Findings of the WMT 2024 Shared Task Translation into Low-Resource Languages of Spain: Blending Rule-Based and Neural Systems

Felipe Sánchez-Martínez,[†] Juan Antonio Pérez-Ortiz,[*][†]
Aarón Galiano-Jiménez,[†] Antoni Oliver[‡]

[†]Universitat d'Alacant {fsanchez,japerez,aaron.galiano}@ua.es
[*]Valencian Graduate School and Research Network of Artificial Intelligence, ValgrAI
[‡]Universitat Oberta de Catalunya aoliverg@uoc.edu

Abstract

This paper presents the results of the Ninth Conference on Machine Translation (WMT24) Shared Task "Translation into Low-Resource Languages of Spain". The task focused on the development of machine translation systems for three language pairs: Spanish–Aragonese, Spanish–Aranese, and Spanish–Asturian. 17 teams participated in the shared task with a total of 87 submissions. The baseline system for all language pairs was Apertium, a rule-based machine translation system that still performs competitively well, even in an era dominated by more advanced non-symbolic approaches. We report and discuss the results of the submitted systems, highlighting the strengths of both neural and rule-based approaches.

1 Introduction

In Spain, a diverse linguistic landscape exists, including, beyond the widely recognized Spanish, other languages such as Basque, Catalan, and Galician. Although Spanish is obviously at the forefront in terms of the volume of resources available for training data-driven machine translation (MT) systems, the capabilities and richness of the other languages should not be underestimated. Basque, Catalan, and Galician, which might have been considered limited in resources in the past, actually possess a significant amount of data that facilitate their integration into modern MT technologies. In fact, these three languages have been recently included among the list of up to 100 languages in well-known multilingual systems such as mBERT[1] (Devlin et al., 2019), XLM-R (Conneau et al., 2020), mBART (Liu et al., 2020), mT5 (Xue et al., 2021) or NLLB-200 (Costa-jussà et al., 2024). However, Spain is home to additional languages with much fewer resources, especially in the form of bilingual data. This task focuses on three of them, namely, Aragonese, Aranese, and Asturian, all of them Romance languages. In particular, participants were asked to submit MT systems from Spanish into any of these three languages.

An interesting fact about our three low-resource languages is that they have open rule-based MT systems available for the Apertium framework. Apertium (Forcada et al., 2011) is a free/open-source rule-based architecture for MT that consists of a pipeline of modules performing morphological analysis, part-of-speech tagging, lexical transfer, lexical selection, chunk-level or recursive structural transfer, and morphological generation.

Another important aspect is that the target languages of the shared task have undergone various orthographic conventions and standards, and the datasets, as well as the MT systems, available may not necessarily adhere to the current conventions adopted by the language academies and used in the test sets.

Submission platform. We utilized the open-source OCELoT platform[2] to collect translation submissions. The platform offers anonymized public leaderboards and has been employed in several previous WMT tasks. Submission privileges were restricted to registered and verified teams with accurate contact information, and each team was limited to a maximum of seven submissions per test set.

Main goals. The primary objectives of this shared task can be summarized as follows:

- To push the boundaries of MT system development when the amount of resources is extremely scarce.
- To explore the transferability among low-resource Romance languages when translating from Spanish.

[1] https://huggingface.co/google-bert/bert-base-multilingual-cased

[2] https://github.com/AppraiseDev/OCELoT

Proceedings of the Ninth Conference on Machine Translation, pages 684–698
November 15-16, 2024 ©2024 Association for Computational Linguistics

- To find the best way to use pre-trained models of any kind for the translation between Spanish and low-resource Romance languages.
- To create publicly available corpora for MT development and evaluation.

Main findings. The main conclusions of the shared task and insights gained are outlined next:

- The best systems result in automatic evaluation scores that are statistically significantly higher than the baseline system, namely, the Apertium rule-based system.
- However, the absolute differences in BLEU and chrF2 scores are not very large (up to 2 BLEU and 1 chrF2 points) in the case of Aragonese and Aranese, which suggests that the rule-based system may still play a role in the translation of these languages, given the fact that they are considerably less resource-hungry than the neural counterparts.
- In the case of Asturian, while the best systems generally maintain a similar range of differences with Apertium, there is one standout system that extends the gap significantly, achieving up to 5 BLEU and more than 3 chrF2 points higher. It is worth noting that this winning system primarily leverages a commercial large language model (LLM) through few-shot learning, sampling a new output if the LLM generates a translation that is unexpectedly short or long compared to the source. This underscores the increasing potential of cutting-edge LLMs and the implications for smaller, specialized systems, which may soon be outpaced by new models, even for low-resource languages like Asturian.

The structure of the paper is as follows. Sec. 2 provides an overview of the three target languages. Sec. 3 then outlines the different submission categories based on the resources used, whereas Sec. 4 describes the training data and resources provided to participants, as well as the development and test data used. Sec. 5 briefly describe the systems submitted within each category. The automatic evaluation results are then reported and discussed in Sec. 6. Finally, Sec. 7 concludes the paper with summarizing remarks.

2 Languages

Aragonese (Glottocode[3] `arag1245`), a Romance language mainly spoken in the Pyrenees valleys of Aragón, is primarily used in rural communities and among older generations; intergenerational transmission is severely at risk. It has around 25 000 speakers (Reyes et al., 2017, Table 5).[4] Although recognized as cultural heritage, it does not hold official status, which hampers its broader use and preservation. Despite these challenges, efforts to revitalize the language continue, supported by educational initiatives and cultural programs.

Aranese (Glottocode `aran1260`) is a variety of the Occitan language spoken in the Val d'Aran, Catalonia, where it holds official status alongside Catalan and Spanish. It is spoken by approximately 4 500 people (Generalitat de Catalunya, 2019, page 4), though its use has been declining due to the dominance of Spanish and Catalan in the region. Despite its small number of speakers, Aranese remains protected by local laws, and efforts to promote its use in education and public life are ongoing.

Asturian (Glottocode `astu1246`), another Romance language, is spoken by around 250 000 people (Llera Ramo, 2018, Figure 6) in Asturias, though it lacks official status. Like the other languages, Asturian is recognized and protected as cultural heritage, and there are efforts to increase its presence in schools and public life. Many speakers have a passive understanding of the language, and there is a strong cultural identity linked to it.

3 Submission Categories

Participants could submit their work in one of three categories,[5] depending on the corpora used, the models employed, and the reproducibility of the results: *constrained*, *open*, and *closed*.

Constrained submissions. These submissions are limited to using only the resources (corpora, dictionaries, Apertium-based systems or data, and orthographic conventions) listed in Section 4.1. Participants may also use publicly available pre-trained language or translation models, as long as their size does not exceed 1 billion parameters (1B),

[3] `https://glottolog.org`

[4] Here and in the following figures, we provide the data for the number of people who can speak the language, including those with even a basic level of proficiency.

[5] The organizers reserved the right to assign submissions to the appropriate category if there was any uncertainty.

as specified in their model cards.[6] This size restriction also applies to neural systems used for auxiliary purposes, such as generating synthetic data. The developed systems could be either bilingual or multilingual, and do not necessarily needed to cover all the target languages.

Open submissions. Submissions in this category can utilize any resources (corpora, pre-trained models, etc.) in any language, with no size restrictions, as long as the resources are publicly available under open-source licenses to ensure reproducibility. MT systems or large language models available online also fall into this category, provided that the resulting outputs are made available to the public.

Closed submissions. Closed submissions face no restrictions on the availability of resources (corpora, pre-trained models, etc.) used for training.

4 Data and Resources

This section describes the training corpora and resources provided to the participants for the *constrained* submissions (Sec. 4.1), as well as the development and test corpora used for all submission categories (Sec. 4.2).

4.1 Training Corpora and Resources

Training data. The shared task included a *constrained* submission category that restricted the resources participants could use to develop their systems, as outlined in Sec. 3. In addition to the FLORES+ dev set (see Sec. 4.2), which could be used for training or validation, participants in this category were provided with the following resources:

- Any resource from OPUS, particularly the largely uncurated resources available for Spanish–Aragonese,[7] Spanish–Occitan,[8] and Spanish–Asturian.[9] Using data from OPUS included monolingual data on the source or target sides or any other bilingual corpus.

- Data from the PILAR dataset (Galiano-Jiménez et al., 2024b), a collection of low-resource language corpora from the Iberian Peninsula. PILAR contains monolingual and parallel resources for research and development in Romance languages, with data for Aragonese (monolingual web crawled and literary texts), Aranese (bilingual Spanish–Aranese legal provisions from the Diari Oficial de la Generalitat de Catalunya, web crawled texts, and classic literary works), and Asturian (literary and popular science writings).

Systems submitted to the other categories (open and closed) could use the resources listed above, but they were not restricted to them.

Language identification. Participants also had access to tools such as Idiomata Cognitor (Galiano-Jiménez et al., 2024a), a highly accurate language identifier for the target languages and other Romance languages.[10]

Apertium data. For participants interested in integrating linguistic data into their systems or generating synthetic data, links were provided to Apertium's resources for Aragonese,[11] Spanish–Aragonese,[12] Aragonese–Catalan,[13] Spanish–Asturian,[14] Asturian,[15] Occitan–Spanish,[16] and Occitan–Catalan.[17]

Other MT systems. In addition to Apertium-based MT systems, participants were informed of other available MT systems, which could also follow different orthographic conventions to those used in the test sets: the *traduze*[18] system for Aragonese–Spanish; the *Softcatalà*[19] neural Aranese–Catalan system; and the *eslema*[20] MT system for Asturian–Spanish.

Dictionaries. Dictionaries, whether monolingual or bilingual, could serve as valuable complementary resources for participants. The following dictionaries were suggested as potential sources:

[6] For example, NLLB-200-600M, among others, meets this requirement: https://huggingface.co/facebook/nllb-200-distilled-600M.

[7] https://opus.nlpl.eu/results/es&an/corpus-result-table

[8] https://opus.nlpl.eu/results/es&oc/corpus-result-table

[9] https://opus.nlpl.eu/results/es&ast/corpus-result-table

[10] https://github.com/transducens/idiomata_cognitor

[11] https://github.com/apertium/apertium-arg

[12] https://github.com/apertium/apertium-spa-arg

[13] https://github.com/apertium/apertium-arg-cat

[14] https://github.com/apertium/apertium-spa-ast

[15] https://github.com/apertium/apertium-ast

[16] https://github.com/apertium/apertium-oci-spa

[17] https://github.com/apertium/apertium-oci-cat

[18] https://traduze.aragon.es/

[19] https://github.com/Softcatala/nmt-softcatala

[20] https://eslema.it.uniovi.es/comun/traductor.php

Diccionari der aranés[21] by Institut d'Estudis Aranesi;[22] and the *Diccionariu de la Llingua Asturiana*, available online[23] with a limit of 500 query results.

Orthographic standards. Participants were informed that the target languages have exhibited various orthographic conventions over time. The evaluation and test sets adhere to contemporary standards, as supported by their respective language academies. The following documents reflect these standards: *Normes ortogràfiques*[24] by the Academia de la Llingua Asturiana; *Ortografía de l'aragonés*[25] by the Academia Aragonesa de la Lengua; and *Gramatica der occitan aranés*[26] published by the Institut d'Estudis Aranesi.

4.2 Development and Test Data

Ad-hoc versions of the FLORES+ datasets were purposefully created for the three languages in the shared task. FLORES+ is a multilingual translation benchmark that began with a limited set of languages (Guzmán et al., 2019), was later expanded to 101 languages (Goyal et al., 2022), and most recently to 200 languages (Costa-jussà et al., 2024). In late 2023, the Open Language Data Initiative[27] (OLDI) took over leadership in extending the dataset to new languages and renamed it FLORES+. Specifically, OLDI proposed a shared task[28] to extend FLORES+ to more languages for the Ninth Conference on Machine Translation (WMT24). The Aragonese, Aranese and Asturian versions of FLORES+ used in this shared task were submitted to the OLDI's task as well.

The sentences in FLORES+ are translations of English sentences sampled equally from Wikinews (an international news source), Wikijunior (a collection of age-appropriate non-fiction books), and Wikivoyage (a travel guide). The dataset consists of a development set (dev) of 997 sentences and a development test set (devtest) of 1012 sentences.

Participants in this shared task were initially provided with the FLORES+ dev set in March 2024 and encouraged to use it for system development, as it closely mirrors the test set in terms of orthographic, grammatical, and domain aspects. Participants had a deadline of July 12, 2024, to submit translations of the Spanish side of the devtest set. Only after that deadline, was the devtest set for Aragonese, Aranese, and Asturian publicly released.

The following provides a brief overview of the FLORES+ datasets for each language, whereas a more detailed explanation of the creation process is available in the paper by Pérez-Ortiz et al. (2024).

For the Aragonese and Aranese datasets, a first draft of the dev and devtest sets were initially generated using the Spanish–Aragonese and Catalan–Aranese Apertium (Forcada et al., 2011) rule-based system. These machine translations were post-edited by language experts and then reviewed by native speakers, including members of the Academia Aragonesa de la Lengua[29] and the Institut d'Estudis Aranesi.[30] The post-editing step is justified by three factors: the lack of resources to hire qualified translators for a from-scratch translation, the common practice of post-editing for these languages, and the high degree of similarity between Spanish and these languages, which makes Apertium translations reliable and less prone to unnatural *translationese*. In the case of Asturian, professional translations originally included in FLORES-101 were reviewed twice by native speakers, including members of the Academia de la Llingua Asturiana.[31]

Pérez-Ortiz et al. (2024, Table 2) report the extent of changes made to the dev and devtest sets after both the initial and final revisions by the language academies. The data reveal significant modifications to the output of Apertium, with a TER score[32] of approximately 26% for Aragonese, 64% for Aranese, and 7% for Asturian, after the two rounds of revision.

[21] https://www.diccionari.cat/cerca/diccionari-der-aranes

[22] A PDF version can be downloaded from http://www.institutestudisaranesi.cat/wp-content/uploads/2021/04/DICCIONARI-DER-ARANÉS.pdf.

[23] https://diccionariu.alladixital.org/

[24] https://alladixital.org/wp-content/uploads/2024/01/Normes-Ortografiques-8a-edicion-FINAL-3.pdf

[25] https://academiaaragonesadelalengua.org/sites/default/files/ficheros-pdf/ortografia-aragones.pdf

[26] http://www.institutestudisaranesi.cat/wp-content/uploads/2021/04/gramatica-aranes.pdf

[27] https://oldi.org

[28] https://www2.statmt.org/wmt24/open-data.html

[29] https://academiaaragonesadelalengua.org

[30] http://www.institutestudisaranesi.cat

[31] https://www.academiadelallingua.com

[32] The translation error rate (TER) metric (Snover et al., 2006) is employed here to quantify the number of edits needed to transform the sentences from the initial versions into their corresponding counterparts in the final corpus.

5 Teams Participating in the Shared Task

We received a total of 87 submissions from 17 different teams. Table 1 lists the teams that participated in the shared task, along with the language pairs they worked on and the reference, if available, to their system description paper.

Along with their translations of the test set, participants submitted an extended abstract describing their systems and the resources used. Based on that information, we provide a brief overview of the systems developed by each of the participants.

CUNI-GA. The CUNI-GA team's contribution (Hrabal et al., 2024) for the three language pairs in the shared task involved the QLoRA fine-tuning of two open-source large language models (LLMs): Aya-23-8B and Command-R 35B. They used a small back-translated dataset, specifically the literary section of the PILAR corpus (Galiano-Jiménez et al., 2024b), which was back-translated using Apertium. Both LLMs were fine-tuned with a single joint model covering all the languages.

CycleL. The Dublin City University presented two systems (Spanish–Aragonese and Spanish–Asturian) to the constrained task (Dréano et al., 2024). They employed CycleGN, a fully self-supervised NMT framework that does not rely on parallel data. For this shared task, they exclusively used the PILAR corpus, applying sentence permutations to ensure the dataset remained non-parallel.

Helsinki-NLP. This team submitted models (de Gibert et al., 2024) exclusively to the unconstrained open track. Alongside the data provided in the task, such as PILAR, they utilized additional monolingual resources like Wikipedia dumps and dictionary definitions. To enhance their training data, they generated synthetic parallel data through back-translation using an OPUS-MT model. Their data filtering process incorporated language identification using the Idiomata Cognitor tool, as well as the OpusCleaner (Bogoychev et al., 2023) and OpusFilter (Aulamo et al., 2021) tools, to clean and refine their datasets.

For their models, Helsinki-NLP considered various initial systems, including OPUS-MT models (Tiedemann et al., 2024) and different sizes of NLLB-200 models, ranging from 600M to 3.3B parameters. They ultimately chose a multilingual OPUS-MT model based on the transformer-big architecture and produced an ensemble model after fine-tuning. Their other submissions used sequence-level distillation to train smaller student models that integrated rule-based translation. This was done by translating parallel sentences using both their neural best system and Apertium, selecting the output with the best chrF score relative to the reference, and training smaller transformer-based models on the distilled data. The sizes of distilled models ranged from the transformer *base* architecture to even smaller models obtained via the OpusDistillery tool.[33] Their different models showed statistically significant differences in most cases, except for Asturian, with the distilled models providing competitive translation performance.

HW-TSC. The Huawei Translation Service Center participated in the *constrained category* by submitting three systems, one for each of the target languages. Their submissions (Luo et al., 2024) were based on a transformer-big architecture with an expanded number of encoder layers (25). They started by training multilingual systems on sampled training data to obtain both one-to-many and many-to-one pre-trained models, which were then further trained on the original bilingual data to create translation models between Spanish and Aragonese, Aranese, and Asturian in both directions. Additionally, they utilized synthetic corpora generated via Apertium (forward translation) and through back-translation using the aforementioned multilingual models. LaBSE (Feng et al., 2022) denoising was applied to filter out noisy parallel sentences from both the provided training data and the generated synthetic data. Finally, transductive ensemble learning was employed to aggregate multiple models for inference.

ILENIA-MT. For the constrained submission (Sant et al., 2024), the team leveraged synthetic corpus generation through Apertium, primarily using data from OPUS and PILAR. Synthetic data was generated by translating from Spanish to Aragonese and Aranese (pivoting through Catalan in this case) using Apertium, while for Asturian, the team directly used NLLB-200-600M. Additional monolingual data was sourced from orthography dictionaries as supplementary resources. A comprehensive data filtering process was applied, involving the removal of noisy sentences using LABSE-based

[33] https://github.com/Helsinki-NLP/OpusDistillery

Submission Name	Language Pairs	System Description
Apertium (baseline)	Aragonese, Aranese, Asturian	(Forcada et al., 2011)
CUNI-GA	Aragonese, Aranese, Asturian	(Hrabal et al., 2024)
CycleL	Aragonese, Asturian	(Dréano et al., 2024)
Helsinki-NLP	Aragonese, Aranese, Asturian	(de Gibert et al., 2024)
HW-TSC	Aragonese, Aranese, Asturian	(Luo et al., 2024)
ILENIA-MT	Aragonese, Aranese, Asturian	(Sant et al., 2024)
imaxin	Asturian	(González, 2024)
LCT-LAP	Aragonese, Aranese, Asturian	(Bär et al., 2024)
Mora translate	Asturian	(Menan et al., 2024)
SJTU-MT	Aragonese, Aranese, Asturian	(Hu et al., 2024)
SRPH-LIT	Aragonese, Aranese, Asturian	(Velasco et al., 2024)
Stevens Inst. of Tech.	Aragonese	(no associated paper)
TAN-IBE	Aragonese, Aranese, Asturian	(Oliver, 2024)
TIM-UNIGE	Aragonese, Aranese	(Mutal and Ormaechea, 2024)
TRIBBLE	Aragonese, Aranese, Asturian	(Kuzmin et al., 2024)
UAlacant	Aragonese, Aranese, Asturian	(Galiano-Jiménez et al., 2024)
Vicomtech	Aragonese, Aranese, Asturian	(Ponce et al., 2024)
Z-AGI Labs	Aragonese, Aranese, Asturian	(no associated paper)

Table 1: Participants in the WMT24 Shared Task "Translation into Low-Resource Languages of Spain". Apertium has its own row, but it is not an actual participant; it rather serves as the baseline system.

embeddings (Feng et al., 2022), sentence length filtering, and language recognition with the Idiomata Cognitor tool. The NLLB-200-600M model was then fine-tuned with all the resulting parallel and synthetic data. To handle unsupported languages in NLLB, new language tags were added for Aragonese and Aranese, initialized with embeddings from Spanish and Occitan, respectively.

For the open submission (Sant et al., 2024), ILENIA-MT used Apertium to generate a large amount of synthetic data, translating 30 million sentences sourced from Spanish monolingual corpora. A transformer model was then trained from scratch. This approach resulted in slightly lower scores than the constrained submission.

imaxin software. This team presented an improved version of the Apertium system for the Spanish–Asturian language pair (González, 2024). The team has enhanced Apertium both in terms of syntax, by developing new constraint grammar and transfer rules, and in the lexical domain, by expanding the dictionaries.

LCT-LAP. The University of the Basque Country submitted three systems to the *constrained category* (Bär et al., 2024). These systems were obtained by fine-tuning OPUS-MT pre-trained models for two high-resource Romance languages: Spanish–Galician was used as the starting point for Spanish–Asturian, and Spanish–Catalan was used for Spanish–Aragonese and Spanish–Aranese. The fine-tuning was conducted on OPUS corpora, with noisy parallel sentences removed from the provided training data, and on synthetic corpora generated with Apertium by translating monolingual corpora in PILAR. Before utilizing the OPUS corpora, Idiomata Cognitor was employed to remove parallel sentences not in the desired language, and Apertium was then employed to translate one side of the parallel corpus, followed by BLEU scoring to filter out low-quality parallel sentences.

Mora translate. This team participated in the Spanish–Asturian language pair with a constrained submission (Menan et al., 2024). Their main contribution is a dual-stage data filtering system that combines statistical methods for both bilingual and monolingual data, along with a filtering method based on Jensen-Shannon divergence (Lin, 1991). They used the filtered CCMatrix and Wikimedia corpora, and utilized the PILAR corpus for Asturian and the Spanish portion of the English–Spanish Wikimedia corpus as monolingual data. Training was conducted in two phases: (1) training the entire model using the filtered Spanish–Asturian CCMatrix, and (2) fine-tuning the best model by unfreezing only the decoder. For fine-tuning, several datasets were combined, including monolingual data translated with NLLB-200-600M.

SJTU-MT. The systems submitted by this team (Hu et al., 2024) are based on strategies that differ significantly from traditional methods. The submissions covered all three target languages with notable variations in approach for each language pair.

For Aragonese and Aranese, the team generated a pseudo-parallel corpus using Apertium. They sampled one million Spanish sentences from the NLLB Spanish corpus in OPUS, then translated these into Aragonese and Aranese using Apertium to create a synthetic parallel corpus. Both models used a small LLM, Qwen2-0.5B,[34] which was first fine-tuned on this synthetic corpus.

For Aragonese, the model underwent an additional step of few-shot fine-tuning. This involved assembling five-shot examples using sentences from the FLORES+ dev set, providing these as context before training the model to translate sentences. At inference time, when a new sentence is inputted for translation, they use the BM25 ranking function (Robertson and Zaragoza, 2009) to identify the five most relevant examples from the FLORES+ dev set to replicate the same few-shot format introduced during training.[35]

For Aranese, after supervised fine-tuning of Qwen2 on the pseudo-corpus, an additional step involved applying the recently proposed *contrastive preference optimization* (CPO) algorithm (Xu et al., 2024). This method, which moves beyond standard training that replicates a reference translation, employs reinforcement learning loss functions to push models towards preferred translations while steering away from suboptimal ones. In order to apply CPO to their model, the team used Apertium translations as the least preferred and the target references from FLORES+ as the most preferred. Despite discouraging Apertium-like translations at times during training, this process improved the system's performance for Aranese.[36]

For Aragonese and Aranese, if the generated translations were significantly shorter or longer than the input, they were replaced with Apertium translations.

For Asturian, the team employed a completely different approach and participated in the open track. They used the large language model Claude 3.5 Sonnet,[37] utilizing a simple prompting strategy: when translating a new FLORES+ devtest sentence, they retrieved the 20 most similar examples from the FLORES+ dev set using the BM25 ranking function, providing these as suggestions to the model. If the translations produced were significantly shorter or longer than the input, rather than relying on Apertium as before, they simply resampled the model's output until achieving a translation within an acceptable length range.

SRPH-LIT. Samsung R&D Institute Philippines submitted three translation systems to the *constrained category* (Velasco et al., 2024), each addressing one of the three language pairs with a standard sequence-to-sequence transformer architecture. For each language pair, three systems were trained and combined using a noisy-channel reranking strategy to enhance output selection during decoding. The training data included filtered OPUS corpora —using ratio-based and LaBSE-based embedding methods— as well as synthetic data generated through back-translation with Apertium. Due to limited direct translation support in Apertium, translations for Aragonese–Spanish followed the path Aragonese–Catalan, Catalan–Interlingua, Interlingua–Spanish, while Aranese–Spanish followed the path Aranese–Catalan, Catalan–Spanish.

Stevens The Stevens Institute of Technology participated with a *constrained* model for Spanish–Aragonese. They leveraged NLLB-200 via a multi-stage fine-tuning process applied to both Aragonese–Spanish and Spanish–Aragonese translations. To supplement the limited available parallel data, they developed a back-translation system, generating synthetic parallel data and refining it by selecting the top 20% based on L2 cosine similarity. This iterative process enhanced both the back-translation model and the final forward translation system. The final system was an ensemble, created by averaging the weights of the two best-performing models on the development corpus.

[34] https://github.com/QwenLM/Qwen

[35] Notably, this approach leveraged the dev set not only during training but also during inference, as opposed to many other systems.

[36] This confirms that the use of post-editing of Apertium translations as an initial step in obtaining the FLORES+ data for our target languages did not overly bias the translations toward an Apertium-like style.

[37] https://www.anthropic.com/news/claude-3-5-sonnet

TAN-IBE. The TAN-IBE team (Oliver, 2024) presented systems for all language pairs in the shared task. To address the lack of resources and the low quality of existing corpora, the team: (a) cleaned the existing corpora; (b) created new corpora from Wikipedia; (c) experimented with back-translation and synthetic corpora; and (d) explored multilingual systems. All training was conducted using a transformer Marian-NMT model. For the Spanish–Asturian pair, they submitted an open system using a cleaned version of the NLLB corpus and the newly created Wikipedia corpus. For Spanish–Aragonese and Spanish–Aranese, they submitted constrained systems, using the cleaned existing corpora, the new Wikipedia corpus, and synthetic corpora generated with Apertium.

TIM-UNIGE. This team started by generating synthetic data for both forward and back-translation (Mutal and Ormaechea, 2024). They employed a two-phase synthetic data generation strategy using the BLOOMZ-560M model (Muennighoff et al., 2023) to fit within the constraints of the task. In the first phase, they fine-tuned BLOOMZ[38] on monolingual data from the target and related languages, using the task of predicting the next token. The FLORES+ dataset served as the validation set. This approach allowed the model to generate additional monolingual text, which they obtained by sampling various prefix lengths from the FLORES+ sentences and completing them through the model.

This synthetic monolingual text was then passed through Apertium to create additional parallel data, which was used to train their models. The training involved either building a transformer from scratch, using a pretrained Helsinki model (72M parameters), or the NLLB-200-600M model. A curriculum learning strategy was employed during training, where multiple phases gradually incorporated smaller subsets of higher-quality data, reduced the learning rate, and shortened the training steps. The final step involved fine-tuning exclusively on the FLORES+ dataset.

TRIBBLE. Universitat Pompeu Fabra and the Polish Academy of Science jointly submitted a model in the *constrained category* for translating into the three languages addressed in the shared task (Kuzmin et al., 2024). Their system, built on the NLLB-200-600M model, was trained on corpora sampled from OPUS and PILAR, along with synthetic data generated using Apertium. They further refined the data by utilizing Idiomata Cognitor and fastText to filter out sentence pairs in undesired languages. To reduce noise in the parallel corpus, they translated the Spanish side into the target language using Apertium and computed similarity scores based on the Levenshtein distance, discarding low-similarity bilingual sentence pairs.

UAlacant. The models submitted by Universitat d'Alacant (Galiano-Jiménez et al., 2024) use both parallel and monolingual corpora, supplemented by synthetic corpora, in the three translation directions of this task. The systems use corpora from OPUS and PILAR, together with synthetic data generated by Apertium. All submissions are classified as open due to the use of the NLLB-200-1.3B model, which exceeds the 1B parameter limit. For each translation direction, they submitted three models by fine-tuning NLLB-200.

The first approach combines parallel corpora for translation with monolingual data used in a denoising task, helping the system to learn from target language corpora even in the absence of sufficient bilingual data. A second approach introduces synthetic data, including both back-translation, where monolingual target language texts are translated into Spanish, and synthetic corpora generated by translating Spanish texts into the target languages. The third approach trains on multiple language pairs simultaneously, including Spanish, Aragonese, Asturian, Aranese and related Romance languages, such as Catalan, Galician and Valencian, using their linguistic similarities to enhance knowledge transfer and improve performance across languages. This results in a multilingual system capable of translating between all the languages.

Vicomtech. This team submitted systems for both the *constrained* and *open categories* (Ponce et al., 2024). For the *constrained category*, they exploited synthetic data generation using Apertium, like many other participants, combining it with the available parallel data. A filtering process that included language identification using Idiomata Cognitor, cross-lingual embeddings, and sentence length ratios was considered to clean the training data. Their neural translation models were built using transformer-base architectures (6 layers in the encoder and 6 in the decoder), alongside the NLLB-

[38] BLOOMZ has the characteristic of having been exposed to the FLORES datasets during its training.

Rank	id	Team	BLEU
Open submission			
1	636	ILENIA-MT	62.7
	637	ILENIA-MT	62.6
2	—	Apertium	61.1
	633	Vicomtech	61.0
3	554	UAlacant	60.2
4	495	UAlacant	59.8
5	577	Helsinki-NLP	52.7
6	549	Helsinki-NLP	51.5
7	563	Helsinki-NLP	50.6
8	523	Helsinki-NLP	49.1
9	548	UAlacant	37.8
10	647	CUNI-GA	36.1
Constrained submission			
1	663	SJTU-MT	63.2
	607	HW-TSC	63.0
2	529	ILENIA-MT	62.3
	558	ILENIA-MT	62.2
	634	ILENIA-MT	62.2
	642	TIM-UNIGE	61.9
3	526	ILENIA-MT	61.6
	—	Apertium	61.1
	504	Vicomtech	61.1
	644	TIM-UNIGE	61.1
4	586	TIM-UNIGE	60.7
	539	TIM-UNIGE	60.5
5	649	Stevens	59.8
6	613	Stevens	57.5
	584	TAN-IBE	57.3
7	622	TRIBBLE	49.2
8	530	LCT-LAP	38.9
	662	Stevens	38.7
	513	Stevens	37.5
9	507	SRPH-LIT	28.2
10	534	Z-AGI Labs	24.3
11	533	Z-AGI Labs	22.1
12	591	CycleL	0.2

Table 2: BLEU scores computed over the FLORES+ devtest set for Spanish–Aragonese.

200-600M model, finding the non-pretrained neural models to work slightly better.

For the *open category*, a key highlight of their approach, as emphasized by the authors, was the use of the Llama3-8B LLM (Dubey, 2024) to generate synthetic data in the reverse direction, i.e., from the target low-resource languages into Spanish. This approach allowed their MT systems to exploit whatever knowledge of the target languages the LLMs might have, even if this knowledge is likely limited due to exposure to only small amounts of text. Their *open category* models combined data from Apertium, other NMT systems, and LLM-generated data, resulting in slightly better scores for Asturian over the constrained models.

Rank	id	Team	BLEU
Open submission			
1	—	Apertium	28.8
	627	Vicomtech	28.8
2	555	UAlacant	28.5
3	587	ILENIA-MT	27.3
	656	CUNI-GA	27.1
	552	UAlacant	27.0
4	578	Helsinki-NLP	24.3
5	562	Helsinki-NLP	22.4
6	550	Helsinki-NLP	22.1
7	524	Helsinki-NLP	21.6
Constrained submission			
1	621	SJTU-MT	30.4
	641	TIM-UNIGE	30.2
	527	ILENIA-MT	30.1
	619	TIM-UNIGE	30.1
	617	TIM-UNIGE	30.0
	640	TIM-UNIGE	29.9
	625	Vicomtech	29.8
2	575	TIM-UNIGE	28.9
	—	Apertium	28.8
3	494	TIM-UNIGE	28.2
4	610	TAN-IBE	26.9
5	608	HW-TSC	26.3
6	623	TRIBBLE	23.9
7	531	LCT-LAP	21.8
8	581	SRPH-LIT	7.7
9	536	Z-AGI Labs	3.8
10	535	Z-AGI Labs	3.7

Table 3: BLEU scores computed over the FLORES+ devtest set for Spanish–Aranese.

Z-AGI Labs. This team participated in all language pairs of the shared task. They fine-tuned the NLLB and Helsinki-NLP/OpusMT models using the OPUS dataset provided on the shared task website.

6 Results and Discussion

We measured the translation quality of the different systems submitted to the shared task when translating the FLORES+ devtest dataset by means of

BLEU (Papineni et al., 2002)[39] and chrF2 (Popović, 2015).[40]. We did not use neural-based metrics, such as COMET (Rei et al., 2020), as they are not available for the target languages. Neither did we conduct a manual evaluation because of the lack of resources to hire qualified translators.

Rank	id	Team	BLEU
Open submission			
1	576	SJTU-MT	23.2
2	551	Helsinki-NLP	18.2
	564	Helsinki-NLP	18.0
	579	Helsinki-NLP	18.0
	568	TAN-IBE	18.0
	629	Vicomtech	18.0
3	525	Helsinki-NLP	17.9
4	553	UAlacant	17.4
5	—	Apertium	17.0
	556	UAlacant	16.9
	497	UAlacant	16.8
6	632	ILENIA-MT	16.7
7	648	CUNI-GA	15.2
Constrained submission			
1	606	HW-TSC	19.8
2	528	ILENIA-MT	18.4
3	624	TRIBBLE	17.9
	547	Mora translate	17.6
	630	Vicomtech	17.6
	532	LCT-LAP	17.5
	546	SRPH-LIT	17.5
	522	Mora translate	17.4
	590	Mora translate	17.4
	519	Mora translate	17.4
	512	Mora translate	17.4
	543	Mora translate	17.3
4	—	Apertium	17.0
5	538	Z-AGI Labs	7.6
6	537	Z-AGI Labs	6.4
7	597	CycleL	0.1
Closed submission			
1	580	imaxin	17.6
2	—	Apertium	17.0

Table 4: BLEU scores computed over the FLORES+ devtest set for Spanish–Asturian.

As already mentioned in the introduction, the three language pairs have an Apertium MT system available; we therefore include Apertium among the systems evalauted in this section. The specific versions of Apertium used for each language are: Spanish–Aragonese 0.6.0,[41] Spanish–Aranese 1.0.8,[42] Spanish–Asturian 1.1.1.[43]

Tables 2, 3 and 4 show the BLEU scores attained by each system for Aragonese, Aranese and Asturian, respectively. Similarly, tables 5, 6 and 7 show the results obtained with chrF2. Each table reports, in addition to the BLEU or chrF2 scores, a ranking of the systems from best (#1) to worse. This ranking was derived using a statistical significance test conducted through paired approximate randomization (Riezler and Maxwell, 2005) with SacreBLEU.[44] Systems within the same rank do not exhibit statistically significant differences.

The ranking process involved an iterative approach. We began by selecting the best system for each metric as the control translation. The translations provided by other systems were then compared to this control to determine if the differences were statistically significant. Systems whose output did not differ significantly from the control were associated with it and removed from the pool of translations. The next best system then became the new control translation, and the process was repeated. This iterative process continued until no system remained in the pool.

In the **Spanish–Aragonese** translation task (Tables 2 and 5), the *open* submission results show the ILENIA-MT team achieving the highest BLEU score of 62.7. Notably, ILENIA-MT's performance is consistent, as their second submission scores nearly identical at 62.6. The Apertium baseline and the submission by Vicomtech closely follow, with BLEU scores of 61.1 and 61.0, respectively. UAlacant's entries, which rank 3rd and 4th with scores of 60.2 and 59.8, demonstrate strong competitiveness as well, outperforming the submissions from Helsinki-NLP, which rank lower.

In the *constrained* submission category, SJTU-MT and HW-TSC lead with BLEU scores of 63.2 and 63.0, respectively, surpassing the top scores from the *open* submissions —a difference that is statistically significant—. SJTU-MT's approach

[39] SacreBLEU BLEU signature: `nrefs:1 | case:mixed | eff:no | tok:13a | smooth:exp | version:2.0.0`

[40] SacreBLEU chrF2 signature: `nrefs:1 | case:mixed | eff:yes | nc:6 | nw:0 | space:no | version:2.0.0`

[41] https://github.com/apertium/apertium-spa-arg/releases/tag/v0.6.0

[42] https://github.com/apertium/apertium-oc-es/releases/tag/v1.0.8

[43] https://github.com/apertium/apertium-spa-ast/releases/tag/v1.1.1

[44] https://github.com/mjpost/sacrebleu

Rank	id	Team	chrF2
Open submission			
1	636	ILENIA-MT	80.0
	637	ILENIA-MT	80.0
2	—	Apertium	79.3
	633	Vicomtech	79.3
3	554	UAlacant	78.9
4	495	UAlacant	78.8
5	577	Helsinki-NLP	75.9
6	549	Helsinki-NLP	75.6
7	563	Helsinki-NLP	75.4
8	523	Helsinki-NLP	74.6
9	548	UAlacant	67.5
10	647	CUNI-GA	67.8
Constrained submission			
1	607	HW-TSC	80.3
	663	SJTU-MT	80.1
	529	ILENIA-MT	79.9
	558	ILENIA-MT	79.9
	634	ILENIA-MT	79.9
2	526	ILENIA-MT	79.5
	642	TIM-UNIGE	79.5
	—	Apertium	79.3
	504	Vicomtech	79.3
3	586	TIM-UNIGE	79.0
	539	TIM-UNIGE	79.0
	644	TIM-UNIGE	79.0
	649	Stevens	78.7
4	584	TAN-IBE	78.1
5	613	Stevens	77.2
6	622	TRIBBLE	73.6
7	530	LCT-LAP	68.6
8	513	Stevens	67.4
9	662	Stevens	62.0
	534	Z-AGI Labs	61.8
10	533	Z-AGI Labs	60.6
11	507	SRPH-LIT	58.4
12	591	CycleL	13.7

Table 5: chrF2 computed over the FLORES+ devtest set for Spanish–Aragonese.

stands out as one of the most innovative in the task, employing strategies that diverge significantly from traditional methods, whereas HW-TSC used the largest number of layers in the encoder (25) of all the submissions. ILENIA-MT continues to perform strongly, with their top entry scoring 62.3, closely followed by TIM-UNIGE, Apertium and Vicomtech.

Rank	id	Team	chrF2
Open submission			
1	—	Apertium	49.4
	627	Vicomtech	49.4
2	555	UAlacant	49.3
3	587	ILENIA-MT	48.8
	656	CUNI-GA	48.5
	552	UAlacant	48.3
4	578	Helsinki-NLP	46.6
5	562	Helsinki-NLP	45.7
6	550	Helsinki-NLP	45.1
7	524	Helsinki-NLP	45.0
Constrained submission			
1	527	ILENIA-MT	50.1
	621	SJTU-MT	49.9
	619	TIM-UNIGE	49.8
	617	TIM-UNIGE	49.7
	625	Vicomtech	49.8
2	641	TIM-UNIGE	49.6
	—	Apertium	49.4
	640	TIM-UNIGE	49.3
	575	TIM-UNIGE	49.2
3	494	TIM-UNIGE	48.8
	610	TAN-IBE	48.8
4	608	HW-TSC	47.9
5	623	TRIBBLE	46.1
6	531	LCT-LAP	45.5
7	581	SRPH-LIT	34.8
8	536	Z-AGI Labs	32.8
9	535	Z-AGI Labs	31.9

Table 6: chrF2 computed over the FLORES+ devtest set for Spanish–Aranese.

For the **Spanish–Aranese** pair (Tables 3 and 6), the *open* submission category presents a narrower range of BLEU scores compared to Aragonese. Apertium and Vicomtech share the top position with a BLEU score of 28.8, closely followed by UAlacant with 28.5.

In the *constrained submission* category, SJTU-MT once again leads, achieving a BLEU score of 30.4. This time, several other teams — TIM-UNIGE, ILENIA-MT, and Vicomtech— join SJTU-MT at the top, all with scores outperforming the *open* submissions by a statistically significant margin.

The lower BLEU scores for the **Spanish–Asturian** language pair (Tables 4 and 7), are likely due to the way the FLORES+ dev and devtest datasets were constructed, with translations originating from English rather than Spanish. In

Rank	id	Team	chrF2
Open submission			
1	576	SJTU-MT	55.2
2	551	Helsinki-NLP	51.6
	564	Helsinki-NLP	51.6
	579	Helsinki-NLP	51.5
	568	TAN-IBE	51.6
	629	Vicomtech	51.6
3	525	Helsinki-NLP	51.4
4	556	UAlacant	50.9
	497	UAlacant	50.9
	—	Apertium	50.8
	553	UAlacant	50.7
5	632	ILENIA-MT	50.5
6	648	CUNI-GA	48.9
Constrained submission			
1	606	HW-TSC	52.2
	528	ILENIA-MT	52.1
2	547	Mora translate	51.4
	630	Vicomtech	51.2
	519	Mora translate	51.2
	590	Mora translate	51.2
	522	Mora translate	51.0
	512	Mora translate	51.0
	543	Mora translate	51.0
	—	Apertium	50.8
3	532	LCT-LAP	50.7
	624	TRIBBLE	50.5
	546	SRPH-LIT	50.0
4	538	Z-AGI Labs	44.4
5	537	Z-AGI Labs	42.7
6	597	CycleL	15.9
Closed submission			
1	580	imaxin	51.2
2	—	Apertium	50.8

Table 7: chrF2 computed over the FLORES+ devtest set for Spanish–Asturian.

the *open* submission category, SJTU-MT leads with a score of 23.2, significantly outperforming the second-best system, Helsinki-NLP, by 5 BLEU points, with the latter's scores clustering around 18.0.

The *constrained* submission results show HW-TSC leading with a BLEU score of 19.8, followed by ILENIA-MT at 18.4. Despite the constrained environment, HW-TSC's results indicate that their extensive use of encoder layers and synthetic data generation proved beneficial.

7 Conclusions

This paper has presented the outcomes of the Ninth Conference on Machine Translation (WMT24) Shared Task on Translation into Low-Resource Languages of Spain. The challenge centred on building MT systems for three Romance language pairs: Spanish–Aragonese, Spanish–Aranese, and Spanish–Asturian. In total, 17 teams took part in this shared task.

Across all three language pairs, there is some variability in performance both between and within the categories (open, constrained, and closed). Top-performing teams such as SJTU-MT, ILENIA-MT, and HW-TSC consistently achieved high rankings across multiple pairs. The results also underscore the challenges posed by low-resource languages, where factors such as data availability and the choice of methods —e.g., synthetic data generation, fine-tuning strategies, or transformer model size— significantly affect performance. Notably, most of the best-performing systems utilized the Apertium rule-based system to generate synthetic data, highlighting the ongoing relevance of these approaches in complementing neural methods.

Acknowledgements

This task would have not been possible without the support and help in data acquisition and in the translation of the FLORES+ sentences from Academia Aragonesa de la Lengua (Instituto de l'Aragonés), Academia de la Llingua Asturiana and Institut d'Estudis Aranesi.

This shared task is part of the R+D+i projects PID2021-127999NB-I00 (*LiLowLa: Lightweight neural translation technologies for low-resource languages*) and PID2021-124663OB-I00 (*TAN-IBE: Neural Machine Translation for the Romance Languages of the Iberian Peninsula*), founded by MCIN /AEI /10.13039/501100011033 / FEDER, UE.

References

Mikko Aulamo, Sami Virpioja, Yves Scherrer, and Jörg Tiedemann. 2021. Boosting neural machine translation from Finnish to Northern Sámi with rule-based backtranslation. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 351–356, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Martin Bär, Elisa Forcada Rodríguez, and María García-Abadillo Velasco. 2024. Robustness of fine-tuned

LLMs for machine translation with varying noise levels: Insights for Asturian, Aragonese and Aranese. In *Proceedings of the Ninth Conference on Machine Translation*, Miami. Association for Computational Linguistics.

Nikolay Bogoychev, Jelmer van der Linde, Graeme Nail, Barry Haddow, Jaume Zaragoza-Bernabeu, Gema Ramírez-Sánchez, Lukas Weymann, Tudor Nicolae Mateiu, Jindřich Helcl, and Mikko Aulamo. 2023. Opuscleaner and opustrainer, open source toolkits for training machine translation and large language models.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Jeff Wang, and NLLB Team. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Ona de Gibert, Mikko Aulamo, Yves Scherrer, and Jörg Tiedemann. 2024. Hybrid distillation from RBMT and NMT: Helsinki-NLP's submission to the Shared Task on Translation into Low-Resource Languages of Spain. In *Proceedings of the Ninth Conference on Machine Translation*, Miami. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sören Dréano, Derek Molloy, and Noel Murphy. 2024. Exploration of the CycleGN framework for low-resource languages. In *Proceedings of the Ninth Conference on Machine Translation*, Miami. Association for Computational Linguistics.

Abhimanyu et al. Dubey. 2024. The Llama 3 herd of models.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

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(2):127–144.

Aarón Galiano-Jiménez, Víctor M Sánchez-Cartagena, Juan Antonio Pérez-Ortiz, and Felipe Sánchez-Martínez. 2024. Universitat d'Alacant's submission to the WMT 2024 Shared Task on Translating into Low-Resource Languages of Spain. In *Proceedings of the Ninth Conference on Machine Translation*, Miami. Association for Computational Linguistics.

Aarón Galiano-Jiménez, Felipe Sánchez-Martínez, and Juan Antonio Pérez-Ortiz. 2024a. Idiomata cognitor.

Aarón Galiano-Jiménez, Felipe Sánchez-Martínez, and Juan Antonio Pérez-Ortiz. 2024b. Pan-Iberian Language Archival Resource.

Generalitat de Catalunya. 2019. *Els usos lingüístics de la població de l'Aran: Principals resultats de l'Enquesta d'usos lingüístics de la població. 2018*. Generalitat de Catalunya, Barcelona.

Sofía García González. 2024. Enhanced Apertium system: Translation into low-resource languages of Spain Spanish–Asturian. In *Proceedings of the Ninth Conference on Machine Translation*, Miami. Association for Computational Linguistics.

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.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Miroslav Hrabal, Josef Jon, Martin Popel, Nam H Luu, Danil Semin, and Ondřej Bojar. 2024. CUNI at WMT24 general translation task: LLMs, (Q)LoRA, CPO and model merging. In *Proceedings of the Ninth Conference on Machine Translation*, Miami. Association for Computational Linguistics.

Tianxiang Hu, Haoxiang Sun, Ruize Gao, Jialong Tang, Pei Zhang, Baosong Yang, and Rui Wang. 2024. SJTU system description for the WMT24 Low-Resource Languages of Spain task. In *Proceedings of the Ninth Conference on Machine Translation*, Miami. Association for Computational Linguistics.

Igor Kuzmin, Piotr Przybyła, Euan McGill, and Horacio Saggion. 2024. TRIBBLE - TRanslating IBerian languages Based on Limited E-resources: System description. In *Proceedings of the Ninth Conference on Machine Translation*, Miami. Association for Computational Linguistics.

J. Lin. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pretraining for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Francisco J. Llera Ramo. 2018. *III Estudio Sociolingüístico de Asturias 2017: Avance de resultados*. Academia de la Llingua Asturiana, Uvieu. Estaya Sociollingüística, colección 7.

Yuanchang Luo, Zhanglin Wu, Daimeng Wei, Hengchao Shang, Zongyao Li, Jiaxin Guo, and Zhiqiang et al. Rao. 2024. Multilingual transfer and domain adaptation for low-resource languages of Spain. In *Proceedings of the Ninth Conference on Machine Translation*, Miami. Association for Computational Linguistics.

Velayuthan Menan, Dilith Jayakody, Nisansa de Silva, Aloka Fernando, and Surangika Ranatunga. 2024. Back to the stats: Rescuing low resource neural machine translation with statistical methods. In *Proceedings of the Ninth Conference on Machine Translation*, Miami. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Jonathan Mutal and Lucía Ormaechea. 2024. TIM-UNIGE translation into low-resource languages of Spain for WMT24. In *Proceedings of the Ninth Conference on Machine Translation*, Miami. Association for Computational Linguistics.

Antoni Oliver. 2024. TAN-IBE participation in the Shared Task: Translation into Low-Resource Languages of Spain. In *Proceedings of the Ninth Conference on Machine Translation*, Miami. Association for Computational Linguistics.

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.

Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Víctor M. Sánchez-Cartagena, Miquel Esplà-Gomis, Aarón Galiano-Jiménez, Antoni Oliver, Claudi Aventín-Boya, Cristina Valdés, Alejandro Pardos, and Juan Pablo Martínez. 2024. FLORES+ datasets for Aragonese, Aranese, Asturian and Valencian. In *Proceedings of the Ninth Conference on Machine Translation*, pages 00–00, Miami. Association for Computational Linguistics.

David Ponce, Harritxu Gete, and Thierry Etchegoyhen. 2024. Vicomtech@WMT 2024: Shared Task on Translation into Low-Resource Languages of Spain. In *Proceedings of the Ninth Conference on Machine Translation*, Miami. 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.

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

Anchel Reyes, Chabier Gimeno, Miguel Montañés, Natxo Sorolla, Pep Espluga, and Juan Pablo Martínez. 2017. *L'aragonés y lo catalán en l'actualidat: analisi d'o censo de población y viviendas de 2011*. Seminario Aragonés de Sociolingüística, Asociación Aragonesa de Sociolochía, Universidad de Zaragoza. Primera parte, febrero 2017.

Stefan Riezler and John T. Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 57–64, Ann Arbor, Michigan. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Aleix Sant, Daniel Bardanca Outeiriño, José Ramom Pichel Campos, Francesca De Luca Fornaciari, Carlos Escolano, Javier García Gilabert, Pablo Gamallo Otero, Audrey Mash, Xixian Liao, and Maite Melero. 2024. Training and fine-tuning NMT models for low-resource languages using Apertium-based synthetic corpora. In *Proceedings of the Ninth Conference on Machine Translation*, Miami. Association for Computational Linguistics.

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.

Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato, Yves Scherrer, Raúl Vázquez, and Sami Virpioja. 2024. Democratizing neural machine translation with opus-mt. *Language Resources and Evaluation*, 58(2):713–755.

Dan John Velasco, Manuel Antonio Rufino, and Jan Christian Blaise Cruz. 2024. Samsung R&D Institute Philippines @ WMT 2024 Low-Resource Languages of Spain Shared Task. In *Proceedings of the Ninth Conference on Machine Translation*, Miami. Association for Computational Linguistics.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 55204–55224. PMLR.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.