Empirical Evaluation of Galician Machine Translation: Spanish–Galician and English–Galician Systems

Sofía García González
UPV/EHU
imaxin|software
sofia.garcia@imaxin.com

German Rigau Claramunt
IXA Group
UPV/EHU
german.rigau@ehu.eus

José Ramom Pichel Campos
Centro de Investigación
en Tecnoloxías da Información
(CiTIUS)
jramon.pichel@usc.gal

Abstract

This paper establishes an empirical evaluation of English–Galician and Spanish–Galician machine translation in legal, health and general domains. The evaluation of the current MT systems was conducted using various metrics and an error analysis. In addition, the first health domain Spanish–Galician test and a reference test for each language pair were developed.

1 Introduction

Neural machine translation (NMT) has become the state of the art in the field, usually outperforming rule-based (RBMT) and statistical machine translation (SMT) in most language pairs (Mohamed et al., 2021). However, the large amount of parallel data necessary to train NMT models is a major challenge for low-resource languages. Lately, some studies have shown that multilingual translation models outperform bilingual models in low-resource translation pairs (Haddow et al., 2022). This is mainly due to transfer learning and the ability of multilingual language models to benefit from high-resource language knowledge to improve the translation of low-resource ones (Ranathunga et al., 2023). Interestingly, other research indicates that, between similar languages, as Spanish–Galician, even RBMT remains competitive with NMT models for low-resource languages (Bayón and Sánchez-Gijón, 2019).

Besides, the evaluation of MT is also a challenging task due to the lack of standard test datasets. This prevents not only the accurate evaluation of existing translation systems, but also the comparison between different studies and experiments (Goyal et al., 2022).

The main motivation of this paper is to do an empirical study of Galician MT in two language pairs, English–Galician and Spanish–Galician. We have chosen these ones because they are the two pairs most developed for Galician in MT. Our focus is on the translation into Galician as we aimed to evaluate the translation quality into a minority language across various system types and language pairs. This direction of translation is often less researched than the reverse direction from the high-resource language to the low-resource one.[1] Thus, taking Galician as a paradigmatic case of a low-resource language, we will evaluate:

1. The efficiency of multilingual and bilingual NMT models in distant (English–Galician) and close (Spanish–Galician) language pairs in general and specific domains.

2. The performance gap between NMT and RBMT systems, especially in Spanish–Galician translation pair.[2]

3. The MT system translations through an error analysis.

To the best of our knowledge, this is the first study comparing the performance of different machine translation systems, in different domains and different linguistic closeness pairs for Galician.

2 Background

In 2012, García-Mateo and Arza (2012) pointed out that "The situation of Galician in terms of linguistic technological support gives rise to cautious optimism", although they also argued that a great deal of development of language technology resources was necessary. Ten years later, there has been an increase in resources and corpora created,

[1] Although experiments have been conducted to evaluate both translation directions for the Spanish–Galician and English–Galician pairs, this paper will only present the results for the translation towards Galician. However, we aim to present the results for the other direction in a future publication.

[2] No SMT system has been included in the experimental part of this article since, to the best of our knowledge, there is no SMT Spanish–Galician or English–Galician model available.

Domain	Dataset	Number of Sentences
Legal Domain	TaCon	1100
Health Domain	Covid-19-HEALTH Wikipedia	957
General Domain	Flores200-devtest	1012
	Tatoeba v2022-03-03	1018
	Nos_MT_Gold-EN–GL_1	1777
	Nos_MT_Gold-EN–GL_2	1777
Combined En-Gl Test Set		7651

Table 1: English–Galician test datasets sizes

Domain	Dataset	Number of Sentences
Legal Domain	TaCon	1100
Health Domain	New Spanish–Galician Health Test	959
General Domain	Flores200-devtest	1012
	Tatoeba v2022-03-03	3121
	Nos_MT_Gold-ES–GL_1	1998
	Nos_MT_Gold-ES–GL_2	1998
Combined Es-Gl Test Set		10198

Table 2: Spanish–Galician test datasets sizes

especially textual resources, but not in tools and services (Sánchez and Mateo, 2022). Lately, in 2021, *O Proxecto Nós*[3] (The Nós Project) raised, an initiative promoted by the Galician Government, aimed at providing the Galician language with openly licensed resources, tools, demonstrators and use cases in the area of intelligent language technologies (de Dios-Flores et al., 2022). In the last two years, they have developed corpora and models for Galician in different NLP areas, as well as machine translation among them. The following subsections will detail the resources currently available for the English–Galician and Spanish–Galician pairs. This will include evaluation datasets (Section 2.1) and MT systems (Section 2.2). Additionally, a brief explanation of the current metrics for MT evaluation will be provided (Section 2.3).

2.1 MT Evaluation Datasets

As any low-resource language, there is a great scarcity of datasets for Galician MT evaluation. In the generic domain, Galician is one of the languages included in the Tatoeba (Tiedemann, 2020) and the Flores200 (Goyal et al., 2022) test sets. These are two multilingual MT evaluation benchmarks that include a wide variety of languages, 100 and 200 respectively, most of them medium and low-resource languages. Lately, The Nós Project has developed evaluation datasets for English–Galician and Spanish–Galician language pairs. For each language pair there are two gold-standard test sets (Nos_MT_Gold_1 and Nos_MT_Gold_2) and a test suite[4] (Nós_MT_Test-suite). The difference between Nos_MT_Gold_1 and Nos_MT_Gold_2 in both language pairs is the Galician part. In Nos_MT_Gold_1 Galician is syntactically and morphologically closer to Spanish, whereas in Nos_MT_Gold_2 it is more similar to Portuguese. Finally, the Nos_MT_Test-suite contains sentences classified based on linguistic phenomena both in Spanish–Galician and English–Galician pairs. These phenomena can be lexical ambiguity between languages, for example words that exist both in Spanish and Galician but with different meanings, grammatical structures that change between Galician and English, etc.

As regards specific domains,[5] there are datasets that can also be used as evaluation tests. In the legal and administrative domain, the TaCon,[6] a multilingual open-source evaluation dataset (Spanish, English, Galician, Catalan and Basque) of the Spanish Constitution includes both language pairs. Furthermore, LEGA[7] is a legal-administrative Spanish–

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

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

[5] The specific domains evaluated in this project are legal and health domains. Considering that, these are the domains considered in this paper.

[6] https://live.european-language-grid.eu/catalogue/corpus/19785/overview/

[7] https://live.european-language-grid.eu/catal

Galician parallel corpus included in the CLUVI.[8]

Finally, in the health domain, Galician is included in the multilingual corpus of COVID-19[9] that includes an English–Galician bilingual corpus obtained from Wikipedia. There is no Spanish–Galician evaluation dataset in the health domain.

2.2 MT Systems for Galician Language

Galician is included in different MT systems such as: the RBMT system Apertium and the neural models: opusMT[10] (Tiedemann and Thottingal, 2020), mBART[11] (Tang et al., 2020), M2M100,[12] (Fan et al., 2021) No-Language-Left-Behind (NLLB200[13]) (Costa-jussà et al., 2022), the Spanish–Galician neural model developed by the *Plan de Tecnologías del Lenguaje – Gobierno de España* (Language Technology Plan-Spanish Government) (PlanTL[14]) and the Spanish–Galician[15] and English–Galician[16] neural models developed by the Nós Project in both directions (Ortega et al., 2022).

1. **Apertium:**[17] Apertium is an open-source machine translation system, which uses the RBMT paradigm, and is particularly suitable for close or very close languages. It was created by the *Universitat d'Alacant* (Alacant University), the *Universidade de Vigo* (University of Vigo) and other public and private institutions in 2006 (Forcada et al., 2011). Nowadays, this is the system used by OpenTrad,[18] implemented in the automatic translator GAIO[19] of *Xunta de Galicia* (Galician Government). The language pairs available nowadays for Galician in this system are Spanish–Galician, Portuguese–Galician and English–Galician.

2. **OpusMT:**[20] OpusMT is a neural machine translation system for different languages trained on OPUS data based on Marian-NMT architecture. Additionally, the opus-mt-en-ROMANCE[21] multilingual model, is capable of translating from English to various romance languages.

3. **mBART:**[22] mBART is a multilingual sequence-to-sequence architecture that extends the capabilities of the BART model. It is pre-trained with a large multilingual corpus, in order to perform different tasks in 50 languages. This model has three different versions depending on the configuration: many-to-many-mmt, one-to-many-mmt and many-to-one-mmt (Tang et al., 2020).

4. **M2M:**[23] M2M is a sequence-to-sequence non-English-centric open source multilingual translation model that can translate directly between any pair of 100 languages. There are three different M2M models depending on the number of training parameters: m2m100_418M,[24] m2m100_1.2B[25] and m2m100_12B.[26]

5. **NLLB:**[27] NLLB-200 is a multilingual MNT model, specifically designed for low-resource language integration, capable of translating between 200 languages. As the M2M systems, there are different models depending on the number of training parameters: nllb-200-distilled-600M,[28]

ogue/corpus/12187

[8] *Corpus Lingüístico da Universidade de Vigo* (University of Vigo Linguistic Corpus). Open-Source multilingual and parallel dataset from the University of Vigo, https://ilg.usc.gal/cluvi/

[9] https://live.european-language-grid.eu/catalogue/corpus/3538

[10] https://huggingface.co/Helsinki-NLP/opus-mt-es-gl

[11] https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt

[12] https://huggingface.co/facebook/m2m100_418M

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

[14] https://huggingface.co/PlanTL-GOB-ES/mt-plantl-es-gl

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

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

[17] https://github.com/apertium

[18] https://opentrad.com/ open-source machine translation service platform of the company **imaxin**|software

[19] https://tradutorgaio.xunta.gal/TradutorPublico/traducir/index

[20] https://huggingface.co/Helsinki-NLP

[21] https://huggingface.co/Helsinki-NLP/opus-mt-en-ROMANCE

[22] https://huggingface.co/docs/transformers/model_doc/mbart

[23] https://huggingface.co/docs/transformers/model_doc/m2m_100

[24] https://huggingface.co/facebook/m2m100_418M

[25] https://huggingface.co/facebook/m2m100_1.2B

[26] https://huggingface.co/facebook/m2m100-12B-last-ckpt

[27] https://huggingface.co/docs/transformers/model_doc/nllb

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

`nllb-200-distilled-1.3B`,[29] `nllb-200-1.3B`[30] and `nllb-200-3.3B`.[31]

6. **`PlanTL`:**[32] `PlanTL` is a machine translation system implemented in the Ministry of Public Administration of the Government of Spain, specifically designed for translation between Spanish and the other official Spanish languages (Galician, Basque and Catalan).

7. **`Nos_MT-OpenNMT`:**[33] `Nos_MT-OpenNMT` are two open-source NMT bilingual models specifically designed for English–Galician and Spanish–Galician machine translation developed for the OpenNMT neural machine translation platform.

2.3 Evaluation Metrics

The MT evaluation is a challenging task that can be divided into two main categories: human evaluation and automatic evaluation.[34]

According to Lee et al. (2023) automatic evaluation metrics can be categorised as: lexical-based metrics, embedding-based metrics and supervised-metrics.

Lexical-based metrics measure the overlap between the hypothesis and the reference at a lexical level (word, phrase, character, etc.). Such metrics can measure the n-gram matching at word level as BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002) or METEOR (Metric for Evaluation of Translation with Explicit ORdering) (Banerjee and Lavie, 2005), and at character level as chrF (Character n-gram metric) (Popović, 2015). Moreover, lexical-based metrics can measure the edit distance between the reference and the hypothesis measuring, on the one hand, the number of insertions, deletions and substitutions necessary to convert one word into another as WER (Word Error Rate) (Tomás et al., 2003) or the number of edit operations that an hypothesis requires to match a reference translation including shift of word sequences apart from insertion, deletion and substitution of words as TER (Translation Error Rate) (Snover et al., 2006).

Regarding embedding-based metrics, they capture the similarity between hypothesis and reference using the embedding of language models (Lee et al., 2023). The main embedding-based metrics are BERTScore (Zhang* et al., 2020) and the current state-of-the-art MT evaluation metric, COMET (Crosslingual Optimized Metric for Evaluacion of Translation) (Rei et al., 2022).

Finally, the supervised metrics are the ones trained by machine learning or deep learning methods using labeled data (Lee et al., 2023). Two examples of this type of metric to MT evaluation are BERT for MTE (Machine Translation Evaluation) (Takahashi et al., 2020) and BLEURT (Sellam et al., 2020).[35]

3 Methodology

To determine the current state of the art of English–Galician and Spanish–Galician MT, we have collected some of the previously mentioned MT evaluation datasets in legal, health and general domains for both language pairs (Section 3.1) to evaluate all available MT systems (Section 3.2) with the main MT metrics (Section 3.3).

3.1 Evaluation Datasets

Table 1 and Table 2 display the chosen test set sizes for the English–Galician and Spanish–Galician pairs respectively. As it can be seen in both tables, the TaCon test is the one used to evaluate the legal domain, while Flores200-devtest, Tatoeba v2022-03-03, Nós_MT_Gold_1 and Nós_MT_Gold_2 are used to evaluate the general domain. The two gold standards created by the Nós project have enabled the comparison between the two Galician language solutions.[36]

In the health domain we used, on the one hand, the Covid19-HEALTH-Wikipedia in the English–Galician pair and, on the other hand, we created our own Spanish–Galician test set by selecting 1000 random sentences from the Spanish Biomedical

[29] https://huggingface.co/facebook/nllb-200-distilled-1.3B

[30] https://huggingface.co/facebook/nllb-200-1.3B

[31] https://huggingface.co/facebook/nllb-200-3.3B

[32] https://administracionelectronica.gob.es/ctt/verPestanaGeneral.htm?idIniciativa=plata

[33] https://huggingface.co/proxectonos

[34] This paper presents a section on error analysis 5.1 as human evaluation, but it is primarily focusing on automatic evaluation with reference-based metrics. Thus, this section will highlight the main MT evaluation metrics.

[35] Supervised methods are dependent on annotated data and, as mentioned by Lee et al. (2023), these are metrics difficult to use in low-resource languages and specific domains, thus they are not included in this article.

[36] The test-suites are not included in this paper as they are very small datasets focused on very specific phenomena.

Crawled Corpus[37] (Carrino et al., 2021). After cleaning the corpus, we were left with 959 sentences that were manually translated into Galician by professional linguists.

Finally, we have compiled a final test set for each language pair that encompasses all six preceding datasets. This comprehensive final test set allows for a conclusive evaluation of the MT models, the combined test set.

3.2 Translation Systems

Taking into account the systems referred to in sub-section 2.2, the ones used in this paper to carry out the evaluations are: the RBMT system, `Apertium`[38] and the bilingual and multilingual neural models: `opus-mt-en-gl`,[39] `opus-mt-es-gl`,[40] `opus-mt-en-ROMANCE`,[41] `mbart-large-50-many-to-many-mmt`,[42] the `M2M` and `NLLB` models,[43] `Nos_MT-OpenNMT-es-gl`,[44] `Nos_MT-OpenNMT-en-gl`,[45] `mt-plantl-es-gl`.[46]

To facilitate and speed up the translation process, we have used the Easy-Translate script.[47] This script allows the translation of large amounts of text in a single command. It is built on top of Transformers and accelerate PyTorch library (García-Ferrero et al., 2022).

3.3 Evaluation Metrics

According to the metrics mentioned in section 2.3, we have evaluated the performance of MT sytems using three lexical-based metrics: one word-based metric (BLEU), one character-based metric (chrF) and one edit-distance based metric (TER) using the SacreBleu script[48] as recommended by Post (2018). And, furthermore, one embedding-based metric that includes Galician in its model, COMET. We have chosen the default reference-based `wmt22-comet-da` model available in COMET webpage.[49]

4 Results

To facilitate the visualisation and comparison of results across all MT systems and language pairs, there will be a table for each test showing the results of both language pairs (English–Galician and Spanish–Galician).

The legal domain test in table 3, the health domain test in table 4 and the four general domain tests: Flores200 (Table 5), Tatoeba (Table 6), Nos_MT_Gold_1 (Table 7), and Nos_MT_Gold_2 (Table 8). Finally, the results of the combined test constructed from all the preceding data sets are visible in table 9.

The best results for each metric are emphasised in bold and, and in cases where one model outperformed on all metrics, it is also highlighted.

5 Analysis

Some conclusions can be drawn from the results of the analyses. Firstly, the Spanish–Galician models results tend to be twice as good as the English–Galician ones in BLEU, chrF and TER metrics. This pattern deviates only in the Flores200 (Table 5) and Nos_MT_Gold_2 (Table 8) tests, which will be analysed in the Error Analysis section, 5.1. Thus, the closer the language pair, the better the results in n-gram matching metrics.

Secondly, the difference in the results obtained between bilingual (`PlanTL` and `Nos_NMT`) and all the large multilingual NMT models types (`M2M100` and `NLLB200`) is not remarkable considering the difference in size. In fact, increasing the parameters in multilingual models leads to better results, however, the difference is not as significant as expected. On the other hand, the smaller multilingual models, such as `OpusMT` or `mBART`, achieve poor results in most test sets, especially in the English–Galician pair. Furthermore, `Apertium` seems to be competitive with the bilingual and the largest

[37] https://zenodo.org/record/5510033#.ZA5i_BzMJH5

[38] The OpenTrad website versions owned by **imaxin**|software were used for both translation pairs in this paper. However, free versions of Apertium for both pairs, Spanish–Galician (https://github.com/apertium/apertium-es-gl and English–Galician (https://github.com/apertium/apertium-en-gl) are available on GitHub.

[39] https://huggingface.co/Helsinki-NLP/opus-mt-en-gl

[40] https://huggingface.co/Helsinki-NLP/opus-mt-es-gl

[41] https://huggingface.co/Helsinki-NLP/opus-mt-en-ROMANCE

[42] https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt

[43] We have used all the available models from both M2M and NLLB

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

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

[46] https://huggingface.co/PlanTL-GOB-ES/mt-plantl-es-gl

[47] https://github.com/ikergarcia1996/Easy-Translate

[48] https://pypi.org/project/sacrebleu/

[49] https://github.com/Unbabel/COMET/blob/master/README.md

English–Galician Models	BLEU	chrF	TER	COMET	Spanish–Galician Models	BLEU	chrF	TER	COMET
opus-mt-en-ROMANCE	17.9	53.4	67.2	0.79	mBART-large-50-many-to-many	5.6	32.4	194.5	0.55
mBART-large-50-many-to-many	20.5	55.4	61.6	0.87	M2M_418M	70.1	89.4	14.3	0.94
M2M_418M	30.1	61.7	52.8	0.88	M2M_1.2B	72.5	90.4	12.9	0.94
M2M_1.2B	35.5	66.1	47.8	**0.90**	M2M_12B	74.5	90.1	12.5	0.94
M2M_12B	**39.3**	**68.3**	**45.2**	**0.90**	NLLB_200-600M	56.9	80.8	26	0.93
NLLB_200-600M	32.2	62.7	50.0	0.89	NLLB_200-distilled-1.3B	53.4	76.1	29.7	0.86
NLLB_200-distilled-1.3B	31.6	64.1	57.1	0.83	NLLB_200-1.3B	62.1	83.2	21.5	0.93
NLLB_200-1.3B	35	66	48.4	**0.90**	NLLB_200-3.3B	64.6	84.3	19.4	0.94
NLLB_200-3.3B	37.7	67.5	46.9	**0.90**	Apertium	74.6	90	10.7	**0.95**
Apertium	18.3	53.1	64.8	0.77	opus-mt-es-gl	68.9	87.2	14.6	0.94
opus-mt-en-gl	16.6	47.5	70.3	0.69	Nos_MT-OpenNMT-es-gl	81.5	92	11.4	**0.95**
Nos_MT-OpenNMT-en-gl	37.7	67.4	46.1	0.89	**mt-plantl-es-gl**	**84.3**	**93.5**	**8.6**	**0.95**

Table 3: Results in English–Galician and Spanish–Galician models in the legal domain test (TaCon)

English–Galician Models	BLEU	chrF	TER	COMET	Spanish–Galician Models	BLEU	chrF	TER	COMET
opus-mt-en-ROMANCE	21.8	52.4	66.9	0.77	mBART-large-50-many-to-many	33.2	60.3	55.5	0.77
mBART-large-50-many-to-many	26.8	57.2	71.6	0.83	M2M_418M	79.3	90.5	11.9	0.92
M2M_418M	32.1	60.1	57.8	0.84	M2M_1.2B	82.3	91.9	10.5	**0.93**
M2M_1.2B	36.6	62.7	54.8	**0.86**	M2M_12B	81.9	91.4	11.2	0.92
M2M_12B	37.5	63.4	55.1	**0.86**	NLLB_200-600M	63.3	82.6	23.1	0.91
NLLB_200-600M	34.6	61.7	56.2	**0.86**	NLLB_200-distilled-1.3B	65.9	83.4	21.7	0.91
NLLB_200-distilled-1.3B	36.6	63	54.3	**0.86**	NLLB_200-1.3B	66.4	83.7	21.4	0.91
NLLB_200-1.3B	36.4	63.4	55.4	**0.86**	NLLB_200-3.3B	68.2	84.3	20.3	0.92
NLLB_200-3.3B	37.4	63.6	53.4	**0.86**	Apertium	82.5	92.4	10.1	**0.93**
Apertium	14.3	48.2	74.3	0.64	opus-mt-es-gl	76.8	90.2	13.3	0.92
opus-mt-en-gl	16.6	44.5	72.1	0.60	Nos_MT-OpenNMT-es-gl	82.5	92.3	11.1	**0.93**
Nos_MT-OpenNMT-en-gl	**42**	**65.5**	**52.6**	0.85	**mt-plantl-es-gl**	**84**	**92.8**	**9.3**	**0.93**

Table 4: Results in English–Galician and Spanish–Galician models in health domain tests

English–Galician Models	BLEU	chrF	TER	COMET	Spanish–Galician Models	BLEU	chrF	TER	COMET
opus-mt-en-ROMANCE	20	54.2	67.9	0.77	mBART-large-50-many-to-many	12.2	42.9	82.2	0.73
mBART-large-50-many-to-many	25.7	56.9	59.6	0.83	M2M_418M	21.7	52.1	66	0.86
M2M_418M	29.4	60	56.7	0.82	M2M_1.2B	22.4	52.6	65.9	0.86
M2M_1.2B	33.8	63	52.4	0.85	M2M_12B	22.4	52.9	66.5	0.86
M2M_12B	35	63.7	**50.7**	0.86	NLLB_200-600M	22.1	52.8	66.8	0.86
NLLB_200-600M	31.9	62.2	54.8	0.86	NLLB_200-distilled-1.3B	**23.9**	**53.9**	**64.5**	0.86
NLLB_200-distilled-1.3B	34.9	64	51.2	**0.87**	NLLB_200-1.3B	23.3	53.5	64.6	0.86
NLLB_200-1.3B	34.9	63.8	51.2	**0.87**	NLLB_200-3.3B	23.8	53.6	64.6	**0.87**
NLLB_200-3.3B	**35.6**	**64.4**	**50.7**	**0.87**	Apertium	18.9	50.6	66.6	0.84
Apertium	16.0	50.3	71.6	0.66	opus-mt-es-gl	20.8	51.7	65.7	0.85
opus-mt-en-gl	19.3	51.7	68.2	0.66	Nos_MT-OpenNMT-es-gl	21.5	51.9	68	0.85
Nos_MT-OpenNMT-en-gl	31.6	62.3	55.8	0.83	mt-plantl-es-gl	21.9	52.3	64.7	0.86

Table 5: Results in English–Galician and Spanish–Galician models in general domain:Flores200-devtest

English–Galician Models	BLEU	chrF	TER	COMET	Spanish–Galician Models	BLEU	chrF	TER	COMET
opus-mt-en-ROMANCE	25.3	50.1	59	0.78	mBART-large-50-many-to-many	27.1	51.3	59.5	0.78
mBART-large-50-many-to-many	37.0	59.6	47.6	0.83	M2M_418M	53.8	71.1	32.3	0.88
M2M_418M	37.5	58.7	48.3	0.83	M2M_1.2B	55.4	72.2	32.3	0.88
M2M_1.2B	41.9	62.9	44.4	0.85	M2M_12B	50.6	67.9	37.6	0.87
M2M_12B	41.5	61.8	45.3	0.86	NLLB_200-600M	50.1	69	34.7	0.88
NLLB_200-600M	42.7	64.3	42.7	0.87	NLLB_200-distilled-1.3B	54.6	72.3	31	0.89
NLLB_200-distilled-1.3B	47	67.7	40	**0.88**	NLLB_200-1.3B	53.1	71.2	32.1	0.89
NLLB_200-1.3B	46.4	67	40.2	**0.88**	NLLB_200-3.3B	56.9	73.4	29.5	0.89
NLLB_200-3.3B	48.4	68.8	**38.5**	**0.88**	Apertium	**68.4**	81	**19.8**	**0.91**
Apertium	27.2	51.9	57.8	0.76	opus-mt-es-gl	67.8	**81.3**	20.4	**0.91**
opus-mt-en-gl	37.4	60.2	47.6	0.87	Nos_MT-OpenNMT-es-gl	61.4	76.9	27.2	0.89
Nos_MT-OpenNMT-en-gl	**48.6**	**69.8**	39.8	0.81	mt-plantl-es-gl	66.1	79.2	22.6	**0.91**

Table 6: Results in English–Galician and Spanish–Galician models in general domain:Tatoeba

English–Galician Models	BLEU	chrF	TER	COMET	Spanish–Galician Models	BLEU	chrF	TER	COMET
opus-mt-en-ROMANCE	21.1	54.4	63.6	0.79	mBART-large-50-many-to-many	30.6	60	55.5	0.76
mBART-large-50-many-to-many	26	56.2	57.3	0.84	M2M_418M	72.9	85.3	19.7	0.88
M2M_418M	32.1	60.6	52.3	0.85	M2M_1.2B	77.1	87.2	17.5	0.89
M2M_1.2B	38.1	64.6	47.1	0.87	M2M_12B	77.6	87.2	17.4	0.89
M2M_12B	**39.4**	**65.4**	**46.2**	**0.88**	NLLB_200-600M	58.5	77.9	29.1	0.88
NLLB_200-600M	35.4	62.9	48.7	0.87	NLLB_200-distilled-1.3B	62.2	79.7	26.6	0.88
NLLB_200-distilled-1.3B	38.1	64.9	46.8	**0.88**	NLLB_200-1.3B	62.7	80.1	26.4	0.88
NLLB_200-1.3B	38	64.9	46.7	**0.88**	NLLB_200-3.3B	65.5	81.2	24.5	0.89
NLLB_200-3.3B	38.7	65.2	46.3	**0.88**	Apertium	78.7	88	16.3	**0.90**
Apertium	17.6	49.2	66.5	0.69	opus-mt-es-gl	71.3	85	20	0.89
opus-mt-en-gl	20.1	50.8	64.1	0.72	Nos_MT-OpenNMT-es-gl	79	88.3	16.8	**0.90**
Nos_MT-OpenNMT-en-gl	35.6	63.4	50.8	0.85	**mt-plantl-es-gl**	**79.6**	**88.6**	**15.6**	**0.90**

Table 7: Results in English–Galician and Spanish–Galician models in general domain: NOS Gold Standard 1

English–Galician Models	BLEU	chrF	TER	COMET	Spanish–Galician Models	BLEU	chrF	TER	COMET
opus-mt-en-ROMANCE	31.9	63.8	49.1	0.81	mBART-large-50-many-to-many	23.4	53.4	63.4	0.76
mBART-large-50-many-to-many	33	61.5	48.2	0.85	M2M_418M	41.8	67	42.2	0.87
M2M_418M	43.6	68.3	39.2	0.87	M2M_1.2B	43.2	67..8	41.4	0.88
M2M_1.2B	50	72	34.8	0.89	M2M_12B	43.2	67.6	41.6	0.88
M2M_12B	49.6	72	35.2	**0.90**	NLLB_200-600M	42.1	67.7	42	0.88
NLLB_200-600M	48.1	71.3	35.3	0.89	NLLB_200-distilled-1.3B	44.1	68.6	40.6	0.88
NLLB_200-distilled-1.3B	50	72.1	34.8	**0.90**	NLLB_200-1.3B	43.4	68.3	41.1	0.88
NLLB_200-1.3B	48.7	71.6	35.4	**0.90**	**NLLB_200-3.3B**	**44.6**	**68.8**	**40.1**	**0.89**
NLLB_200-3.3B	**50.8**	**72.8**	**33.7**	**0.90**	Apertium	42.9	67.4	41.6	0.88
Apertium	25.2	56	55.9	0.71	opus-mt-es-gl	41.3	67.1	42.3	0.88
opus-mt-en-gl	29.2	58.1	52.4	0.75	Nos_MT-OpenNMT-es-gl	43.2	67.9	41.4	0.88
Nos_MT-OpenNMT-en-gl	45.9	69.9	40.2	0.87	mt-plantl-es-gl	43.3	67.9	41.1	0.88

Table 8: Results in English–Galician and Spanish–Galician models in general domain: NOS Gold Standard 2

English–Galician Models	BLEU	chrF	TER	COMET	Spanish–Galician Models	BLEU	chrF	TER	COMET
opus-mt-en-ROMANCE	23.5	54.6	59.6	0.79	mBART-large-50-many-to-many	21.4	46.9	69.1	0.74
mBART-large-50-many-to-many	27.7	56.8	56.2	0.84	M2M_418M	56.5	74.9	32.1	0.89
M2M_418M	34.7	61.8	50.3	0.85	M2M_1.2B	58.8	76	31	0.89
M2M_1.2B	39.9	65.2	45.7	0.87	M2M_12B	58.6	75.7	31.9	0.89
M2M_12B	41.3	66.2	45.4	**0.88**	NLLB_200-600M	48.7	70.7	37	0.88
NLLB_200-600M	37.8	64	47.1	**0.88**	NLLB_200-distilled-1.3B	51.4	72	35.8	0.88
NLLB_200-distilled-1.3B	40.1	65.3	46.2	0.87	NLLB_200-1.3B	51.7	72.2	34.9	0.89
NLLB_200-1.3B	40.1	65.5	45.4	**0.88**	NLLB_200-3.3B	53.7	73.4	33.6	**0.90**
NLLB_200-3.3B	**41.6**	**66.4**	**44.5**	**0.88**	Apertium	60.7	77.5	29	**0.90**
Apertium	18.5	50.6	65.4	0.70	opus-mt-es-gl	56.9	75.7	31	**0.90**
opus-mt-en-gl	22	49.9	62.2	0.71	Nos_MT-OpenNMT-es-gl	60.9	77.1	30.4	**0.90**
Nos_MT-OpenNMT-en-gl	39.8	65	47	0.86	**mt-plantl-es-glmt-plantl-es-gl**	**62.2**	**78**	**28.7**	**0.90**

Table 9: Results in English–Galician and Spanish–Galician models in the combined test datasets

multilingual NMT models in all Spanish–Galician test sets, demonstrating that RBMT is still efficient in closely related language pairs, even more than some multilingual models such as mBART.

Finally, all the metrics are consistent with each other. That is, they all give fair results depending on the quality of the models; if one model gives poor results, or the quality between models is similar, this is reflected in all the metrics without there being much variation between them. Within this pattern, however, COMET requires a separate analysis. The COMET results are higher than the other metrics, the lowest being 0.55 in the Spanish–Galician mbart-many-to-many model in the TaCon test, Table 3, which obtains very poor results in the other metrics and which will also be analysed later in 5.1. Moreover, the results between models do not vary as much as for the other metrics. Thus, many models achieve the same result in all tests, usually in those models that do not show significant variations in the other metrics. In fact, in some test sets the consistency with the other metrics disrupts. For example, in the legal domain, Table 3, the English–Galician M2M_1.2B, M2M_12B, NLLB_200-1.3B and NLLB_200-3.3B models get the same results in COMET, although between M2M_1.2B, NLLB_200-1.3B, NLLB_200-3.3B there is a difference of almost four points in the other metrics. Regarding the health domain and Tatoeba tests, table 4, the Nos_NMT-EN-GL model achieves the best results in all the metrics except COMET. Also in the Tatoeba test, table 6, the opusmt-es-gl, Apertium and PlanTL achieve the same COMET results, although there is a difference of almost two points between Apertium and PlanTL in the other metrics. Finally, in the combined test, table 9, the PlanTL achieves the same result in COMET as NLLB_200-3.3B in the Spanish–Galician pair, although there is a difference of almost ten points in the other metrics between these two models. This last discrepancy will be analysed in section 5.1. Therefore, to accurately interpret COMET results accurately, it is essential to consider its punctuation range in comparison to other metrics.

5.1 Error Analysis

In this section we will analyze the results previously highlighted: the difference between the results of the English–Galician and Spanish–Galician models on the Nos_MT_Gold_2 and Flores200 tests; the poor performance of the mBART model in the Spanish–Galician TaCon test and the discrepancy between the COMET results in the Spanish–Galician combined test between PlanTL and NLLB200_3.3B. For this analysis we have selected, for each test set, 100 random sentences from source, reference and the translations of the models selected.

With regard to the Nos_MT_Gold_2 and Flores200 test sets, the results seem to be determined by the linguistic characteristics of Galician. As already mentioned, the Nos_MT_Gold_2 Galician is syntactically and lexically closer to Portuguese than Nos_MT_Gold_1. In general, all the MT systems translate maintaining the word order and syntactic structure of the source language. This is therefore the reason for the drop in performance in the Nos_MT_Gold-ES–GL_2 compared to Nos_MT_Gold-ES–GL_1. All the translation models, preserve either the Spanish or the English structures, and therefore the results are very different. In the Spanish–Galician MT models, Galician is translated preserving the Spanish syntax and vocabulary, which explains the results in Nos_MT_Gold-ES–GL_2. Regarding the results in the English–Galician pair in the Nos_MT_Gold-EN–GL, the results between test 1 and 2 are more similar because the MT systems are mantaining English syntactic structures, which give poorer translations in Nos_MT_Gold-EN–GL_1 compared to Nos_MT_Gold-ES–GL_1. Flores200 presents a similar issue. The Spanish and Galician sections of Flores200 are based on non literal translations of the original English sentences, resulting in meaning that matches the originals but not their form. Consequently, the metric scores are low in both language pairs. This also clarifies why the Spanish–Galician results are generally better than English–Galician ones in all test sets.The closer the languages are, the easier it is for a literal translation to be correct. Although this closeness also presents challenges in multilingual models that tend to mix the languages.

On the other hand, we analysed the mBART-large-50-many-to-many-mmt Spanish–Galician model's translation in the TaCon test. The translation errors include omissions, hallucinations and a significant amount of language mixing. Short sentences like *artículo 1* (article 1) or *partido político* (political party) result in the model hallucination providing an unrelated legal paragraph, often the same one. It is possibly a paragraph from a legal text with which the model has been trained. However, in some instances where the meaning of the original sentence is

maintained, a mixture of Spanish or English terms with the Galician translation is used. On other occasions, no translation is provided at all. As a result, the model's translation in this legal domain is unsatisfactory. Although in other domains the translation quality of this model seems slightly better, e.g. the health domain in table 4.

Finally, we compared the translation of `planTL` and `nllb-3.3B` in the Spanish–Galician combined test. The multilingual model had some notable errors, including the insertion of Spanish terms in the Galician translation, misconjugation of certain verb tenses (e.g. *comeste* instead of *comiches* (You ate)), and mid-sentence omissions. It is worth considering whether COMET can accurately assess a term's translation when translated into the wrong language, or if the sentence contains errors in conjugation or construction, particularly in low-resource languages with which it has been trained. Discrepancies in test sets compared to other metrics may be due to such factors.

To conclude, additional errors were discovered in the reference sentences of the tests during the error analysis. The inaccuracies in Galician were evident in the Tatoeba test for both the Spanish–Galician and English–Galician pairs. Such errors include written terms in Spanish, like the personal pronoun *él* ('he') that should not have an accent in Galician, *el*; inaccurately conjugated verb forms which do not exist in Galician — such as *contraxo* instead of *contraeu* ('contracted')— and also the omission of important information from the original sentence. It is recognized that these errors are particularly serious in a MT benchmark.

6 Conclusions & Future Work

To summarize, based on the analysis provided, we can conclude that MT models often provide a literal translation of the original sentence. As a consequence, distant language pairs such as English–Galician may result in unsatisfactory translations due to this language distance. In contrast, similar language pairs, such as Spanish–Galician, do not present that issue due to the greater linguistic proximity. For this reason, in close language pairs, an RBMT model remains competitive despite the errors inherent to this type of systems. However, in the case of Galician, which has two valid linguistic solutions , the translations of Spanish–Galician models maintain structures and vocabulary similar to Spanish, leading to the gradual loss of genuine Galician linguistic phenomena. On the other hand, it was shown that only very large multilingual models outperform or even out the bilingual NMT models in both language pairs. Thus, NMT bilingual models can outperform the multilingual ones even in low-resource language pairs. Given these results, it is not only important to point out the competitiveness of bilingual neural models and, in the case of the Spanish–Galician pair, an RBMT system with large multilingual models in terms of translation quality, but also their environmental impact. As Shterionov and Vanmassenhove (2023) point out, an RBMT system does not require a large investment in computational resources, whereas neural models require a large consumption of energy both in their training and at the time of translation.[50]

Finally, it is worth noting the importance of ensuring the linguistic correctness of specific test sets used as benchmarks for MT evaluation. It is crucial to identify and rectify linguistic errors, as well as implementing measures to enhance the structure, syntax, morphology, and vocabulary. For this reason, we will release the first Spanish–Galician health test, along with reference MT tests for the English–Galician and Spanish–Galician pairs.

As part of future work, we will include other language pairs such as Portuguese–Galician, use additional metrics like BERTScore, conduct a more thorough analysis of each metric and a comprehensive review of the linguistic errors made by each model.

Acknowledgements

We would like to express our gratitude to the Nós project members for their assistance and guidance during the development of the methodological part of the project. Additionally, computational resources for this research were provided by UPV/EHU and **imaxin**|software. Finally, we acknowledge the funding received from the following projects:

(i) DeepKnowledge (PID2021-127777OB-C21) and ERDF A way of making Europe.

(ii) DeepR3 (TED2021-130295B-C31) and European Union NextGeneration EU/PRTR.

[50] In this paper we have not conducted a study of the computational and energy consumption required by each model when translating, however we plan to incorporate it in future work.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

María Do Campo Bayón and Pilar Sánchez-Gijón. 2019. Evaluating machine translation in a low-resource language combination: Spanish-Galician. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pages 30–35, Dublin, Ireland. European Association for Machine Translation.

Casimiro Pio Carrino, Jordi Armengol-Estapé, Ona de Gibert Bonet, Asier Gutiérrez-Fandiño, Aitor Gonzalez-Agirre, Martin Krallinger, and Marta Villegas. 2021. Spanish biomedical crawled corpus: A large, diverse dataset for spanish biomedical language models. *arXiv preprint arXiv:2109.07765*.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Iria de Dios-Flores, Carmen Magariños, Adina Ioana Vladu, John E. Ortega, José Ramom Pichel, Marcos García, Pablo Gamallo, Elisa Fernández Rei, Alberto Bugarín-Diz, Manuel González González, Senén Barro, and Xosé Luis Regueira. 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, Marseille, France. European Language Resources Association.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *The Journal of Machine Learning Research*, 22(1):4839–4886.

Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25:127–144.

Iker García-Ferrero, Rodrigo Agerri, and German Rigau. 2022. Model and data transfer for cross-lingual sequence labelling in zero-resource settings. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6403–6416, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Carmen García-Mateo and Montserrat Arza. 2012. *O idioma galego na era dixital – The Galician Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Springer, Heidelberg, New York, Dordrecht, London. Georg Rehm and Hans Uszkoreit (series editors).

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.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Seungjun Lee, Jungseob Lee, Hyeonseok Moon, Chanjun Park, Jaehyung Seo, Sugyeong Eo, Seonmin Koo, and Heuiseok Lim. 2023. A survey on evaluation metrics for machine translation. *Mathematics*, 11(4):1006.

Shereen A Mohamed, Ashraf A Elsayed, YF Hassan, and Mohamed A Abdou. 2021. Neural machine translation: past, present, and future. *Neural Computing and Applications*, 33:15919–15931.

John E Ortega, Iria de Dios-Flores, Pablo Gamallo, and José Ramom Pichel. 2022. A neural machin e translation system for galician from transliterated portuguese text. In *Proceedings of the Annual Conference of the Spanish Association for Natural Language Processing. CEUR Workshop Proceedings*, volume 3224, pages 92–95.

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.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55:1–37.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Dimitar Shterionov and Eva Vanmassenhove. 2023. *The Ecological Footprint of Neural Machine Translation Systems*, volume 4, pages 185–213. Springer Nature Switzerland AG, Switzerland. 25 pages, 3 figures, 10 tables Copyright © 2023, The Author(s), under exclusive license to Springer Nature Switzerland AG.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

José Manuel Ramírez Sánchez and Carmen García Mateo. 2022. Deliverable D1.15 Report on the Galician Language. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Kosuke Takahashi, Katsuhito Sudoh, and Satoshi Nakamura. 2020. Automatic machine translation evaluation using source language inputs and cross-lingual language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3553–3558, Online. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

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.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Jesús Tomás, Josep Àngel Mas, and Francisco Casacuberta. 2003. A quantitative method for machine translation evaluation. In *Proceedings of the EACL 2003 Workshop on Evaluation Initiatives in Natural Language Processing: are evaluation methods, metrics and resources reusable?*, pages 27–34, Columbus, Ohio. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.