Procesamiento del Lenguaje Natural, Revista nº 75, septiembre de 2025, pp. 297-315 recibido 06-04-2025 revisado 11-05-2025 aceptado 17-05-2025

Machine Translation for Low-Resource Languages: Performance Trade-offs Between Seq2Seq and Generative Approaches

Traducción Automática para Lenguas de Bajos Recursos: Comparativa de Rendimiento entre Modelos Seq2Seq y Generativos

Saúl Buján,[1] **Daniel Bardanca,**[1] **Pablo Gamallo,**[1]
Iria de-Dios-Flores,[2] **José Ramom Pichel**[1]
[1]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS)
Universidade de Santiago de Compostela
[2]Department of Translation and Language Sciences
Universitat Pompeu Fabra
{saul.bujan,danielbardanca.outeirino,pablo.gamallo,jramon.pichel}@usc.gal
iria.dedios@upf.edu

Abstract: This study evaluates two machine translation paradigms—sequence-to-sequence (seq2seq) models and generative language models (LLMs)—for translating Spanish-Galician (closely related) and English-Galician (distant) language pairs. The seq2seq models include bilingual and multilingual models trained from scratch, and NLLB-200 as-is and fine-tuned. The generative models involve both pre-trained and fine-tuned large language models. The evaluation is conducted using quantitative metrics (BLEU and COMET) and qualitative analysis, which includes an ad hoc test suite designed to assess linguistic accuracy. Results show that fine-tuned generative models outperform seq2seq models for the distant language pair (English-Galician), whereas bilingual seq2seq models remain competitive for closely related languages (Spanish-Galician). The study highlights the trade-offs between both approaches and provides insights into optimizing translation strategies for low-resource languages like Galician.
Keywords: low-resource languages, qualitative evaluation, sequence-to-sequence models, generative models.

Resumen: Este estudio evalúa dos paradigmas de la traducción automática—modelos sequence-to-sequence (seq2seq) y modelos del lenguaje generativos (LLMs)—para traducir los pares de lenguas español-gallego (cercanas) e inglés-gallego (más distantes). Los modelos seq2seq incluyen modelos bilingües y multilingües entrenados desde cero, así como el modelo NLLB-200 en su versión original y ajustada. Los modelos generativos incluyen grandes modelos del lenguaje tanto preentrenados como ajustados. La evaluación se lleva a cabo mediante métricas cuantitativas (BLEU y COMET) y un análisis cualitativo, que incluye un test suite diseñado ad hoc para valorar la precisión lingüística. Los resultados muestran que los modelos generativos ajustados superan a los modelos seq2seq para el par de lenguas distante (inglés-gallego), mientras que los modelos bilingües seq2seq siguen siendo competivos para lenguas próximas (español-gallego). Este estudio resalta los pros y contras de ambos enfoques y ofrece perspectivas para optimizar estrategias de traducción para lenguas de bajos recursos como el gallego.
Palabras clave: lenguas de bajos recursos, evaluación cualitativa, modelos sequence-to-sequence, modelos generativos.

1 Introduction

Machine translation (MT) is a vital tool for communication between different linguistic communities, with uses in global trade, public services, and intercultural exchange. While MT quality has improved significantly for major languages, low-resource languages still pose challenges Ranathunga et al. (2023). Transla-

ISSN 1135-5948 DOI 10.26342/2025-75-22

tion quality is affected not only by the scarcity of resources but also by the linguistic distance between source and target languages. Distant pairs like English-Galician involve major syntactic and lexical differences, whereas closely related pairs like Spanish-Galician present subtler issues such as false friends and fine-grained semantic shifts Puduppully et al. (2023). Additionally, the dominance of Spanish in Galicia Freixeiro Mato (2006) introduces lexical and morphosyntactic interference into Galician texts, further complicating translation.

Recent advances in neural MT include both sequence-to-sequence (seq2seq) and generative (decoder-only) models, each with distinct strengths. Seq2seq models, trained in bilingual or multilingual settings, are tailored for translation tasks and benefit from shared cross-lingual patterns Johnson et al. (2017). Generative models (also known as Large Language Models or LLMs) offer greater generalization through massive pretraining Brown et al. (2020), and can be adapted via few-shot prompting or fine-tuning. However, their relative effectiveness remains underexplored, especially in low-resource settings and in the context of language distance. Furthermore, evaluation practices often rely on automatic metrics such as BLEU and COMET, which may not fully capture linguistic nuances.

This study addresses these gaps by evaluating a diverse set of translation models on Spanish-Galician (ES–GL) and English-Galician (EN–GL) tasks, comparing both seq2seq and generative approaches:

Seq2seq: Four different types of models corresponding to four seq2seq strategies were trained: 1) Bilingual models trained with our parallel corpora, namely two bilingual Spanish-Galician (ES-GL[1], GL-ES[2]) and two English-Galician (EN-GL[3], GL-EN[4]) models. 2) A model trained with our parallel corpora for a small set of five languages, including Galician, giving rise to a controlled multilingual setting. We call this model Multilingual-5 [5]. 3) A pre-trained multilingual model with a very large number of languages as-is: NLLB-200 Koishekenov et al. (2023) (distilled version). 4) A version of NLLB-200 fine-tuned using our parallel corpora.

Generative: Two different training settings with LLMs were used: 1) In context learning with few-shot prompting, using two pre-trained LLMs: Carballo-8B[6] and Salamandra-2B.[7] 2) LLMs fine-tuned with specific parallel corpora, namely EUlang-7B and SalamandraTA-2B.[8]

Our objective is to assess the models' performance across close vs. distant language pairs, combining automatic metrics and qualitative evaluations. We also design a test suite to examine fine-grained linguistic phenomena and analyze correlations between metrics and human judgment. By examining the relationship between quantitative and qualitative metrics, we provide insights into the strengths and limitations of each approach and their suitability for different translation scenarios. The results of our experiment show that for close languages, the most basic models (bilingual seq2seq) are sufficient, while for distant languages, the fine-tuned generative LLMs perform better. Therefore, this study contributes to the ongoing debate on how to optimize MT strategies for a resource-poor language such as Galician in relation to close and distant languages in a wider range of linguistic contexts. The main contributions of this paper are: the creation of evaluation datasets through multiple methodologies; a comparative study of seq2seq models and LLMs across language distances; and the public release of all models, corpora, and evaluation data to support reproducibility and further research on under-resourced languages.

This paper is organized as follows. Section 2 reviews related work on MT approaches. Section 3 provides the details of our training corpora and evaluation datasets. Section 4 outlines the training setups for different translation models, while in Section 5, we report the results. Finally, we discuss the results in Section 6 and conclude with some notes on future work and limitations in Section 7.

[1]https://huggingface.co/proxectonos/Nos_MT-OpenNMT-es-gl.

[2]https://huggingface.co/proxectonos/Nos_MT-OpenNMT-gl-es.

[3]https://huggingface.co/proxectonos/Nos_MT-OpenNMT-en-gl.

[4]https://huggingface.co/proxectonos/Nos_MT-OpenNMT-gl-en.

[5]https://huggingface.co/proxectonos/Nos_MT-OpenNMT-multilingual.

[6]https://huggingface.co/proxectonos/Llama-3.1-Carballo.

[7]https://huggingface.co/BSC-LT/salamandra-2B.

[8]https://huggingface.co/BSC-LT/salamandraTA-2B.

2 Related work

2.1 Neural MT for Low-Resource Languages

Some studies Koehn and Knowles (2017) emphasized the limitations of traditional neural MT systems in handling languages with limited training data. Approaches such as back-translation Sennrich et al. (2016) and data augmentation Campos et al. (2009) have been proposed to mitigate these issues, demonstrating improvements in translation quality for low-resource scenarios. Galician, being a low-resource language, shares these challenges, making such studies highly relevant González et al. (2024).

Galician-specific research has been relatively sparse, although efforts have been made to develop bilingual and multilingual resources for the language. For example, Tiedemann et al., (Tiedemann (2012) have contributed significantly to the creation of parallel corpora for European languages, including Galician. Furthermore, work such as de Dios-Folores et al., (de Dios-Flores et al. (2022) on the development of tools offers an interesting first step for future work in that language.

2.2 Multilingual and Transfer Learning Approaches

Multilingual models have significantly advanced MT in low-resource scenarios by exploiting cross-lingual transfer. Zero-shot translation was introduced Johnson et al. (2017) using multilingual MT, showing the benefits of training on multiple language pairs. Among the main models, mBART Liu et al. (2020) enables fine-tuning for MT through multilingual denoising pretraining, while M2M-100 Fan et al. (2021) expanded support to 100+ languages for direct translation across 10,000+ pairs.

More recently, the Meta AI No Language Left Behind (NLLB) project Costa-jussà et al. (2022) further improved coverage to 200+ languages using curriculum learning and back-translation, with special emphasis on low-resource languages. One of the strategies we describe in the present work builds on NLLB by fine-tuning its distilled model on specific corpora for ES-GL and EN-GL translation, evaluating its potential in improving low-resource translations.

2.3 Large Language Models and Their Role in Translation

Unlike traditional MT, generative LLMs can learn translation from self-supervised training on monolingual data. A study described in Zhu et al. (Zhu et al. (2024) compares eight LLMs and finds that GPT-4 surpasses the strong supervised baseline of NLLB in 40.91% of translation directions, but still lags behind commercial translation systems like Google Translate, especially in low-resource languages. Another study Wang et al. (2023) highlights the benefits of context-aware prompts, showing LLMs' strong performance at document-level MT. Furthermore, Xu et al. (Xu et al. (2023) proposed a two-stage fine-tuning method (monolingual followed by limited parallel data) achieving results beyond NLLB. Finally, Enis and Hopkins (Enis and Hopkins (2024) examine the translation capabilities of Claude 3 Opus, a large language model released by Anthropic in March 2024, reporting it outperforms other LLMs in MT, particularly for low-resource languages translated into English. However, the application of generative LLMs to low-resource languages like Galician has not been explored. Models such as those belonging to the Carballo family Gamallo et al. (2024), which focus on regional or less-resourced languages, offer promising directions, complementing traditional MT approaches.

2.4 Evaluation of MT

A number of metrics compete in an ecosystem where there is no longer a dominant default metric Kocmi et al. (2024). The most widely used are metrics based on n-grams such as BLEU Papineni et al. (2002) and newer neural metrics such as COMET Rei et al. (2020). However, quantitative metrics often fail to capture linguistic nuances, motivating researchers to design qualitative evaluation frameworks. Work such as Isabelle et al. (Isabelle et al. (2017) emphasizes the importance of targeted test suites for assessing specific linguistic phenomena, which aligns with the qualitative evaluation methodology used in this study.

3 Textual resources

This section describes both the training corpora and the evaluation datasets specifically developed to assess EN-GL and ES-GL translation performance. We present two types of evaluation resources: gold standards for broad

automatic evaluation, and test suites targeting specific linguistic phenomena. Together they enable comprehensive quantitative and qualitative analysis. They are described below.

3.1 Training corpora

This training dataset comprises two types of corpora for the ES-GL and EN-GL language pairs: authentic and synthetic. The authentic corpus includes datasets sourced from the OPUS[9] repository Tiedemann (2009), which have been processed to remove noise and non-target language phrases using automated language detection tools. To supplement this with synthetic corpora, the Spanish-Portuguese corpus was converted into ES-GL by translating from Portuguese to Galician using the rule-based Apertium translator Forcada and Tyers (2016) enhanced with a transliteration module for spelling those words that were not recognized by the translator. Similarly, the English-Portuguese corpus was translated into EN-GL by applying the same method. Each language pair (ES-GL and EN-GL) contains approximately 70 million sentences.

3.2 Evaluation datasets

All the datasets created for this project [10], defined below, are composed of parallel translations that have been meticulously designed and rigorously reviewed. This ensures that each dataset serves as a high-quality tool for quantitative evaluations using standard automatic metrics (e.g. BLEU or COMET). We also use other well-known multilingual parallel datasets such as Flores Goyal et al. (2022), NTREX Federmann et al. (2022), Tatoeba Tiedemann (2020) and TaCon de Gibert Bonet et al. (2022). Furthermore, the test suites incorporated within these datasets offer a means to conduct detailed qualitative evaluations, allowing for an in-depth analysis of critical translation phenomena. This dual approach ensures a comprehensive assessment of MT systems, highlighting both their strengths and areas of improvement.

3.2.1 Gold standards: quantitative scores

We introduce four new openly available gold standards for Galician, two for each language pair (EN-GL and ES-GL), which serve as reliable references for evaluating the output quality of MT systems through automatic metrics such as BLEU, COMET, and ChrF. For each pair, the two gold standards share identical source sentences but differ in their construction process. The first set (gold standard 1) was created by compiling parallel texts from diverse websites and online resources, followed by a meticulous human review by professional translators to ensure high-quality alignment. This dual approach allows us to compare MT system behavior on both naturalistic and post-edited data, providing a richer evaluation context. The second set (gold standard 2) was generated through a hybrid process involving automatic translation and post-editing. Source sentences were first translated into Portuguese and then into Galician using specific MT systems (DeepL[11] and OpenTrad for ES-GL Loinaz et al. (2006); Google Translate and OpenTrad for EN-GL). A transliteration tool was applied to adapt Portuguese-influenced spellings into Galician. This automated process was followed by a thorough human review. All gold standards are made publicly available to support reproducibility and further research in Galician MT evaluation.

While the two gold standards are in perfect alignment with the rules of standard Galician, the version created through Portuguese offers an interesting translation alternative, tapping into the linguistic similarities between Galician and Portuguese and providing translations that might be considered closer to a traditional or purer form of Galician, as opposed to the more heavily influenced by Spanish syntax and vocabulary.

3.2.2 Test suites: qualitative scores

In addition to the gold standards, we developed two comprehensive test suites to evaluate the EN-GL and ES-GL MT systems, consisting of manually crafted sentences targeting linguistic phenomena that we have observed to challenge MT systems in the language pairs of interest. They allow us to perform fine-grained, linguistically motivated evaluation beyond what standard automatic metrics can detect.

The test suites for EN-GL and ES-GL MT evaluations are structured into 12 broad categories for the two pairs (e.g. ambiguity, agreement, proper nouns or verbal system), al-

[9] https://opus.nlpl.eu/

[10] https://github.com/proxectonos/corpora/blob/main/README_English.md

[11] https://www.deepl.com/en/translator

though they are not the same in each pair. Within each of the broad categories, linguistic phenomena are further divided into subcategories, resulting in a detailed taxonomy customized to address specific translation complexities. These categories were chosen based on common MT errors observed in preliminary system outputs and are inspired by known challenges in Romance language processing.

For example, the ambiguity category covers lexical and pronominal ambiguities, while agreement addresses issues of number and gender concord. An instance of lexical ambiguity is the sentence *The cat climbed up to the top of the tree and now he can't go down*, which in Galician translates to *O gato subiu á copa da árbore e agora non sabe baixar. Copa* in Galician can mean both *top of a tree* and *cup*, so the translation models are being tested on their ability to discern these two meanings. Pronominal ambiguity can be seen in the sentence *Did you all attend the conference last week?*, which translates in Galician to *Todos vós asististes á conferencia a semana pasada?*. Here, the English pronoun *you* should be correctly translated to the second-person plural Galician pronoun *vós* and not to the second-person singular *ti*.

The verbal system is the most extensive category, which includes phenomena such as subject-verb agreement, as well as periphrases and verb tenses. Accurate handling of these verbal constructions ensures that the translated text maintains the correct tense, aspect, and mood.

Overall, the EN-GL test suite evaluates 42 distinct linguistic phenomena (and contains 364 sentence pairs), and the ES-GL suite covers 38 linguistic phenomena (and contains 334 sentence pairs). Their structured taxonomy enables detailed analysis of MT system performance across different linguistic dimensions. The evaluation process involves translating the test suite sentences using the MT system under evaluation and comparing the output with expert translations. Pre-designed regular expressions facilitate a semi-automatic assessment by identifying key linguistic phenomena in the sentences, and manual scoring follows to ensure accuracy, as it involves assessing the correct translation of specific linguistic phenomena that cannot always be reliably evaluated through automated methods alone. This strategy provides granular insights into MT systems' strengths and weaknesses, complementing automatic evaluation metrics. More details on the evaluation processes are provided below.

4 Model details

We have conducted a series of experiments to identify the best translation strategies for our languages of interest. These experiments are grouped into two main strategies: seq2seq and generative models. Notably, we used the same ES-GL and EN-GL training corpora across all experiments. In what follows, we describe the models we have evaluated and the evaluation methods.

4.1 Seq2seq models

Below, we present the details of the different types of seq2seq models evaluated.

4.1.1 Bilingual Models

We trained transformer-based seq2seq models for each language pair (GL-EN and ES-GL) on a single A100 GPU using AdamW ($\beta_1 = 0.9$), ($\beta_2 = 0.998$), ($\epsilon = 10^{-8}$), with OpenNMT-py 3.2's default learning rate[12]. The models have 12 encoder/decoder layers, 16 attention heads, a hidden size of 512, and a 30K vocabulary. Training used a batch size of 2048 sentences (max length: 150 tokens) for 20 epochs. The GL-EN model was trained on 43.5M tokens, and ES-GL on 70M tokens. Models were converted to ctranslate2 [13] format to optimize the inference time and size. GL-EN models were trained on corpora containing 43.5 million sentences, and ES-GL models were trained on a dataset containing 70 million sentences. These bilingual models are the simplest and smallest architectures developed for these experiments, with a size of 500 MB.

4.1.2 Multilingual-5

We have trained a multilingual transformer-based seq2seq model from scratch in the four official languages of the Kingdom of Spain (Spanish, Catalan, Basque, Galician) and English, on a carefully curated selection of texts from the original datasets used for the bilingual models, ensuring an even representation of all languages within the model. The Catalan and Basque corpora were collected from the ILENIA project.[14] The size of the corpus was limited to 30M sentences per language to

[12] https://pypi.org/project/OpenNMT-py/
[13] https://pypi.org/project/ctranslate2/
[14] https://proyectoilenia.es/en/models/

prevent any single language pair from dominating the training process, thus promoting a more balanced performance across the included language combinations. The model was trained on a single A100 GPU for 40 epochs and has 12 encoder and 12 decoder layers, 16 attention heads, and a hidden size of 512 dimensions. The batch size was set to 4096 sentences, with a maximum length of 150 tokens per sentence. We used AdamW optimizer ($\beta_1 = 0.9$), ($\beta_2 = 0.998$), ($\epsilon = 10 - 8$), and the default learning rate set by OpenNMT-py 3.2. The Vocabulary size was set to 25K. Testing was performed with the transformed ctranslate2 version of the model.

4.1.3 NLLB-200

We also evaluated an existing larger multilingual model, the NLLB-200-distilled-600M model,[15] both without and with fine-tuning, a model that supports over 200 languages, including Spanish, English, and Galician. This distilled version, comprising 600M parameters, is particularly effective in translating between low-resource languages, as well as widely used language pairs like Spanish or English. Trained on extensive multilingual datasets and built upon the Transformer architecture, the model notably excels at zero-shot translation, enabling it to perform translations between language pairs even in the absence of direct training data.

The model was also fine-tuned with the same corpora used to train the bilingual and small multilingual models, in order to enhance translation quality for our specific language pairs (EN-GL and ES-GL). Fine-tuning involves additional training on domain-specific or language-focused datasets, enabling the model to better handle linguistic nuances, idiomatic expressions, and cultural context. This process is especially beneficial for low-resource languages like Galician, where high-quality parallel data is limited. By adapting the model to specific datasets with fine-tuning, we expect to improve accuracy and fluency.

4.2 Generative models

Given LLMs' generalization abilities, we evaluated generative models as alternatives to traditional seq2seq methods, focusing on cases where pre-trained knowledge aids translation. We assessed both base and instructed models.

4.2.1 Foundation LLMs

The models used in this group are from the Salamandra and Carballo families. The Salamandra family includes models in three sizes: 2B, 7B, and 40B parameters. We focused on the 2B version (see footnote 7), pre-trained from scratch on a vast corpus comprising 12.875 trillion tokens in 35 European languages and code. Its architecture features 24 layers, a hidden size of 2048, and 16 attention heads, totaling 2.25 billion parameters. The vocabulary size is 256k, and the model employs flash attention optimization. Carballo-8B is a highly specialized transformer-based causal language model with 8 billion parameters, optimized for the Galician language while retaining capabilities in Portuguese, Spanish, Catalan, and English. Developed through continual pre-training of Meta's Llama-3.1-8B model, this variant addresses the lack of large-scale Galician-centric language models (see footnote 6). The model was trained on five nodes, each with two NVIDIA A100 GPUs, at the Galician Supercomputing Center (CESGA) using Causal Modeling and DeepSpeed[16]. The training corpus contains approximately 20 billion tokens, with 5B dedicated to Galician exclusively. Training hyperparameters included a batch size of 4 per device, AdamW optimizer, a linear learning rate schedule, and a learning rate of 1e-4. Gradient accumulation was set to 8 to handle large data efficiently.

To evaluate the translation capabilities of these models, a five-shot prompting strategy was used, providing the model with five carefully selected example translations within the prompt to guide it in generating accurate translations for new input text. All tests were performed with vLLM Kwon et al. (2023).

4.2.2 Instructed LLMs

To represent instructed LLMs, we selected SalamandraTA-2B and Eulang-7B. SalamandraTA-2B is a multilingual MT model capable of handling translations across 30 languages (see footnote 8). Built on the foundational Salamandra 2B model, it underwent continual pre-training on 70 billion tokens of parallel data. The training was conducted on the Marenostrum5 supercomputer.[17] Eulang-7B is a continual pre-training of the Gemma-

[15] https://huggingface.co/facebook/nllb-200-distilled-600M

[16] https://github.com/deepspeedai/DeepSpeed

[17] https://www.bsc.es/ca/marenostrum/marenostrum-

7b-it[18] model, optimized for enhanced performance in general translation tasks.

5 Results

This section presents the results of the different experiments. Model performance was evaluated through quantitative metrics (BLEU and COMET) and a manual qualitative evaluation. The qualitative analysis assesses overall translation and examines whether the models accurately translate specific Galician linguistic phenomena from the test suite. The results of the quantitative evaluations are aggregated by economy and clarity but are disaggregated by datasets (those described in Section 3) and by language pairs/direction in Annex A. The qualitative evaluation is disaggregated by language pairs/direction in Annex B.

5.1 Quantitative Evaluation

The three plots in Figure 1 present the average BLEU score for the models disaggregated for ES-GL and EN-GL pairs, in both translation directions. Figures 1a and 1b show the BLEU scores for ES-GL and EN-GL, respectively, while Figure 1c stands for the overall mean. The scores highlight the effectiveness of different approaches, ranging from models trained from scratch to those fine-tuned with specialized corpora. Blue bars represent the models of the generative paradigm: EU-lang, salam_ta (abbreviation of Salamandra-2B_TA), carballo (abbreviation of Carballo-8B) and salam (abbreviation of Salamandra-2B), And grey bars stand for the seq2seq models: mult5 (referring to the small multilingual model), nllb (abbreviation of NLLB-200-distilled-600M), bil (abbreviation of bilingual), and nllb_f (referring to the fine-tuned NLLB-200-distilled-600M).

Figure 2 shows the same type of evaluation with the same models, averaging on all datasets, using COMET.

5.2 Qualitative Evaluation

A qualitative evaluation of 100 sentences from the test suites (Section 3.2.2) was conducted using two complementary approaches. Figure 3 presents global sentence-level scores, where each translation was judged as correct or incorrect based on overall adequacy and fluency. In contrast, Figure 4 offers a more fine-grained analysis by focusing on specific linguistic phenomena such as verbal morphology, agreement, or word order. Both evaluations used binary scoring (1 = correct, 0 = incorrect), but the targeted analysis in Figure 4 highlights systematic strengths and weaknesses of the models. Although this deviates from standard graded scales, it ensures consistency across annotators and emphasizes the presence or absence of critical errors, particularly in morphosyntactic patterns, rather than relative quality, while reducing subjectivity and simplifying the identification of systematic patterns in model behavior.

The plot in Figure 3 shows the results of the qualitative score regarding global assessment of sentences for the models used: the first and second plots (figures 3a and 3b) present the results for the ES-GL and EN-GL translations, respectively, while the third (Figure 3c illustrates the overall mean performance. Similarly, the three graphs in Figure 4 show the results for the qualitative score regarding specific target phenomena.

Finally, Table 1 presents Pearson's and Spearman's correlation coefficients comparing the global qualitative evaluation (Qual) with the two quantitative metrics (BLEU and COMET) across the three settings: ES-GL, EN-GL, and the overall mean.

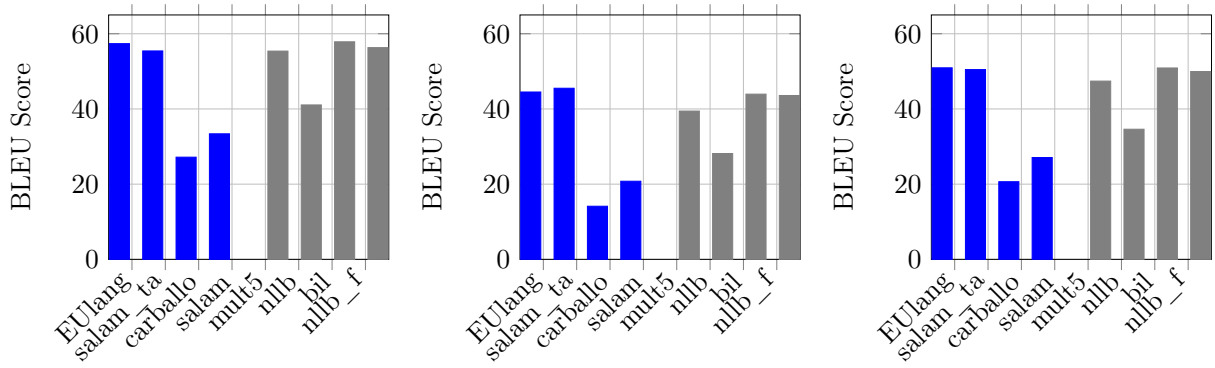
Qual.	Quant.	Pearson	Spearman
es-gl	BLEU	**0.8031**	**0.7413**
es-gl	COMET	0.7146	0.5714
en-gl	BLEU	**0.7978**	**0.7545**
en-gl	COMET	0.4647	0.2857
overall	BLEU	**0.6934**	**0.7619**
overall	COMET	0.5431	0.3851

Table 1: Pearson and Spearman correlation between qualitative and quantitative metrics for ES-GL, EN-GL, and the overall mean.

6 Discussion

In contrast to the results obtained in other studies Glushkova et al. (2023), in both ES-GL and EN-GL settings, the qualitative evaluation shows a stronger correlation with BLEU (0.8031 and 0.7978, respectively) than with COMET (0.7146 and 0.4647). This trend continues, as expected, in the overall mean, where the BLEU correlation (0.6934) is higher than the COMET correlation (0.5431). These results, shown in Table 1, suggest that the qualitative scores align more closely with BLEU,

[18]https://huggingface.co/google/gemma-7b-it

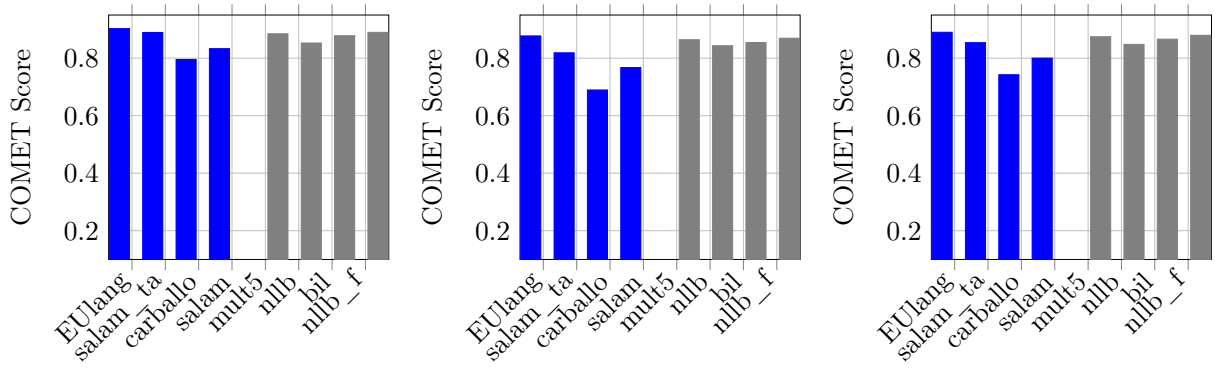


(a) BLEU scores for the ES-GL pair

(b) BLEU scores for the EN-GL pair

(c) Overall BLEU scores

Figure 1: BLEU evaluation scores for different translation models across language pairs.



(a) COMET scores for the ES-GL pair

(b) COMET scores for the EN-GL pair

(c) Overall COMET scores

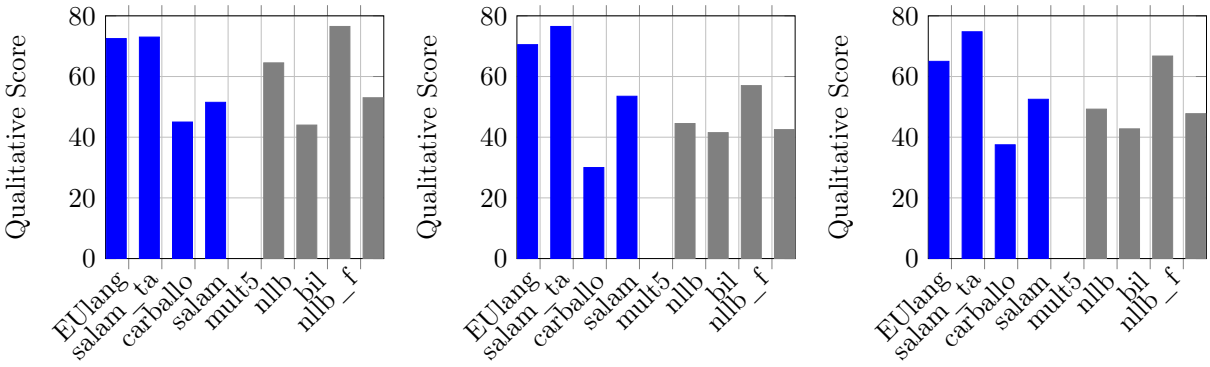
Figure 2: COMET evaluation scores for different translation models across language pairs.

though there is still a moderate relationship with COMET. Because BLEU correlates better, the rest of the discussion will focus on the results of both the qualitative and BLEU evaluation.[19]

Although the correlation between the BLEU metric and the qualitative assessment is high, there are significant disparities. The two main disparities refer to the bilingual models and Salamandra-2B_TA, which are rated higher in the qualitative than in the quantitative evaluation. Another striking disparity is that the difference between NLLB-200 and its fine-tuned version (nllb_f) is diluted in the qualitative assessment, while BLEU gives a clear advantage to the latter.

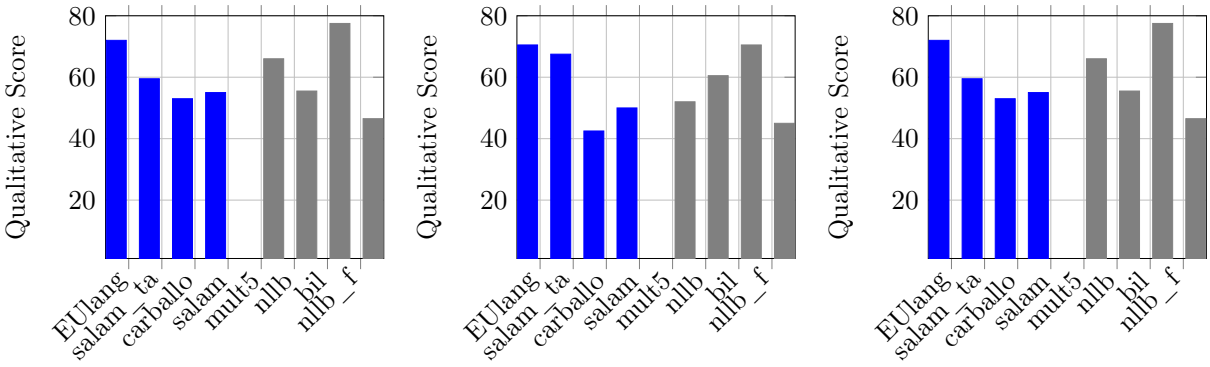
Concerning the overall scores in Figures 1c and 3c, there are no major differences between the best-performing seq2seq model, the bilingual one, and the best-performing LLMs, which are the two fine-tuned LLMs: EUlang and Salamandra-2B_TA, although the latter performs slightly better. However, a clear distinction emerges when comparing seq2seq and LLMs across language pairs in the qualitative global assessment. In the case of the EN-GL pair (see Figure 3b), the two fine-tuned LLMs, particularly Salamadra-2B_TA, surpass all other models. This trend reinforces the advantage of LLMs in handling complex language structures and less directly related languages. Yet, this could also be attributed to the strong English representation in the foundation model before fine-tuning, which may help mitigate linguistic distance. By contrast, for similar languages as ES-GL (Figure 3a), the bilingual seq2seq model achieves the best performance. This indicates that parallel corpora and simple architectures are more reliable for close languages than other architec-

[19] All experiments have also been performed with two other n-gram based metrics: Translation Edit Rate (TER) and Character F-score (chrF). These correlate with BLEU and, therefore do not add new information to the evaluation process.



(a) Scores for the ES-GL pair (b) Scores for the EN-GL pair (c) Overall scores

Figure 3: Qualitative scores for global assessment for different translation models across language pairs.



(a) Scores for the ES-GL pair (b) Scores for the EN-GL pair (c) Overall scores

Figure 4: Qualitative scores for target phenomena for different translation models across language pairs.

tures. Furthermore, among the non-instructed foundation models, the most multilingual, Salamandra-2B, performs better in translation than the less multilingual, Carballo-8B, highlighting the benefits of broader linguistic coverage in pre-training in spite of the difference in size (number of parameters).

Unlike the qualitative evaluation, BLEU does not clearly distinguish which is the best type of model in any of the scenarios: the two fine-tuned LLMs have results similar to the best seq2seq models (see Figure 1c). Focusing on BLEU and the overall qualitative evaluation of the multilingual seq2seq models (figures 1c and 3c), there are no major differences among the three variants (multilingual-5 and NLLB-200 as-is and fine-tuned). However, qualitative global assessment slightly favors the smallest multilingual model (multilingual-5), suggesting that the amount of language pairs covered does not necessarily lead to improved translation quality.

Concerning the qualitative scores for target linguistic phenomena, Figure 4c shows that the seq2seq bilingual model achieved the best results, while EUlang was the best generative LLM. A detailed analysis of the linguistic errors made by the bilingual models reveals that the greatest translation difficulties are found mainly with pronouns and inflectional verb forms. See the four tables in Annex C for more details about the types of linguistic errors made by the bilingual models. The fact that the seq2seq bilingual models achieved the best results for target linguistic phenomena suggests that language-specific parallel corpora are highly effective in capturing and reproducing the differences in linguistic structure between the source and target languages. Bilingual seq2seq models remain strong contenders for MT tasks, especially when linguistic accuracy is a priority.

Finally, we must point out that, in the case of generative LLMs, we have detected impor-

tant differences in performance depending on the prompting used by each type of model, which makes their use very unstable.

7 Conclusions and Future Work

7.1 Conclusions

In this study, we conducted a thorough evaluation of two MT paradigms, both sequence-to-sequence and generative, for two language pairs, a distant and a close pair, and several evaluation strategies. Our findings reveal that the qualitative evaluation shows a stronger correlation with the BLEU metric than with COMET, and provides insights that BLEU alone cannot capture, particularly in distinguishing model performance for specific linguistic phenomena. We also found that bilingual seq2seq models, despite the fact of being the simplest and smallest architectures, remain strong MT systems, especially for similar languages leveraging their shared linguistic features to achieve high translation accuracy. However, when it comes to more distant language pairs such as English and Galician, fine-tuned generative models demonstrated superior performance, likely due to their broader pre-training and fine-tuning.

Overall, our work contributes to the development of more accurate and reliable MT systems for low-resource languages, highlighting the value of qualitative evaluation metrics and comprehensive test suites in assessing the performance of translation models.

7.2 Future Work

Building on the findings of this study, future research should explore several key areas to further improve MT for low-resource languages. Investigating the performance of translation models on additional low-resource language pairs, particularly those with significant structural differences from the source languages, would provide insight into the ability of models to make generalizations and to adapt to different linguistic landscapes. Furthermore, incorporating linguistic phenomena into training corpora containing morphological and syntactic structures that are not easy to translate could enhance their ability to manage complex linguistic sentences, allowing them to provide more accurate translations. Finally, to avoid the cost of human evaluation, more automatic methods should be defined in qualitative evaluation without losing their finesse in identifying linguistic problems at different levels. The use of LLMs as judges for this purpose will be explored.

Limitations

Despite the contributions of this study, several limitations should be acknowledged. We used synthetic texts to ensure sufficient data for a low-resource language such as Galician. Although synthetic corpora provide useful data, their use in training can introduce biases and limit the models' ability to generalize to new linguistic patterns. Furthermore, the development of more advanced evaluation metrics remains a challenge. While the metrics used in this study are comprehensive, they may not fully capture the subtleties of translation quality—semantic-based metrics like BLEURT and BERTScore, for instance, were not considered. Moreover, the focus on sentence-level evaluation does not fully assess the capacity of LLMs to handle broader discourse and inter-sentential context. To address this, we propose extending the evaluation to multi-sentence or paragraph-level inputs in future work.

Finally, the study centers on EN-GL and ES-GL pairs, which means the findings may not generalize to low-resource languages with different linguistic traits, particularly if test suites are to be replicated. By addressing these limitations and pursuing the proposed research directions, future work can further advance MT for under-resourced languages.

Acknowledgements

This research has received financial support from the Agencia Estatal de Investigación (LingUMT, grant PID2021-128811OA-I00), the Xunta de Galicia - Consellería de Cultura, Educación, Formación Profesional e Universidades (Centro de investigación de Galicia accreditation 2024-2027 ED431G-2023/04). Also, the ILENIA-Nós Project, which is funded by the Spanish Ministry of Economic Affairs and Digital Transformation and by the Recovery, Transformation, and Resilience Plan - Funded by the European Union - NextGenerationEU, with reference 2022/TL22/00215336. Iria de-Dios-Flores is supported by project JDC2022-049433-I, financed by the MCIN/AEI/10.13039/501100011033 and the European Union "NextGenerationEU/PRTR", and grant SGR 2021 00470, financed by AGAUR (Catalan Government).

References

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Campos, J. R. P., Fernández, P. M., Gómez, O. S., Otero, P. G., and García, A. (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.

Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., et al. (2022). No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

de Dios-Flores, I., Magariños, C., Vladu, A. I., Ortega, J. E., Campos, J. R. P., Garcia, M., Gamallo, P., Rei, E. F., Diz, A. B., González, M. G., et al. (2022). The nós project: Opening routes for the galician language in the field of language technologies. In *Proceedings of the workshop towards digital language equality within the 13th language resources and evaluation conference*, pages 52–61.

de Gibert Bonet, O., Kharitonova, K., Calvo Figueras, B., Armengol-Estapé, J., and Melero, M. (2022). Quality versus quantity: Building Catalan-English MT resources. In Melero, M., Sakti, S., and Soria, C., editors, *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.

Enis, M. and Hopkins, M. (2024). From LLM to NMT: Advancing low-resource machine translation with claude. *arXiv preprint arXiv:2404.13813*.

Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., Chaudhary, V., Goyal, N., Birch, T., Liptchinsky, V., Edunov, S., Grave, E., Auli, M., and Joulin, A. (2021). Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(1).

Federmann, C., Kocmi, T., and Xin, Y. (2022). NTREX-128 – news test references for MT evaluation of 128 languages. In Ahuja, K., Anastasopoulos, A., Patra, B., Neubig, G., Choudhury, M., Dandapat, S., Sitaram, S., and Chaudhary, V., editors, *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

Forcada, M. L. and Tyers, F. M. (2016). Apertium: a free/open source platform for machine translation and basic language technology. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation: Projects/Products*, Riga, Latvia. Baltic Journal of Modern Computing.

Freixeiro Mato, X. (2006). O galego e o castelán en galicia: unha convivencia conflitiva.

Gamallo, P., Rodríguez, P., de Dios-Flores, I., Sotelo, S., Paniagua, S., Bardanca, D., Pichel, J. R., and Garcia, M. (2024). Open generative large language models for galician. *arXiv preprint arXiv:2406.13893*.

Glushkova, T., Zerva, C., and Martins, A. F. T. (2023). BLEU meets COMET: Combining lexical and neural metrics towards robust machine translation evaluation. In Nurminen, M., Brenner, J., Koponen, M., Latomaa, S., Mikhailov, M., Schierl, F., Ranasinghe, T., Vanmassenhove, E., Vidal, S. A., Aranberri, N., Nunziatini, M., Escartín, C. P., Forcada, M., Popovic, M., Scarton, C., and Moniz, H., editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 47–58, Tampere, Finland. European Association for Machine Translation.

González, S. G., Claramunt, G. R., and Campos, J. R. P. (2024). Empirical evaluation of galician machine translation: Spanish–galician and english–galician systems. *PROPOR 2024*, page 411.

Goyal, N., Gao, C., Chaudhary, V., Chen, P.-J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzmán, F., and Fan, A. (2022). The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Isabelle, P., Cherry, C., and Foster, G. (2017). A challenge set approach to evaluating machine translation. *arXiv preprint arXiv:1704.07431*.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., et al. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Kocmi, T., Zouhar, V., Federmann, C., and Post, M. (2024). Navigating the metrics maze: Reconciling score magnitudes and accuracies. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1999–2014, Bangkok, Thailand. Association for Computational Linguistics.

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39.

Koishekenov, Y., Berard, A., and Nikoulina, V. (2023). Memory-efficient NLLB-200: Language-specific expert pruning of a massively multilingual machine translation model. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3567–3585, Toronto, Canada. Association for Computational Linguistics.

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., and Stoica, I. (2023). Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Loinaz, I. A., Arantzabal, I., Forcada, M. L., Guinovart, X. G., Padró, L., Pichel, J. R., Waliño, J., et al. (2006). Opentrad: Traducción automática de código abierto para las lenguas del estado español. *Procesamiento del Lenguaje Natural*, (37):357–358.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (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.

Puduppully, R., Kunchukuttan, A., Dabre, R., Aw, A. T., and Chen, N. (2023). DecoMT: Decomposed prompting for machine translation between related languages using large language models. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4586–4602, Singapore. Association for Computational Linguistics.

Ranathunga, S., Lee, E.-S. A., Prifti Skenduli, M., Shekhar, R., Alam, M., and Kaur, R. (2023). Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37.

Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. (2020). COMET: A neural framework for MT evaluation. *arXiv preprint arXiv:2009.09025*.

Sennrich, R., Haddow, B., and Birch, A. (2016). Edinburgh neural machine translation systems for WMT 16. *arXiv preprint arXiv:1606.02891*.

Tiedemann, J. (2009). *News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces*, volume V, pages 237–248.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218. Citeseer.

Tiedemann, J. (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.

Wang, L., Lyu, C., Ji, T., Zhang, Z., Yu, D., Shi, S., and Tu, Z. (2023). Document-level machine translation with large language models. In Bouamor, H., Pino, J., and Bali,

K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.

Xu, H., Kim, Y. J., Sharaf, A., and Awadalla, H. H. (2023). A paradigm shift in machine translation: Boosting translation performance of large language models. *ArXiv*, abs/2309.11674.

Zhu, W., Liu, H., Dong, Q., Xu, J., Huang, S., Kong, L., Chen, J., and Li, L. (2024). Multilingual machine translation with large language models: Empirical results and analysis. In Duh, K., Gomez, H., and Bethard, S., editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

A Quantitative results broken down by datasets

In this annex, we show the results of the eight translation models (one per table) across all the datasets and for the two language pairs in the two directions. Our three datasets are in blue.

Dataset	Direction	BLEU	COMET
test suite	en-gl	44.2	0.8783
	es-gl	70.2	0.9220
	gl-es	77.7	0.9384
	gl-en	53.5	0.8894
gold1	en-gl	34.3	0.9263
	es-gl	75.1	0.9396
	gl-es	75.6	0.9389
	gl-en	38.6	0.9315
gold2	en-gl	50.9	0.8836
	es-gl	43.3	0.8625
	gl-es	49.8	0.8639
	gl-en	62.4	0.8816
flores	en-gl	35.7	0.8695
	es-gl	21.8	0.8665
	gl-es	24.3	0.8656
	gl-en	39.2	0.8791
ntrex	en-gl	38.3	0.8413
	es-gl	34.2	0.8663
	gl-es	36.9	0.8696
	gl-en	36.4	0.8551
taCon	en-gl	41.2	0.8744
	es-gl	79.9	0.9494
	gl-es	83.2	0.9447
	gl-en	38.4	0.8643
tatoeba	en-gl	52.3	0.8742
	es-gl	63.5	0.8952
	gl-es	67.7	0.9163
	gl-en	57.8	0.8356

Table 2: Quantitative results for EUlang.

Dataset	Direction	BLEU	COMET
test suite	en-gl	41.4	0.8712
	es-gl	64.7	0.8902
	gl-es	69.3	0.9198
	gl-en	53.6	0.9046
gold1	en-gl	41.6	0.7594
	es-gl	78.3	0.8821
	gl-es	79.8	0.8674
	gl-en	47.4	0.6683
gold2	en-gl	50.1	0.7638
	es-gl	43.0	0.8801
	gl-es	48.8	0.8681
	gl-en	62.3	0.6901
flores	en-gl	34.9	0.7072
	es-gl	22.0	0.8648
	gl-es	23.7	0.8559
	gl-en	39.9	0.8785
ntrex	en-gl	35.4	0.8216
	es-gl	34.3	0.8641
	gl-es	35.7	0.8578
	gl-en	37.6	0.8630
taCon	en-gl	40.4	0.8964
	es-gl	75.3	0.9349
	gl-es	81.8	0.9382
	gl-en	37.5	0.8633
tatoeba	en-gl	47.6	0.8744
	es-gl	60.1	0.9126
	gl-es	59.2	0.9152
	gl-en	67.4	0.8999

Table 3: Quantitative results for Salamandra_TA.

Dataset	Direction	BLEU	COMET
test suite	en-gl	6.0	0.6851
	es-gl	23.7	0.8084
	gl-es	39.3	0.8342
	gl-en	30.6	0.8219
gold1	en-gl	10.7	0.6324
	es-gl	34.0	0.8114
	gl-es	42.8	0.7834
	gl-en	17.1	0.5923
gold2	en-gl	11.0	0.7165
	es-gl	33.9	0.8057
	gl-es	42.8	0.8232
	gl-en	28.4	0.6451
flores	en-gl	13.5	0.7394
	es-gl	15.9	0.8276
	gl-es	19.9	0.8371
	gl-en	21.4	0.8350
ntrex	en-gl	5.5	0.6176
	es-gl	19.2	0.8110
	gl-es	26.4	0.8180
	gl-en	17.4	0.7841
taCon	en-gl	4.4	0.5880
	es-gl	11.6	0.7202
	gl-es	10.6	0.6158
	gl-en	3.6	0.5344
tatoeba	en-gl	4.5	0.6460
	es-gl	27.3	0.8055
	gl-es	33.1	0.8354
	gl-en	23.7	0.8138

Table 4: Quantitative results for Carballo-8B.

Dataset	Direction	BLEU	COMET
test suite	en-gl	24.2	0.8235
	es-gl	36.2	0.8191
	gl-es	41.9	0.8786
	gl-en	24.8	0.8204
gold1	en-gl	18.8	0.7114
	es-gl	41.6	0.8240
	gl-es	45.8	0.8125
	gl-en	20.9	0.6165
gold2	en-gl	23.5	0.7165
	es-gl	34.8	0.8294
	gl-es	34.8	0.8318
	gl-en	27.5	0.6394
flores	en-gl	18.4	0.8190
	es-gl	17.6	0.8378
	gl-es	19.2	0.8416
	gl-en	20.8	0.8406
ntrex	en-gl	17.0	0.7889
	es-gl	23.5	0.8286
	gl-es	26.5	0.8316
	gl-en	19.4	0.8116
taCon	en-gl	8.9	0.7816
	es-gl	31.1	0.8447
	gl-es	30.8	0.7806
	gl-en	10.3	0.6651
tatoeba	en-gl	22.4	0.8207
	es-gl	37.3	0.8491
	gl-es	46.3	0.8550
	gl-en	34.0	0.8888

Table 5: Quantitative results for Salamandra-2B.

Dataset	Direction	BLEU	COMET
test suite	en-gl	36.5	0.8431
	es-gl	68.9	0.9121
	gl-es	70.1	0.9128
	gl-en	39.8	0.8536
gold1	en-gl	36.8	0.8791
	es-gl	78.0	0.9122
	gl-es	80.6	0.9213
	gl-en	41.4	0.8837
gold2	en-gl	55.9	0.8837
	es-gl	42.3	0.8816
	gl-es	48.2	0.8763
	gl-en	42.3	0.8732
flores	en-gl	31.4	0.8413
	es-gl	21.8	0.8152
	gl-es	23.7	0.8152
	gl-en	34.4	0.8762
ntrex	en-gl	31.6	0.8553
	es-gl	33.9	0.8725
	gl-es	35.9	0.8645
	gl-en	32.0	0.8672
taCon	en-gl	33.3	0.8817
	es-gl	73.2	0.9243
	gl-es	78.7	0.9241
	gl-en	33.8	0.8562
tatoeba	en-gl	48.3	0.8511
	es-gl	60.9	0.8821
	gl-es	59.1	0.8742
	gl-en	54.6	0.8532

Table 6: Quantitative results for Multilingual5.

Dataset	Direction	BLEU	COMET
test suite	en-gl	28.5	0.8108
	es-gl	45.6	0.8524
	gl-es	46.5	0.8647
	gl-en	25.6	0.8178
gold1	en-gl	28.7	0.8601
	es-gl	54.2	0.8920
	gl-es	58.2	0.9023
	gl-en	20.5	0.8715
gold2	en-gl	46.6	0.8852
	es-gl	34.7	0.8573
	gl-es	35.1	0.7766
	gl-en	35.4	0.8455
flores	en-gl	20.5	0.8142
	es-gl	14.6	0.7942
	gl-es	16.0	0.7951
	gl-en	24.7	0.8435
ntrex	en-gl	19.8	0.8375
	es-gl	24.3	0.8574
	gl-es	25.4	0.8430
	gl-en	20.5	0.8425
taCon	en-gl	24.6	0.8653
	es-gl	59.3	0.8925
	gl-es	60.2	0.8929
	gl-en	20.7	0.8353
tatoeba	en-gl	38.5	0.8472
	es-gl	52.6	0.8536
	gl-es	48.3	0.8631
	gl-en	39.5	0.8291

Table 7: Quantitative results for NLLB-200.

Dataset	Direction	BLEU	COMET
test suite	en-gl	42.4	0.8340
	es-gl	73.4	0.9015
	gl-es	77.0	0.8974
	gl-en	44.4	0.8372
gold1	en-gl	36.9	0.8716
	es-gl	79.6	0.9025
	gl-es	81.7	0.9332
	gl-en	42.1	0.8813
gold2	en-gl	49.0	0.8836
	es-gl	42.9	0.8735
	gl-es	49.9	0.8715
	gl-en	59.1	0.8803
flores	en-gl	33.5	0.8253
	es-gl	21.2	0.8016
	gl-es	23.8	0.8098
	gl-en	37.8	0.8406
ntrex	en-gl	35.2	0.8422
	es-gl	32.2	0.8649
	gl-es	33.1	0.8561
	gl-en	34.5	0.8570
taCon	en-gl	40.5	0.8758
	es-gl	84.3	0.9218
	gl-es	86.1	0.9318
	gl-en	41.7	0.8467
tatoeba	en-gl	58.9	0.8543
	es-gl	64.2	0.8637
	gl-es	60.7	0.8621
	gl-en	59.1	0.8324

Table 8: Quantitative results for the four bilingual models (en-gl,es-gl,gl-es,gl-en).

Dataset	Direction	BLEU	COMET
test suite	en-gl	37.6	0.8735
	es-gl	60.9	0.8680
	gl-es	68.4	0.9027
	gl-en	45.3	0.8528
gold1	en-gl	38.1	0.8453
	es-gl	77.0	0.9342
	gl-es	79.5	0.9433
	gl-en	43.6	0.9024
gold2	en-gl	47.7	0.8870
	es-gl	42.5	0.8755
	gl-es	48.8	0.8715
	gl-en	60.1	0.8893
flores	en-gl	34.0	0.8475
	es-gl	22.0	0.8711
	gl-es	24.5	0.8690
	gl-en	40.8	0.8725
ntrex	en-gl	35.7	0.8482
	es-gl	34.0	0.8628
	gl-es	36.6	0.8672
	gl-en	40.2	0.8611
taCon	en-gl	34.8	0.8518
	es-gl	76.9	0.9255
	gl-es	85.4	0.9447
	gl-en	40.8	0.8709
tatoeba	en-gl	51.4	0.8675
	es-gl	64.5	0.8854
	gl-es	67.4	0.8360
	gl-en	60.0	0.9012

Table 9: Quantitative results for fine-tuned NLLB-200.

B Qualitative results broken down by model

In this annex, we show the qualitative evaluation (global and by linguistic target phenomena) disaggregated by language pairs and directions.

Dataset	**en-gl**	**es-gl**	**gl-es**	**gl-en**
EUlang	70	67	77	71
salam_ta	56	55	64	79
carballo	50	46	60	61
salam	43	47	63	57
mult5	56	68	64	48
nllb	58	54	57	63
bil	72	76	79	69
nllb_f	44	41	52	46

Table 10: Target phenomena scores by language pairs. Percentage of correctness.

Dataset	**en-gl**	**es-gl**	**gl-es**	**gl-en**
EUlang	61	68	77	54
salam_ta	78	73	73	75
carballo	36	40	50	57
salam	55	47	56	52
mult5	41	68	61	27
nllb	40	38	50	44
bil	61	62	81	53
nllb_f	45	49	57	40

Table 11: Global assessment scores by language pairs. Percentage of correctness.

C Target phenomena: counting linguistic errors per category

The tables below show the errors made by the best model on sentences containing different types of linguistic features. Each table is focused on one particular language pair and a specific direction. The first column in each table specifies the linguistic phenomenon, the second column shows the number of errors made, and the third one stands for the total number of examples per linguistic feature.

Linguistic Feat.	**Errors**	**Total Ex.**
Ambiguity	1	7
Agreement	1	2
Grammatical structure	3	4
Sentence structure	2	3
Negation	0	2
Proper name	0	4
Numeration	0	2
Pronoun	3	15
Punctuation	2	19
Verbal inflection	8	38
Elliptical subject	0	2
Passive voice	1	2

Table 12: Linguistic errors in bilingual model EN-GL.

Linguistic Feat.	**Errors**	**Total Ex.**
Ambiguity	2	7
Agreement	1	2
Grammatical structure	1	4
Sentence structure	2	3
Negation	0	2
Proper name	1	4
Numeration	0	2
Pronoun	3	15
Punctuation	3	19
Verbal inflection	12	38
Elliptical subject	0	2
Passive voice	1	2

Table 13: Linguistic errors in bilingual model GL-EN.

Linguistic Feat.	**Errors**	**Total Ex.**
Adverb	0	1
Ambiguity	2	5
Agreement	0	3
Grammatical structure	1	3
Sentence structure	0	3
Proper name	1	5
Pronoun	5	16
Punctuation	1	18
Social register	1	3
Verbal inflection	4	40
Elliptical subject	0	2
Passive voice	0	1

Table 14: Linguistic errors in bilingual model GL-ES.

Linguistic Feat.	**Errors**	**Total Ex.**
Adverb	1	1
Ambiguity	1	5
Agreement	1	3
Grammatical structure	1	3
Sentence structure	0	3
Proper name	1	5
Pronoun	1	16
Punctuation	4	18
Social register	0	3
Verbal inflection	4	40
Elliptical subject	0	2
Passive voice	0	1

Table 15: Linguistic errors in bilingual model ES-GL.