Training and Fine-Tuning NMT Models for Low-Resource Languages using Apertium-Based Synthetic Corpora

Aleix Sant[1], **Daniel Bardanca Outeiriño**[2], **José Ramom Pichel Campos**[2]
Francesca De Luca Fornaciari[1], **Carlos Escolano**[1], **Javier García Gilabert**[1]
Pablo Gamallo Otero[2], **Audrey Mash**[1], **Xixian Liao**[1], **Maite Melero**[1]
[1] Barcelona Supercomputing Center (BSC),
[2] Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS)
Universidade de Santiago de Compostela
{aleix.santsavall, francesca.delucafornaciari, carlos.escolano
javier.garcia1, audrey.mash, xixian.liao, maite.melero}@bsc.es
{daniel.bardanca, jramon.pichel, pablo.gamallo}@usc.gal

Abstract

In this paper, we present the two strategies employed for the WMT24 Shared Task on Translation into Low-Resource Languages of Spain. We participated in the language pairs of Spanish-to-Aragonese, Spanish-to-Aranese, and Spanish-to-Asturian, developing neural-based translation systems and moving away from rule-based approaches for these language directions. To create these models, two distinct strategies were employed. The first strategy involved a thorough cleaning process and curation of the limited provided data, followed by fine-tuning the multilingual NLLB-200-600M model (Constrained Submission). The other strategy involved training a transformer from scratch using a vast amount of synthetic data (Open Submission). Both approaches relied on generated synthetic data and resulted in high ChrF and BLEU scores. However, given the characteristics of the task, the strategy used in the Constrained Submission resulted in higher scores that surpassed the baselines across the three translation directions, whereas the strategy employed in the Open Submission yielded slightly lower scores than the highest baseline.

1 Introduction

This article presents the work done by the ILENIA team, which includes researchers from the Barcelona Supercomputing Center and Proxecto Nós (CiTIUS - Universidade de Santiago de Compostela), for the WMT24 Shared Task on Translation into Low-Resource Languages of Spain[1]. Our participation covered three translation directions: Spanish-to-Aragonese, Spanish-to-Aranese, and Spanish-to-Asturian, all of which are Romance languages.

Aragonese is spoken in several valleys of the Pyrenees in the autonomous community of Aragon. It is one of Europe's smallest language communities, with around 8,500 native speakers and 25,000 total speakers. According to UNESCO, Aragonese is an increasingly endangered language (Moseley, 2010).

Aranese is spoken in Vall d'Aran, in the northwest of Catalonia. It is the native language of this unique region, with approximately 5,090 native speakers. Aranese is a variant of Gascon, one of the main dialects of the Occitan language.

Asturian is the variant of Astur-Leonese spoken in the autonomous community of Asturias, in northern Spain. Currently, around 250,000 people have the ability to understand, speak, read, and write Asturian, representing the 25% of this autonomous community.

For this Shared Task, we participated in two types of submissions. In the **Constrained Submission**, we were allowed to use the specified resources, such as corpora, dictionaries, Apertium-based systems, and documents defining the contemporary orthographic conventions for each language. Regarding the models, we could use publicly available models, provided they did not exceed 1 billion parameters. To meet these requirements, we collected and generated synthetic data from the available resources, built and applied a comprehensive cleaning pipeline to preprocess the data, and fine-tuned three separate NLLB-200-600M (Costa-jussà et al., 2022) on the corresponding Spanish-to-Aragonese, Spanish-to-Aranese, and Spanish-to-Asturian translation directions. BSC researchers conducted the experiments for this submission.

For the **Open Submission**, we were allowed to use any publicly available resources, including corpora and models of any size, as long as the resulting

[1] https://www2.statmt.org/wmt24/romance-task.html

Proceedings of the Ninth Conference on Machine Translation (WMT), pages 925–933
November 15–16, 2024.

outputs were made available. For this submission, we chose to generate large amounts of synthetic corpora from data released by ILENIA and the Proxecto Nós using Apertium, a rule-based translator (Khanna et al., 2021). We then trained three models based on the transformer architecture (Vaswani et al., 2017) from scratch using OpenNMT-py 3.2[2] (Klein et al., 2018). The researchers from Proxecto Nós were responsible for this second submission.

2 Data

2.1 Data Collection and Synthetic Creation

2.1.1 Constrained Submission

OPUS The organizers created the Aragonese and Aranese FLORES+ dev and devtest sets using Apertium, translating the corresponding Spanish texts from the FLORES-200 multilingual dataset (NLLB Team, 2022) into these languages. We consider this to be a limitation of the Shared Task since the reference test sets are biased towards Apertium-generated data. Nevertheless, in order to achieve the highest possible score on our submission for the Shared Task, we decided to use Apertium to generate synthetic Aragonese and Aranese translations from Spanish monolingual data, instead of directly using the parallel data provided by OPUS (Nygaard and Tiedemann, 2003). Specifically, we used the Spanish side of the es-arg and es-oc parallel corpora from OPUS to generate synthetic data. For Aragonese, we used the GNOME, Ubuntu, Wikimatrix, and Wikimedia corpora, and for Aranese, we used these same corpora in addition to Kde4 and NLLB.

For Asturian, we did not generate any synthetic data from OPUS since the Asturian FLORES+ dev and devtest sets were simply enhanced versions of the original FLORES-200. Instead, we downloaded the following es-ast parallel data: GNOME, Kde4, NLLB, Tatoeba, Ubuntu, and Wikimedia.

PILAR[3] We generated synthetic Spanish translations from the monolingual data provided by the organizers in the three respective Romance languages (Galiano-Jiménez et al., 2024b). For Aragonese and Aranese, we used Apertium, while for Asturian, we employed NLLB-200-600M, which fell within the submission's limits. Given the similarity between Aranese and Aragonese to Catalan, we explored whether cascading through Catalan could enhance translation quality. In machine translation, cascading refers to the sequential use of multiple translation systems to improve overall translation accuracy. As demonstrated in Table 1, this method yielded higher scores when translating from Aranese. Consequently, we generated synthetic Spanish translations from Aranese by cascading through Catalan, while for Aragonese and Asturian, we produced the translations directly into Spanish.

	Aragonese → Spanish		Aranese → Spanish	
	ChrF	**BLEU**	**ChrF**	**BLEU**
`Direct`	**80.93**	**66.09**	68.79	45.02
`Cascade`	79.41	63.00	**69.67**	**47.29**

Table 1: Scores obtained with and without cascading through Catalan for FLORES+ dev test.

Provided PDFs We extracted monolingual data in Aragonese and Aranese from the provided `ortografia-aragones.pdf` and `DICCIONARI-DER-ARANÉS.pdf` respectively. After the text extraction, we semi-automatically post-processed the data to obtain a structured and clean corpus. Then we generated the corresponding Spanish translations using Apertium following the same method described in the previous paragraph.

FLORES+ It consists of an extension of FLORES-200 (NLLB Team, 2022), a multilingual English-centric machine translation dataset involving 200 languages, that includes Aragonese, Aranese, and an improved version of Asturian. The FLORES+ dev set for each language served as the validation set during the training phase to optimize our MT engines, while the devtest set was used to evaluate participants' models in the competition. After the final submissions, the devtest set was released, allowing us to obtain scores for the baseline models in this additional set.

2.1.2 Open Submission

For this submission, we used Apertium to generate synthetic data. We created synthetic datasets using a parallel corpus of 30M Galician-Spanish sentence pairs. We only kept the Spanish side of the corpus as the source language data, and then used Apertium to translate it into Aragonese and Asturian, resulting in two 30M sentence parallel corpora (Spanish-Aragonese and Spanish-Asturian).

[2] https://github.com/OpenNMT/OpenNMT-py
[3] https://github.com/transducens/PILAR

In the case of Aranese, we used a high-quality 30M sentence Spanish-Catalan parallel dataset. We translated the Catalan side into Aranese using Apertium, creating a Spanish-Aranese corpus.

2.2 Data Preprocessing and Cleaning

2.2.1 Constrained Submission

For this submission, we dedicated substantial effort to cleaning, curating, and normalizing the provided data. We designed a comprehensive cleaning pipeline that processed all the parallel data described in the previous section, resulting in well-structured parallel corpora for the Spanish-Aragonese, Spanish-Aranese, and Spanish-Asturian language pairs.

Following the automatic cleaning, we curated the resulting data to ensure it aligned to the orthographic standards outlined in the task statement and matched the characteristics of the corresponding FLORES+ sets for each language.

Blank Spaces, Hard- and Soft-Duplicates Removal The initial step involved removing any unnecessary blank spaces and exact duplicates within the corpus. Then, NLPDedup[4] was used to remove near duplicates.

Idiomata Cognitor[5] This language identifier (Galiano-Jiménez et al., 2024a), specifically designed for certain Romance languages, was employed to accurately determine the languages of each data pair and exclude pairs with sentences belonging to other languages. This method ensures that the translator model is trained on appropriate data.

Perl Corpus Cleaner We employed Moses (Koehn et al., 2007) preprocessing script `clean-corpus-n.perl` to further clean the parallel corpus. It eliminates sentences containing more than 150 tokens and discards sentence pairs with a length ratio exceeding 3.

Linguistic Data Normalization Since the released corpora included text in various orthographies, and both the FLORES+ dev and devtest sets adhered to the current standards endorsed by their respective language academies, we ensured that our training data conformed to these established norms through a normalization process. This process was carried out semi-automatically: incorrect patterns in the data were detected and replaced with the correct ones according to the relevant linguistic rules. This normalization was primarily applied to the Aragonese and Aranese monolingual data. For example, in Aragonese, we encountered different types of definite articles used interchangeably, such as "o"/"lo", "a"/"la", "os"/"los", and "as"/"las", as well as various ways to write the word "university", including "unibersidad", "unibersidá", and "universidat". In the case of definite articles, all forms needed to be standardized to "lo"/"la"/"los"/"las", except when following a word ending in 'n', where the forms are "o"/"a"/"os"/"as". For the term "university", the officially accepted word is "universidat".

Data Curation Using the FLORES+ dev set as our reference, we further examined the parallel data to identify misleading translations in the training data. This curation process involved both semi-automatic and manual methods, primarily focusing on the word level. For example, Apertium often leaves unknown words unchanged in the target sentence or produces incorrect translations due to insufficient contextual understanding of the source sentence. These are common behaviors of rule-based translators. We aimed to detect these issues and correct these translation errors.

According to Table 2, a large number of sentence pairs are discarded, mainly due to the high volume of duplicates in the three corpora, with Aranese and, particularly, Asturian exhibiting the highest number of duplicates.

	Aragonese	Aranese	Asturian
`Original`	74,014	1,336,229	6,603,733
`Filtered`	47,521	407,397	704,933

Table 2: Parallel corpus statistics per target language. `Original` refers to all the collected data pairs before going through the pipeline. `Filtered` refers to the number of pairs resulting from the data cleaning pipeline.

LaBSE scoring To evaluate the quality of translations in the parallel datasets, we used a sentence embedding model. Specifically, we employed LaBSE (Feng et al., 2022) to generate embeddings for both source and target sentences and calculated the co-

[4] https://github.com/saattrupdan/NLPDedup

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

		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	σ
Aragonese	**ChrF**	84.6	84.59	84.59	84.61	84.59	**84.63**	84.58	84.62	84.5	0.6752
	BLEU	71.05	71.05	71	71.09	71.03	**71.14**	70.99	71.05	70.84	0.0832
Aranese	**ChrF**	75.95	**76.04**	76.02	75.99	75.97	75.97	75.91	75.84	75.32	0.2218
	BLEU	55.39	55.5	**55.57**	55.43	55.39	55.43	55.3	55.2	54.39	0.3535
Asturian	**ChrF**	52.25	52.25	**52.26**	52.25	52.21	52.19	52.21	52.22	52.19	0.1897
	BLEU	**19.35**	19.34	19.33	19.32	19.28	19.23	19.26	19.27	19.24	0.1737

Table 3: Scores for different LaBSE thresholds on FLORES+ dev set.

sine similarity score between them. Following the approach outlined in (Garcia Gilabert et al., 2024), we obtained scores across various thresholds of cosine similarity to determine the most suitable training dataset for fine-tuning (Table 3). However, the low standard deviation among the results indicates no considerable differences in performance between the sets. We selected a threshold of 0.6 for Aragonese, a threshold of 0.3 for Aranese, and a threshold of 0.1 for Asturian, prioritizing a higher BLEU over a ChrF.

2.2.2 Open Submission

No extra preprocessing or cleaning was performed on the synthetic corpora generated with Apertium for this submission. The source data had previously been processed before publication with their own pipeline [6]. This prepossessing includes fixing encoding issues, deduplication, perplexity filtering and language recognition.

3 Methodology

3.1 Baselines

In Table 4, ChrF and BLEU scores on the FLORES+ dev set for the state-of-the-art models including the directions of interest are shown. Except for the NLLB models, which employ deep learning, all the other evaluated engines are rule-based.

3.2 Constrained Submission: Fine-tuning

3.2.1 Model

NLLB-200-600M is the smallest model from the NLLB family of multilingual machine translation models. It is a dense transformer model distilled from the pre-trained NLLB-200, a 54.5B sparsely gated mixture-of-experts model, designed to support translations between 202 languages, including many low-resource languages. Therefore, it incorporates substantial cross-lingua knowledge, making it suitable for further fine-tuning to other languages. It has a vocabulary size of 256k tokens, plus additional tokens for the language tags corresponding to all the languages supported by the model. With respect to our languages of interest, it just handles Asturian. Occitan is also in the list of languages, but not the Aranese variant.

3.2.2 Fine-tuning

For this approach, we fully fine-tuned three separate NLLB-200-600M models for Spanish-to-Aranese, Spanish-to-Aragonese, and Spanish-to-Asturian, leveraging the cross-lingua knowledge NLLB possesses.

Adding New Language Tags Among the languages of interest, NLLB only supports Asturian natively. To extend NLLB's translation capabilities to translate to Aragonese and to Aranese, we incorporated new tokens referring to their respective language tags (`arn_Latn` and `arg_Latn`), since these languages are not present in NLLB[7]. Language tags enable NLLB to identify the source and target languages for translation. Adding new language tags implies extending the embedding matrix with additional embeddings. These new embeddings were initialized using the embeddings of other language tags already supported by NLLB, which were linguistically close to our target language. Specifically, we used the `spa_Latn` embedding for Aragonese and the `oci_Latn` embedding for Aranese. Finally, we retrained the embedding matrix during the fine-tuning to enable correct Spanish-to-Aragonese and Spanish-to-Aranese translation. For Asturian, since NLLB already supports translation to this language, our objective was simply to improve the model's performance

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

[7] Aranese is a variant of Occitan, but due to observed differences in the test sets, we treated Aranese as a distinct language with its own language tag.

		Spanish → Aragonese		Spanish → Aranese		Spanish → Asturian	
		ChrF	**BLEU**	**ChrF**	**BLEU**	**ChrF**	**BLEU**
Baselines	Apertium	82	65.34	72.63	48.96	50.57	16.66
	Traduze	69.51	37.43	-	-	-	-
	Softcatala	73.97	50.21	58.61	34.43	-	-
	Eslema	-	-	-	-	50.77	17.3
	NLLB-600M	-	-	-	-	49.72	17.23
	NLLB-1.3B	-	-	-	-	50.04	17.44
	NLLB-3.3B	-	-	-	-	50.15	17.96
Constrained Submission	NLLB-600M fine-tuned	84.63	71.14	76.04	55.5	52.26	19.33
Open Submission	Transformer from scratch	81.35	63.95	71.48	45.92	50.37	16.86

Table 4: Evaluations computed on the FLORES+ dev set.

through fine-tuning.

Treatment of "«" and "»" tokens Given that "«" and "»" symbols are not in the NLLB vocabulary and they were present in the reference test sets, we decided to preprocess the training data by replacing "«" and "»" with "<<" and ">>" (as these tokens are in the vocabulary) and then revert this replacement after the model's inference.

3.2.3 Inference experiments

The generated translation is restricted to a length of 512 tokens. Testing on FLORES+ dev, we conducted a grid search on the top-performing model obtained from the LaBSE scoring step. We experimented with various beam sizes (B) and repetition penalty terms (β). We tried all combinations between $B = [3, 5, 10]$ and $\beta = [1, 3, 4]$. Nevertheless, no significant differences in performance were observed for these languages, so we ended up using the same hyperparameters employed during training: $B = 5$ and $\beta = 1$. For detailed results, see Appendix.

3.2.4 Configurations

In the fine-tuning, we used the AdamW optimizer (Loshchilov and Hutter, 2019) with $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-6}$, and $\lambda = 0.001$. The learning rate was set to 3×10^{-4}. We applied an inverse square root scheduler with 15,000 warmup steps.

For Aragonese, the batch size was set to 16, the gradient accumulation steps to 8, and the model was trained for 15 epochs. For Aranese and Asturian, the batch size was set to 8, the gradient accumulation steps to 4, and the models were trained for 10 epochs. All models were fine-tuned using the Transformers[8] library on H100 GPUs. Every 1,000 training steps, the ChrF score was computed on the FLORES+ dev set, and the model checkpoints were saved when the score improved.

3.3 Open Submission: Training

3.3.1 Model

We trained three transformer models from scratch using OpenNMT-py 3.2, each with its own BPE vocabulary. The vocabulary size for each model was set to 20,000 units, based on previous internal research investigating the impact of vocabulary size on BLEU scores (Outeirinho et al., 2024).

3.3.2 Configurations

All three models were trained on a single A100 GPU using the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.9998$, $\epsilon = 10^{-8}$, and learning rate 5×10^{-4}. The batch size was set to 2048 sentences, with a maximum length of 150 tokens per sentence. All models were trained for a maximum of 10 epochs.

3.4 Evaluation

To evaluate the performance of our models during the development phase, we used the FLORES+ dev set, which contains 997 general domain sentences. The results obtained on this test set for the two developed strategies can be seen in Table 4.

[8] https://huggingface.co/docs/transformers/

		Spanish → Aragonese		Spanish → Aranese		Spanish → Asturian	
		ChrF	**BLEU**	**ChrF**	**BLEU**	**ChrF**	**BLEU**
Baselines	Apertium	79.31	61.11	49.42	28.85	50.84	16.99
	Traduze	67.66	35.47	-	-	-	-
	Softcatala	71.99	47.08	48.29	26.07	-	-
	Eslema	-	-	-	-	50.91	17.17
	NLLB-600M	-	-	-	-	49.05	16.21
	NLLB-1.3B	-	-	-	-	49.71	16.54
	NLLB-3.3B	-	-	-	-	50.03	17.09
Constrained Submission	NLLB-600M fine-tuned	79.88	62.32	50.05	30.12	52.14	18.43
Open Submission	Transformer from scratch	78.61	59.76	48.84	27.31	50.54	16.68

Table 5: Evaluations computed on the FLORES+ devtest set.

3.4.1 Constrained Submission

Following this strategy, we surpassed the baselines across the three languages, achieving better performance than the state-of-the-art. We observed an increase of +2.63 in ChrF and +5.8 in BLEU for Aragonese (compared to Apertium), +3.41 in ChrF and +6.54 in BLEU for Aranese (compared to Apertium), and +1.49 in ChrF and +1.37 in BLEU for Asturian (compared to Eslema and NLLB-200-3.3B, respectively). These results suggest that both the thorough curation of data and the cross-lingual knowledge possessed by NLLB contributed to these improvements.

3.4.2 Open Submission

Leveraging transformer models has led to results that are only slightly behind the best rule-based approaches. However, the significance of these results lies in the fact that they enable the community to access and develop neural models that perform competently in a relatively short time compared to developing a new rule-based system from the ground up. These neural models, particularly transformers, offer new possibilities, such as the ability to learn from limited data and improved scalability, which can help prevent languages with fewer speakers from being marginalized in the online world.

4 Results

The FLORES+ devtest set, containing 1,012 sentences, was used to evaluate and rank the participants' models. Once the competition ended, the organizers made the FLORES+ devtest set public. To further expand the evaluation of our new models, we also obtained scores for the baselines using this test set. Consult Table 5 for all the scores. We see the same trend as with the FLORES+ dev set. Using the fine-tuned version of NLLB-200-600M on the cleaned data, we surpass all the baseline models for the three languages, whereas training the models with OpenNMT-py 3.2 lags behind. Specifically, we enhance the scores by +0.57 in ChrF and +1.21 in BLEU for Aragonese (compared to Apertium), +0.63 in ChrF and +1.27 in BLEU for Aranese (compared to Apertium), and +1.23 in ChrF and +1.26 in BLEU for Asturian (compared to Eslema). At the time of writing this paper, the final ranking scores were not available, so no mention of our final positions in the competition is included in this paper.

5 Discussion

Compared to traditional rule-based translation systems, neural models offer greater flexibility, scalability, and adaptability, making them the state-of-the-art in Machine Translation. Hence, our work in developing neural systems for Aragonese, Aranese and Asturian represents an advance in the preservation and promotion of the use of these languages. It also allows the research community to use our models for further advancements in language technology, linguistic research, and the development of more sophisticated and accurate translation systems.

6 Conclusions

This paper summarizes the work done by the ILENIA team for the Shared Task on Translation into Low-Resource Languages of Spain. By participating in this public competition, we have contributed to the creation and improvement of NMT models for Aragonese, Aranese, and Asturian - three minority languages of Spain. Prior to this task, no NMT models were available for Spanish-to-Aragonese and Spanish-to-Aranese translation.

We presented a Constrained and an Open Submission, each employing different approaches. For the Constrained Submission, adhering to data and model restrictions, we fine-tuned the NLLB-200-600M model, with considerable effort devoted to data cleaning and curation. For the Open Submission, we generated a large amount of synthetic data using Apertium, a rule-based MT system, and used it to train a transformer-based model from scratch.

Results on both FLORES+ dev and devtest sets across the three language directions show that the first strategy achieves better performance and improves translation quality compared to the baselines, whereas the second strategy lags slightly behind the best baseline models.

Acknowledgements

This work has been promoted and financed by the Spanish Ministry of Economic Affairs and Digital Transformation and by the Recovery, Transformation, and Resilience Plan (Funded by EU – NextGenerationEU within the framework of the project ILENIA and Nós Project with references 2022/TL22/00215337, 2022/TL22/00215336, 2022/TL22/00215335, 2022/TL22/00215334, by the Government of Catalonia through the Aina project, by the Xunta de Galicia through the collaboration agreements signed in with the University of Santiago de Compostela in 2021 and 2022 and by the Galician Government (ERDF 2014-2020: Call ED431G 2019/04, and ED431F 2021/01), by CIN/AEI/10.13039/501100011033, grants with references PID2021-128811OA-I00 (LingUMT), PLEC2021-007662 (Big-eRisk). Additionally, it has been supported by DeepR3 (TED2021-130295B-C32, TED2021-130295B-C33) (Funded by MCIN/AEI/10.13039/501100011033 and European Union NextGeneration EU/PRTR) and it is part of the project PID2021-123988OB-C33, financed by MCIN/AEI/10.13039/501100011033/FEDER, UE.

References

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, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

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.

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. Pilar.

Javier Garcia Gilabert, Aleix Sant, Carlos Escolano, Francesca De Luca Fornaciari, Audrey Mash, and Maite Melero. 2024. BSC submission to the AmericasNLP 2024 shared task. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 143–149, Mexico City, Mexico. Association for Computational Linguistics.

Tanmai Khanna, Jonathan North Washington, Francis M. Tyers, Sevilay Bayatli, Daniel G. Swanson, Flammie A. Pirinen, Irene Tang, and Hèctor Alòs i Font. 2021. Recent advances in apertium, a free/open-source rule-based machine translation platform for low-resource languages. *Mach. Transl.*, 35(4):475–502.

Guillaume Klein, Yoon Kim, Yuntian Deng, Vincent Nguyen, Jean Senellart, and Alexander M. Rush. 2018. Opennmt: Neural machine translation toolkit.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

C. Moseley. 2010. *Atlas of the World's Languages in Danger*. Memory of peoples Series. UNESCO.

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 NLLB Team, Marta R. Costa-jussà. 2022. No language left behind: Scaling human-centered machine translation.

Lars Nygaard and Jörg Tiedemann. 2003. Opus—an open source parallel corpus. In *Proceedings of the 13th Nordic Conference on Computational Linguistics*.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Appendices

A Spanish-to-Aragonese

		Beam Width		
		3	**5**	**10**
Rep. Pen.	**1**	84.63	84.63	84.63
	3	84.64	84.64	84.66
	4	86.63	84.64	84.63

Table 6: ChrF scores obtained using grid search in inference.

		Beam Width		
		3	**5**	**10**
Rep. Pen.	**1**	71.12	71.14	71.14
	3	71.15	71.16	71.16
	4	71.15	71.16	71.15

Table 7: BLEU scores obtained using grid search in inference.

B Spanish-to-Aranese

		Beam Width		
		3	**5**	**10**
Rep. Pen.	**1**	76.04	76.04	76.04
	3	76.04	76.04	76.04
	4	76.04	76.04	76.04

Table 8: ChrF scores obtained using grid search in inference.

		Beam Width		
		3	**5**	**10**
Rep. Pen.	**1**	55.5	55.5	55.5
	3	55.6	55.6	55.6
	4	55.6	55.6	55.6

Table 9: BLEU scores obtained using grid search in inference.

C Spanish-to-Asturian

		Beam Width		
		3	**5**	**10**
Rep. Pen.	**1**	52.28	52.26	52.23
	3	52.26	52.24	52.22
	4	52.25	52.25	52.23

Table 10: ChrF scores obtained using grid search in inference.

		Beam Width		
		3	**5**	**10**
Rep. Pen.	**1**	19.34	19.33	19.21
	3	19.25	19.24	19.22
	4	19.21	19.24	19.21

Table 11: BLEU scores obtained using grid search in inference.