset, we generate a large-scale synthetic parallel corpus via LLMs, after selecting the most frequent idioms from authentic corpora and manually augmenting them.

3. We train two seq2seq translation models, one without and one augmented with this idiom-enriched corpus.

4. We evaluate the models in two ways: first, we check whether the model retrained with idioms improves on the base model, and then we compare them against the LLM-based translation system SalamandraTA (Garcia Gilabert et al., 2025).

Both the parallel data and the seq2seq models will be publicly released and uploaded to HuggingFace. This paper is organized as follows. Section 2 reviews related work on idiomatic and low-resource MT approaches. Section 3 provides the details of the dataset construction. Section 4 outlines the training setups for our model and its evaluation and reports the results. Finally, we conclude with some notes on future work and limitations in Section 5.

2 Related Work

This section reviews previous work on idioms in machine translation and approaches for low-resource language pairs.

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 980–987

2.1 Idioms and Multiword Expressions in Machine Translation

Idiomatic expressions present a well-known challenge for NMT systems due to their non-compositional semantics (Liu et al., 2023). Previous work has approached this problem through specialized datasets, retrieval methods, and architectural adaptations. In this context, creating high-quality parallel datasets is a foundational step in idiom-aware MT. The introduction of benchmark datasets for idiom translation (Fadaee et al., 2018) highlighted how most idioms in training data appear literally, requiring focused annotation.

Researchers have also developed specialized techniques for MT. The proposal of retrieval augmentation combined with loss weighting to improve idiom translation (Liu et al., 2023) has identified potential idioms during translation and retrieved similar examples to guide generation. More recently, an investigation on how verbal multiword expressions affect MT quality (Liu et al., 2025) found that transformer-based models still struggle with these constructions despite architectural advances.

2.2 MT for Low-Resource Languages

While approaches like back-translation (Sennrich et al., 2016) and data augmentation (Pichel Campos et al., 2009) aim to mitigate data scarcity in low-resource NMT, the need for high-quality curated data remains paramount (Goyle et al., 2023). Therefore, the use of synthetic parallel data has become a common thread in this area. Recent approaches have leveraged LLMs for higher-quality generation (Zhu et al., 2023), showing that LLMs can produce more fluent and contextually appropriate synthetic translations than previous methods, particularly for low-resource language pairs where human-curated parallel data is scarce (Caswell et al., 2021).

Being a low-resource language, Galician shares these challenges. While Galician-specific research has been scarce, we have recently conducted a comprehensive comparison between traditional seq2seq models and generative LLMs for low-resource MT (Buján et al., 2025), suggesting that while LLMs offer strong zero-shot capabilities, fine-tuned seq2seq models can achieve competitive results with less computational overhead.

3 Dataset Construction

This section describes the process of building, filtering, and augmenting our parallel idiom dataset for Spanish-Galician.

3.1 Parallel Idiom Dataset

We compiled an initial set of idioms in Spanish and Galician from existing dictionaries, phraseological databases, and online newspapers. Each idiom pair was manually verified for equivalence in meaning and pragmatic use by native speakers. The resulting dataset of roughly 250 idiom pairs was later refined by a dual-phase process involving frequency analysis followed by manual selection to identify the most relevant items.

Importantly, a crucial methodological consideration is that we do not assume a one-to-one mapping between idioms across languages. Idiomatic equivalence is known to be highly variable depending on register, context, or geographical variety. Accordingly, our dataset does not aim to capture all possible variants of each idiom; instead, we prioritized the most frequent variant based on corpus evidence and native-speaker consensus. Expanding the dataset to include synonymous expressions and diatopic variants remains as future work.

3.2 Frequency-Based Selection

The first step of the selection process aimed to maximize the relevance of our data despite the constraints of Galician's low-resource language status. Consequently, we extracted the most frequent idioms from our original set consulting large Spanish and Galician corpora, such as *CORPES XXI* [1] and *CORGA* [2]. We also incorporated smaller corpora of transcribed oral data, both authentic spontaneous speech and scripted dialogue. Specifically, we included the *ESLORA* (Vázquez Rozas and Blanco, 2023) corpus (for Spanish) and the *TED2020* (Reimers and Gurevych, 2020) and *OpenSubtitles2016* (Lison and Tiedemann, 2016) corpora (for Spanish and Galician). Our frequency measurements were obtained through a combination of manual and automated methods: manual searches were conducted within the native interfaces of the larger corpora to account for all potential variants, while for the smaller corpora, we

[1] https://www.rae.es/corpes/
[2] https://corpus.cirp.gal/corga/

implemented an automated approach using regular expressions to capture all possible instances of each idiom.

The initial frequency analysis revealed generally low ocurrence counts for idioms, even for those considered culturally significant (for instance, the Galician idiom *"outra vaca no millo"* —"another cow in the cornfield"— was attested only six times in the major *CORGA* corpus). To compensate for this limited representativeness in the available corpora, we implemented an additional manual selection phase.

3.3 Manual Augmentation and Expert Validation

This second phase aimed to capture idioms that are widely recognized by speakers but are infrequent in written data, using a human-centered approach. Accordingly, a group of linguists used their native-speaker expertise to identify culturally salient and frequent expressions from the original list that lacked strong corpus attestation. The final idiom set comprised 66 bilingual pairs.

3.4 Synthetic Parallel Corpus Generation

We employed *Deepseek V-2* [3] to generate parallel sentences containing each idiom in context following a two-step prompting strategy:

First, we iteratively applied a meticulously designed prompt, detailed in Appendix A, to generate Galician sentences. For each iteration, the prompt provided the LLM with a single bilingual idiom pair and its meaning, then instructed it to create 100 original sentences, drawing on the model's capacity to emulate a Galician phraseology expert. The prompt specified certain limitations to ensure syntactic and semantic diversity and provided three to five usage examples for each idiom, which were either created manually or extracted from a corpus (*CORGA* or *CORPES XXI*). To ensure grammatical correctness, it explicitly listed and explained a series of recurrent errors to prevent their occurrence in the new sentences.

Second, we used a separate, targeted prompt —provided in Appendix B— to translate the sentences into Spanish. This phase allowed us to manually correct errors, explicitly enforce standard Spanish orthography, and give particular emphasis to specific rules that are frequent sources of error when translating from Galician.

[3] https://www.deepseek.com/

Thus, the model produced Galician and Spanish examples of the uses of each idiom, ensuring contextual diversity. All generated examples were manually verified to assess quality, correctness, and fidelity. The resulting parallel idiom dataset is publicly available [4].

4 Experiments

This section outlines the training setup, evaluation metrics, and comparative results of our idiom-augmented translation models.

4.1 Model Training

We trained two transformer-based seq2seq models for the ES–GL language pair on a single A100 GPU. Both models were optimized with AdamW using the parameters $\beta_1 = 0.9$, $\beta_2 = 0.998$, and $\epsilon = 10^{-8}$, and employed the default learning rate of OpenNMT-py 3.2[5].

The models share the same architecture: 12 encoder and 12 decoder layers, 16 attention heads per layer, a hidden size of 512, and a vocabulary of 30K tokens, the most adequate size according to previous work (Bardanca et al., 2024). Training ran for 20 epochs with batches of 2048 sentences (maximum sequence length of 150 tokens). After training, the models were converted to the CTranslate2 format[6] to reduce memory footprint and improve inference efficiency. Each model contains approximately 120M parameters, resulting in a final size of around 500 MB.

Both systems were trained on bilingual corpora of approximately 70 million sentence ES-GL pairs (de Dios-Flores et al., 2024). Their only difference is the inclusion of the additional idiom-specific dataset: one model was trained without idioms, whereas the second incorporated them during full training. We opted not to perform a dedicated fine-tuning step solely on the idiom dataset, since preliminary experiments showed that this procedure, while improving idiom translation, tends to degrade overall translation quality. Given that these are small models, retraining, despite being more expensive than fine-tuning, is a task that can be completed in a reasonable amount of time: $\sim$ 40 hours using $\sim$ 11G of GPU.

[4] https://huggingface.co/datasets/proxectonos/corpus_paralelo_idioms

[5] https://pypi.org/project/OpenNMT-py/

[6] https://pypi.org/project/ctranslate2/

4.2 Evaluation

We evaluated three systems:

Original Seq2Seq: Trained on standard parallel data only.

Idiom-augmented Seq2Seq: Retrained with the synthetic idiom corpus.

LLM: The state-of-the-art LLM translator *SalamandraTA-7b-instruct* (Garcia Gilabert et al., 2025), fine-tuned with parallel datasets that include the same ES-GL parallel corpora as our seq2seq original model. It is a decoder-only transformer with 7 billion parameters, $\sim$ 60 times larger than our seq2seq models.

We conducted our primary evaluation using BLEU (Papineni et al., 2002) and COMET (Rei et al., 2020). Alternative string-overlap metrics were excluded as their scores showed strong correlation with BLEU (Buján et al., 2025) and thus provided redundant information. The evaluation proceeded along two dimensions: overall translation quality with the two above-mentioned metrics and targeted idiom accuracy with a specific metric developed specifically for this purpose. Furthermore, a qualitative manual evaluation was also conducted on the idiom-specific dataset comparing the original and idiom-augmented seq2seqs.

To assess overall quality, we evaluated the three systems on three ES-GL gold standards we previously developed (Buján et al., 2025), in addition to established multilingual datasets (Flores (Goyal et al., 2022), Tatoeba (Tiedemann, 2020), and TaCon (de Gibert et al., 2022)). Moreover, we also created two specific golden datasets suited for measuring idiom translations: the test partition of the dataset described in Section 3 and the Aya dataset, based on the Spanish version of the Aya Red-Teaming[7] dataset, which we have corrected and translated into Galician and Portuguese, suitable for evaluating idiomatic expressions as it contains informal language. This resource, along with our other materials, has been released under an open-source license[8].

4.3 Results and Discussion

The BLEU and COMET scores displayed in Table 1 confirm our core hypothesis: explicitly augmenting training data with parallel idiom examples significantly improves an NMT system's ability to translate idiomatic expressions correctly. This result was supported by the human comparison on the idiom dataset translation, showing that the augmented model produced a preferable translation compared to the original one in 90.3% of cases, while the opposite was true in only 0.9% of instances, with the remaining cases resulting in equivalent translations. The high accuracy on *seen* idioms demonstrates the effectiveness of our data generation and retraining strategy. However, subsequent testing reveals this does not hold for *unseen* idioms, underscoring the non-compositional nature of idioms and how, without explicit examples, models struggle to infer their correct translation. This points to the need for continued expansion of idiom lexicons and techniques for better cross-idiom generalization.

While the LLM is supposed to have inherent knowledge of idioms, our specialized, augmented seq2seq model outperforms it on most datasets (especially on that containing idioms) with greater efficiency and a smaller environmental footprint.

5 Conclusions and Future Work

This paper addresses key challenges in machine translation: translating multiword expressions and operating in low-resource language settings. After reviewing the state-of-the-art literature, we presented our main contribution: a high-quality, parallel idiom dataset for Spanish–Galician. We then detailed a multi-step construction process, including the use of synthetic data.

We evaluated our resource by training and comparing two transformer-based seq2seq models on a large general-domain corpus, one augmented with our idiom dataset. The results demonstrate that targeted idiom data not only substantially improves the translation of idiomatic expressions but also enhances overall translation quality, as confirmed by the evaluation process.

Nevertheless, we acknowledge limitations in the dataset's scale and the models' capacity. The dataset, while curated, is limited to 66 core idiom pairs. Furthermore, the 120M-parameter seq2seq models, while efficient, are not the most powerful architectures available. We suggest that future work could expand the idiom dataset and evaluate its impact on larger, pre-trained models and large language fine-tuned for translation, to further ex-

[7] https://huggingface.co/datasets/CohereLabs/aya_redteaming

[8] https://huggingface.co/datasets/proxectonos/aya_nos

	BLEU			COMET		
Dataset	**Original**	**Augmented**	**LLM**	**Original**	**Augmented**	**LLM**
gold1	79.5	**79.7**	63.2	0.894	**0.895**	0.873
gold2	**44.0**	43.7	43.5	**0.877**	0.876	0.869
test-suite	75.2	**76.2**	56.1	0.921	**0.925**	0.886
flores	21.8	21.9	**28.1**	0.842	0.842	**0.859**
tatoeba	63.7	**63.8**	54.1	**0.898**	**0.898**	0.880
taCon	82.8	**83.6**	76.2	0.953	**0.954**	0.947
aya	66.4	**66.8**	65.7	0.907	**0.909**	**0.909**
idioms	70.4	**90.2**	51.9	0.811	**0.912**	0.798

Table 1: Comparison of BLEU and COMET scores across datasets for the three models.

plore the trade-offs between performance, data efficiency, and computational cost.

Acknowledgements

This paper was funded by MCIU/AEI (grants with references PID 2021-128811OA-I00, PID2024-161928OB-I00, and AIA2025-163322-C62), by the Galician Government (Research Center of Galicia accreditation 2024-2027 ED431G-2023/04 and GPCE 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

Daniel Bardanca, Pablo Gamallo, 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 Linguistics.

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

Isaac Caswell, Julia Kreutzer, Lisa Wang, Ahsan Wahab, et al. 2021. Quality at a glance: An audit of web-crawled multilingual datasets. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9146–9162.

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 Linguistics.

Ona de Gibert, Ksenia Kharitonova, Blanca Calvo Figueras, Jordi Armengol-Estapé, and Maite Melero. 2022. Quality versus quantity: Building Catalan-English MT resources. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 59–69, Marseille, France. European Language Resources Association.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2018. Examining the tip of the iceberg: A data set for idiom translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 925–929, Miyazaki, Japan. European Language Resources Association (ELRA).

Javier Garcia Gilabert, Xixian Liao, Severino Da Dalt, Ella Bohman, Audrey Mash, Francesca De Luca Fornaciari, Irene Baucells, Joan Llop, Miguel Claramunt, Carlos Escolano, and Maite Melero. 2025. From salamandra to salamandrata: Bsc submission for wmt25 general machine translation shared task. In *Proc. of WMT*, pages 614–637.

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.

Vakul Goyle, Parvathy Krishnaswamy, Kannan Girija Ravikumar, Utsa Chattopadhyay, and Kartikay

Goyle. 2023. Neural machine translation for low resource languages. *arXiv preprint*.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Emmy Liu, Aditi Chaudhary, and Graham Neubig. 2023. Crossing the threshold: Idiomatic machine translation through retrieval augmentation and loss weighting. In *Proc. of EMNLP*, pages 15095–15111.

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

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

José Ramóm Pichel Campos, Pablo Martín, Óscar Senra, Pablo Gamallo, and Alberto García. 2009. Carvalho: Un sistema de traducción estadística inglés-galego construído a partir del corpus paralelo inglés-portugués EuroParl. *Procesamiento del Lenguaje Natural*, 43:379–381.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation*, pages 371–381, Berlin, Germany. Association for Computational Linguistics.

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.

Victoria Vázquez Rozas and Marta Blanco. 2023. El corpus ESLORA de español oral: un recurso para la investigación en ELE. *TEISEL. Tecnologías para la investigación en segundas lenguas*, 2.

Wenhao Zhu, Delin Pang, Liang Wang, Hangbo Chen, et al. 2023. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14266–14281.

Appendix A.1 – Galician Generation Prompt (Original)

Imaxina que es un catedrático en gramática e fraseoloxía galegas. Tendo isto en conta, tes que xerar 100 oracións en galego usando o idiom "idiom_gal" (significado "idiom_gal_sdo") e, posteriormente, terás que darme a súa tradución ao castelán empregando o idiom equivalente "idiom_es" (significado "idioms_es_sdo").

Sigue detalladamente e con precisión todas e cada unha das seguintes indicacións:

- É imprescindible que as oracións que xeres non teñan unha estructura sintáctica similar ou equivalente entre elas; é dicir, deben ser sintáctica e semánticamente diversas e variadas, ademais de ser complexas.
- É imprescindible que parezan creadas por humanos e non xeradas por un LLM.
- É imprescindible que non conteñan erros gramaticais nin ortográficos de ningún tipo.
- Cando xeres frases en galego, non uses nunca o portugués nin o español.

Sobre a redacción:

- Non uses nunca "¡" e/ou "¿" ao comezo das frases interrogativas e/ou exclamativas en galego (é dicir, en galego o correcto sería "onde vai?" e non "¿onde vai?").
- Non poñas "." despois de "?" e/ou "!" en galego (nin en castelán tampouco cando fagas a tradución). Cando un enunciado remata con "?" e/ou "!", comeza sempre con maiúsculas despois de devanditos signos.
- Non acentúes nunca os pronomes interrogativos nin exclamativos en galego (por exemplo, o correcto é "que fixeches?" e non "qué fixeches?").
- Evita as estructuras pasivas tanto en galego coma en castelán cando traduzas as frases (non son tan frecuentes e naturais coma noutras linguas).

- Intenta manter maioritariamente o suxeito antes do verbo na sintaxe das túas frases.
- Presta especial atención aos pronomes en galego e sitúaos correctamente na oración (atención, sobre todo, ao par "te"–"che" e á colocación dos clíticos).

A continuación, vouche proporcionar unhas oracións de exemplo co idiom que tes que usar en contexto. Engade estas frases dentro dese conxunto de 100 e inspírate nelas en calquera aspecto relevante para a xeración das restantes:

- Engadir as 3–5 oracións de exemplo usando "idiom_gal" en contexto.

Último paso: comeza xerando exclusivamente as 100 frases en galego (máis adiante, cando cho diga eu cun prompt distinto, terás que traducilas ao español). Dame todas as oracións xuntas, separadas unha da outra por saltos de liña.

Appendix A.2 – Galician Generation Prompt (English)

Imagine you are a professor of Galician grammar and phraseology. Taking this into account, you must generate 100 sentences in Galician using the idiom "idiom_gal" (meaning "idiom_gal_mean") and, subsequently, you will have to give me their translation into Spanish using the equivalent idiom "idiom_es" (meaning "idioms_es_mean"). Follow in detail and with precision each and every one of the following instructions:

- It is essential that the sentences you generate do not have a similar or equivalent syntactic structure among them; that is, they must be syntactically and semantically diverse and varied, in addition to being complex.
- It is essential that they appear to be created by humans and not generated by an LLM.
- It is essential that they contain no grammatical or spelling errors of any kind.
- When generating sentences in Galician, never use Portuguese or Spanish.

Regarding writing:

- Never use "¡" and/or "¿" at the beginning of interrogative and/or exclamatory sentences in Galician (that is, in Galician the correct form would be "onde vai?" and not "¿onde vai?").
- Do not put a "." after "?" and/or "!" in Galician (nor in Spanish when you do the translation). When a sentence ends with "?" or "!", always start with a capital letter after these marks.
- Never accent interrogative or exclamatory pronouns in Galician (for example, the correct form is "que fixeches?" and not "qué fixeches?").
- Avoid passive structures both in Galician and in Spanish when translating the sentences (they are not as frequent and natural as in other languages).
- Try to keep the subject mostly before the verb in the syntax of your sentences.
- Pay special attention to pronouns in Galician and place them correctly in the sentence (especially, the "te"–"che" pair and the placement of clitics).

Next, I will provide you with some example sentences including the idiom you must use in context. Add these sentences within that set of 100 and be inspired by them in any aspect for the generation of the remaining ones:

- Add 3–5 example sentences using "idiom_gal" in context.

Final step: start by generating exclusively the 100 sentences in Galician (later, when I tell you with a different prompt, you will have to translate them into Spanish). Give me all the sentences, separated by line breaks.

Appendix B.1 – Spanish Translation Prompt (Original)

Has tenido algunos errores, pero te devuelvo a continuación las oraciones corregidas. Ahora, traduce dichas frases al español de modo fiel y perfectamente correcto con respecto a la gramática española (en concreto, tienes que traducir "idiom_gal" por su equivalente "idiom_es"). Además, quiero que añadas "¿" o "¡" si hay alguna oración interrogativa o exclamativa y que acentúes los

pronombres interrogativos y/o exclamativos que aparezcan (es decir, “¡qué guapo!” y no “que guapo!”).

Quiero que me devuelvas una oración después de otra, separadas por saltos de línea. Aquí tienes las oraciones correctas en gallego:

Appendix B.2 – Spanish Translation Prompt (English)

You have made some errors, but I am providing you with corrected sentences below. Now, translate these sentences into Spanish in a faithful and grammatically perfect manner (specifically, you must translate “idiom_gal” with its equivalent “idiom_es”). Additionally, I want you to add “¿” or “¡” if there is any interrogative or exclamatory sentence and to accent any interrogative and/or exclamatory pronouns that appear (that is, “¡qué guapo!” and not “que guapo!”).

I want you to return one sentence after another, separated by line breaks. Here are the correct sentences in Galician: