Vicomtech@WMT 2024: Shared Task on Translation into Low-Resource Languages of Spain

David Ponce[*1,2] and **Harritxu Gete**[*1,2] and **Thierry Etchegoyhen**[1]
[1] Fundación Vicomtech, Basque Research and Technology Alliance (BRTA)
[2] University of the Basque Country UPV/EHU
{adponce,hgete,tetchegoyhen}@vicomtech.org

Abstract

We describe Vicomtech's participation in the WMT 2024 Shared Task on translation into low-resource languages of Spain. We addressed all three languages of the task, namely Aragonese, Aranese and Asturian, in both constrained and open settings. Our work mainly centred on exploiting different types of corpora via data filtering, selection and combination methods, along with synthetic data generated with translation models based on rules, neural sequence-to-sequence or large language models. We improved or matched the best baselines in all three language pairs and present complementary results on additional test sets.

1 Introduction

Despite significant progress in Machine Translation (MT) in recent years, notably with the advent of Neural Machine Translation (NMT) approaches (Bahdanau et al., 2015; Vaswani et al., 2017), translation from and into low-resource languages remains a challenge.

Spain features a large variety of languages beyond Spanish, with varying degrees of MT support. Important quality gains have thus been achieved for the Basque language within the NMT framework (Etchegoyhen et al., 2018), with large public deployments of quality MT systems[1]. For Catalan, a romance language with closer proximity to Spanish, earlier NMT improved over rule-based (RMT) and statistical (SMT) models, although with performance losses on out-of-domain test sets (Costa-jussà, 2017); more recent work on translation between similar languages, that included Catalan-Spanish, showed a predominance of NMT approaches to the task (Akhbardeh et al., 2021).

In addition to the aforementioned languages, there are languages such as Aragonese, Aranese and Asturian which could be viewed as extremely low-resourced in terms of MT technological support. For most, the main technology is still RBMT, based on the Apertium framework (Forcada and Tyers, 2016). The WMT 2024 shared task on translation into low-resourced languages of Spain addresses translation from Spanish into all three of these languages. In this work, we describe Vicomtech's participation in the shared task, where we submitted entries to both the constrained and open tracks.

In the remainder of this paper, we describe our approaches to improve MT performance for the three selected language pairs. We explored data selection, generation, and combination, comparing the use of different types of data to train end-to-end NMT models as well as fine-tuning pretrained multilingual MT models. In addition to typical parallel data curation, where we filtered the available parallel and comparable data according to sentence similarity, length differences and language identification, we also explored the generation of synthetic data along different lines. We notably compared the use of RBMT systems and large language models (LLM) to generate synthetic parallel datasets from the available monolingual data. The latter approach in particular showcased the potential of LLMs to create back-translations from the selected three low-resource Romance languages into the high-resource Spanish language.

2 Methodology

2.1 Parallel Data Curation

Despite the limited amount of data available for the languages addressed in this task, several crawled corpora were made available. However, after manually examining sampled of the data, they appeared to feature large amounts of noise, including poor alignments, language identification errors, or sentence pairs with empty information in one of the

[*] Equal contribution.
[1] https://www.euskadi.eus/traductor/

Proceedings of the Ninth Conference on Machine Translation, pages 934–942

languages. We therefore performed various types of filtering, described below.

Language Identification. We performed language identification with the Idiomata Cognitor tool[2], a Bayesian language identifier specialised on Romance Languages (Galiano-Jiménez et al., 2024a). We filtered all sentence pairs where the identified language on either side mismatched the expected language in the parallel dataset.

Length Ratio. We filtered all sentence pairs where the ratio of lengths, in terms of characters, was above a predefined threshold. Unless otherwise specified, we used a default ratio of 3.0. Our goal with this type of filtering was to remove obvious erroneous alignments rather than determine an optimal threshold in terms of length differences.

Sentence Similarity. We filtered all sentence pairs whose similarity score was below a predefined threshold. Similarity was computed as the cosine similarity of the sentence embeddings for each sentence pair. After preliminary experiments with different models, we opted for the all-MiniLM-L6-v2 model of the Transformers library[3], as it provided sufficient quality for the considered pairs, while also supporting sufficiently fast processing to run multiple filtering experiments. For each language pair, we assigned similarity scores to the parallel corpora after language and length filtering, manually examined samples of the data and determined a similarity threshold accordingly.

2.2 Synthetic Data Creation

For low-resource languages, parallel data are typically scarce and monolingual corpora are a rich source of complementary data. We aimed to explore different approaches to exploit this type of data, generating synthetic data by translating via RBMT systems, NMT models and LLMs (see Frontull and Moser (2024) for a similar approach). Depending on model availability and/or quality, we generated data to be used as either back-translations (BT) (Sennrich et al., 2016) from the low-resource languages into Spanish, or as forward-translations (FT) (Li and Specia, 2019) in the opposite translation direction. In either case, the synthetic data generated from monolingual data were used as parallel data to translate into the low-resource languages. Additionally, we used pivot machine-translation from Catalan to Spanish to complement the Spanish-Aranese parallel datasets, as described below.

RBMT data (BT + FT). As back-translations, we translated the available monolingual corpora in Aragonese and Aranese into Spanish with the corresponding Apertium systems.[4] As forward-translations, we generated synthetic data from Spanish into all three low-resource languages, since Apertium covered all three language pairs in that direction. Our goal in both the BT and FT cases was to evaluate the impact of data translated via transformation rules that tend to closely follow the structure of the original Spanish data.

NMT data (BT). We generated back-translations into Spanish for all three language pairs with baseline NMT models, either trained from scratch or pretrained and fine-tuned, on the curated parallel data, as described in Section 2.3. Considering the low volumes of clean parallel data and the relatively low quality of the baselines, we discarded the use of forward-translations in this scenario. Back-translations are more robust in this type of scenario, as the target language monolingual data are expected to be correct for the decoder to model and the noise in corresponding synthetic source data can be handled relatively well by NMT models in general. Our aim with NMT-based NMT data was to generate synthetic data of relatively fluid translations that would differ from, and could complement, RBMT translations.

LLM data (BT). We also leveraged a general-purpose language model in zero-shot fashion to generate back-translations, querying the model to translate from the low-resource language into Spanish. Our preliminary assessment on the three language pairs was that translation into the low-resource languages could not constitute a reasonable alternative, as most translations from Spanish into either low-resource language were of low quality, irrespective of the size of the selected model. However, in the reverse direction, in all three pairs translation quality was markedly better, indicating that the meaning of the text in the low-resource language could be properly captured by the model, while generating correct output in the high-resource Spanish language.

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

[3]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[4]https://www.apertium.org/ Note that there was no readily available system for Asturian to Spanish.

Pivot MT data. Among the available corpora for the task were data in Catalan-Aranese (see Section 3.1), which could be exploited as well via pivot MT. To this end, we translated the available corpora from Catalan to Spanish with a high-quality in-house NMT model trained on OPUS parallel data (see Appendix C for further details).

2.3 Models & Training

Models. We trained two main types of models: Transformer-base encoder-decoder models trained from scratch on the available parallel, with or without complementary synthetic data, and a pretrained multilingual model fine-tuned with the same data, namely an NLLB-200-600M model (Costa-jussà et al., 2022). By opting for two parallel approaches, we aimed to evaluate the positive or negative impact of accessing pretrained multilingual knowledge on the task. With either type of model, we trained baseline variants on the curated parallel data, which were used used to generate back-translations, as described in Section 2.2. Both types of models were also used on the combined datasets to train final models, as our main aim was to contrast and compare the use of pre-trained multilingual knowledge vs. focused training on a specific low-resource language pair.

Tagging. To train the model variants, we performed several experiments around data tagging, which has been shown to be an efficient approach to training data discrimination, for back-translations (Caswell et al., 2019) or comparable data (Gete and Etchegoyhen, 2022), for instance. We used specific tags, prepended to each training instance, to indicate the type of data at hand, namely <BT> or <FT>. We aimed to investigate in particular whether data tags would be beneficial or detrimental in the case of low amounts of parallel data, combined with larger sets of synthetic data.

3 Experimental Setup

3.1 Corpora

For the constrained track, we selected the parallel corpora for Asturian, Aranese, and Aragonese from the PILAR collection (Galiano-Jiménez et al., 2024b), the monolingual WikiMedia data for Spanish and Asturian, the parallel data for Spanish-Asturian and Spanish-Occitan (with Occitan related to Aranese) from CCMatrix (Schwenk et al., 2021b) and WikiMatrix (Schwenk et al., 2021a) for Spanish-Aragonese, all of which were downloaded from the OPUS repository (Tiedemann, 2012). We performed language identification to keep only the Aranese sentences from Occitan, and also translated the Catalan portion of the Catalan-Aranese dataset via pivot translation into Spanish. For the open track, we included Asturian and Aragonese monolingual data extracted from WikiDump[5], for additional back-translations.

Corpus	Lang.	# Sent.	# Filt.	Constr.
PILAR	ast	38.8K	-	✓
	arg	84.7K	-	✓
	arn	273.2K	-	✓
	cat-arn	64.1K	-	✓
WikiMedia	es	3.9M	2.7M	✓
	ast	65.7K	65.6K	✓
CCMatrix	es-ast	6.5M	533.7K	✓
	es-oci	925.5K	55.5K / 8.9K	✓
WikiMatrix	es-arg	33.7K	19.2K / 13.7K	✓
WikiDump	ast	3.2M	2.1M	✗
	arg	508.5K	255.1K	✗

Table 1: Corpora statistics. We indicate the number of initial (# Sent.) and filtered (# Filt.) sentences and corpus use in the constrained track (Constr.). x/y indicates filtering with similarity thresholds of 0.5 (x) and 0.7 (y).

Excepting the PILAR datasets, which were used as is, we filtered the contents from the parallel corpora using the methods described in Section 2.1. The similarity threshold was set at 0.7 after manually reviewing portions of the data. For Aragonese and Aranese, since significant portions of the datasets were discarded at this threshold, we also created an additional dataset with a 0.5 threshold. For the constrained task, we selected these larger, though noisier, datasets. For the open task, we opted for the smaller, higher-quality datasets, due to the greater availability of data.

As evaluation data, we selected the dev sets available in PILAR, as well as a filtered subset of 3,000 highquality sentence pairs from CCMatrix for Asturian, with a similarity threshold set at 0.9. The latter was created as all models consistently yielded significantly lower scores on the Asturian PILAR dev set, compared to the other language pairs, and Marian models trained with this development set struggled to converge effectively. We report results on the official development set throughout the paper, but discuss additional results on our own development set in Section 5.

Corpora statistics are summarised in Table 1.

[5]https://dumps.wikimedia.org/, accessed June 2024

3.2 Models

In this section, we describe the translation models we used for the task, including the baselines and the models trained on the selected data described in the previous section. Since we generated synthetic data for both forward (from Spanish) and backward (into Spanish) translation, we present each type of model in turn. Training details, including additional model characteristics and training hyper-parameters are described in Appendix A. All models were evaluated in terms of BLEU and chrF, computed with the sacreBLEU toolkit [6]. Statistical significance was computed via bootstrap resampling (Koehn, 2004) for all results. Best results, for $p < 0.05$, are indicated in bold in all tables.

3.2.1 Translation from Spanish

For translation from Spanish, we first assessed the quality of three baseline models not trained on any of the selected data: the rule-based Apertium for each language pair, as a reference MT system for these languages; the multilingual NLLB-200-distilled-600M model, pretrained on a broad range of languages including Asturian and Occitan, as a neural baseline under the constrained track limitation of pretrained models with fewer than one billion parameters; and the Llama3-8B instruction model (AI@Meta, 2024), as an experimental testbed for zero-shot LLM-based translation.

Lang.	Model	Dev	
		BLEU	chrF
ES→AST	Apertium	**17.1**	**50.7**
	NLLB	14.3	44.2
	Llama3	15.2	48.9
ES→ARG	Apertium	**66.0**	**82.**2
	NLLB	7.9	42.1
	Llama3	30.4	64.5
ES→ARN	Apertium	**38.0**	**60.0**
	NLLB	8.5	39.2
	Llama3	4.5	32.6

Table 2: Baseline model results on the development sets for translation from Spanish

Table 2 presents the results for each baseline model in this translation direction, in terms of BLEU (Papineni et al., 2002) and chrF (Popović, 2015). Apertium achieved the highest scores in all three language pairs, demonstrating the value of an RBMT approach for the selected languages. NLLB and Llama3 were notably both outperformed by large margins on ES-ARN; the former performed equally poorly on ES-ARG but the latter achieved a more reasonable performance of 30.4 BLEU points in this case, still far from the scores obtained by the Apertium baseline. The only language pair where all three models achieved relatively similar low scores was ES-AST, which might be due to the specifics of this development set (see Section 5 for further discussion).

Considering these results, we used Apertium to generate forward synthetic data for all ES→XX translation pairs. To prepare the final models, all related to translation from Spanish in the task, we used two types of approaches: fine-tuning the NLLB model on the selected data and training from scratch a Transformer-base model with 6 encoder layers and 6 decoder layers, trained with the Marian NMT toolkit (Junczys-Dowmunt et al., 2018).

3.2.2 Translation into Spanish

Model	Aranese	Aragonese	Asturian
Apertium	**34.8**	66.2	-
NLLB	31.0	55.6	64.1
Llama3	33.1	64.3	71.5
Marian	**34.0**	**69.6**	**86.5**

Table 3: BLEU scores for translation into Spanish on the PILAR development sets for Aranese and Aragonese, and on a custom development set for Asturian.

Translation into Spanish was performed to generate synthetic back-translations. For this task, we used the three baseline approaches described in the previous section (except for AST-ES with Apertium, as it is not currently supported) and trained an additional XX→ES Marian model on the selected parallel and forward-translation data.

Table 3 presents the BLEU scores for these models on the task-provided development sets for Aranese and Aragonese, and on our custom development set for Asturian. For Aranese, there was no statistically significant difference between the Marian and Apertium models, both outperforming NLLB and Llama3; in Aragonese, Marian outperformed all other models, with NLLB performing notably worse; in Asturian, it again significantly outperformed both NLLB and Llama3. Considering these results, we selected the Marian model to generate all back-translations. Additionally, since forward-translations were all generated using Apertium, the incorporation of a neural model could add more variety to the synthetic data.

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

Lang.	Model	Data	# Sent.	Source	Dev		Test	
					BLEU	chrF	BLEU	chrF
ast	Apertium	-	-	-	17.1	50.7	17.0	**50.8**
	NLLB	Parallel	533.7K	CCMatrix	**18.1**	**51.3**	**17.6**	**51.2**
		FT	-	-				
		BT	638.1K	PILAR+CCMatrix+WikiMedia				
arg	Apertium	-	-	-	**66.0**	**82.2**	**61.1**	**79.3**
	Marian	Parallel	19.22K	WikiMatrix	**66.0**	**82.2**	**61.1**	**79.3**
		FT	2.7M	WikiMedia				
		BT	103.9K	PILAR+WikiMatrix				
arn	Apertium	-	-	-	38.0	60.0	28.8	49.4
	Marian	Parallel	-	-	**38.7**	**60.3**	**29.8**	**49.8**
		MT	64.1K	PILAR cat-arn				
		FT	2.7M	WikiMedia [Tagged]				
		BT	392.8K	PILAR + CCMatrix + PILARcat-arn				

Table 4: BLEU and chrF scores for our primary submissions in the constrained track

Lang.	Model	Data	# Sent.	Source	Dev		Test	
					BLEU	chrF	BLEU	chrF
ast	Apertium	-	-	-	17.1	50.7	17.0	50.8
	NLLB	Parallel	533.7K	CCMatrix	**18.6**	**51.6**	**18.0**	**51.6**
		FT	-	-				
		BT	2.7M	WikiDump+PILAR+CCMatrix+WikiMedia				
arg	Apertium	-	-	-	**66.0**	**82.2**	**61.1**	**79.3**
	Marian	Parallel	13.7K	WikiMatrix	**65.9**	**82.2**	**61.0**	**79.3**
		FT	2.7M	WikiMedia				
		BT	353.5K	WikiDump+PILAR+WikiMatrix				
arn	Apertium	-	-	-	**38.0**	**60.0**	**28.8**	**49.4**
	Marian	Parallel	8.9K	CCMatrix (es-oci)	**37.9**	**60.0**	**28.8**	**49.4**
		MT	64.1K	PILARcat-arn				
		FT	2.7M	WikiMedia				
		BT	346.2K	PILAR+CCMatrix+PILARcat-arn				

Table 5: BLEU and chrF scores for our primary submissions in the open track

A notable result are the relatively high scores of Llama3 zero-shot translation into Spanish, confirming our initial assessments of the potential leveraging this type of LLM to translate from low-resource into high-resource languages. Further variants such as few-shot translation might be worth exploring in this type of scenarios.

4 Main Results

The best results for our shared task submissions are summarised in Table 4 and Table 5 for the constrained and open tracks, respectively. We report BLEU and chrF scores on the PILAR development sets and on the task test sets, as reported on the OCELoT website.

4.1 Constrained Track

In the constrained setup, our focus was on optimising translation models within the set limits of OPUS data and pretrained models under one billion parameters. For Asturian, the best results were achieved via a fine-tuning of NLLB using both parallel data from CCMatrix and back-translations generated from the PILAR, CCMatrix and WikiMedia corpora using our custom Marian model.

In the case of Aragonese and Aranese, training Marian models from scratch proved to be the most successful strategy. Given that NLLB was not specifically trained on these languages, this result was not unexpected. For these languages, we also incorporated forward-translations generated using Apertium and back-translations created with our Marian models. For Aranese, the use of parallel

Lang.	Model	Data	# Sent.	Source	Not tagged	Tagged
ast	Apertium	-	-	-	**17.1**	-
	Marian	Parallel	533.7K	CCMatrix	**16.9**	**17.4**
		FT	2.7M	WikiMedia		
		BT	-	-		
arg	Apertium	-	-	-	**66.0**	-
	Marian	Parallel	19.2K	WikiMatrix	**66.0**	46.5
		FT	2.7M	WikiMedia		
		BT	103.9K	PILAR+WikiMatrix		
arn	Apertium	-	-	-	38.0	-
	Marian	Parallel	-	-	37.9	**38.7**
		MT	64.1K	PILAR cat-arn		
		FT	2.7M	WikiMedia		
		BT	392.8K	PILAR + CCMatrix + PILARcat-arn		

Table 6: BLEU scores comparison between models trained with and without tags in the forward-translated data.

data from CCMatrix resulted in lower performance, likely due to the lower quality of these data, which were originally Spanish-Occitan alignments. The inclusion of the pivot translations from Catalan was also beneficial for the Spanish-Aranese pair.

Overall, when comparing our results to the baselines in Table 2, our custom models consistently outperformed the vanilla NLLB across all languages, particularly for Aragonese and Aranese, which were unseen by this model. The models trained for Asturian and Aranese also achieved higher scores than the Apertium baseline. For Aragonese however, our best submission could only match the Apertium baseline scores. This limitation is likely due to the influence of the forward-translations from the rule-based Apertium system, a factor which was not mitigated with data tagging.

4.2 Open Track

Our contributions to the open track were twofold: augmenting the training data by incorporating Asturian and Aragonese Wikipedia content, and generating back-translations using Llama3 in a zero-shot setting.

As shown in Table 5, these additions improved the BLEU score for Asturian by 0.5 points compared to the constrained track. However, the results for Aragonese did not benefit from the extra data, showing a slight decrease of 0.1 BLEU. For Aranese, the use of back-translations from Llama3 appeared to be detrimental, resulting in a performance drop of 0.8 BLEU points.

Overall, the open track models yielded mixed results, as the augmented data generated via back-translation and zero-shot LLM translation resulted in either minor gains or losses. This might be due to the specifics of the development and test sets, in the sense that the augmented data might come from domains of little benefit to improve the translation on these datasets. The results of Section 3.2 are still important in our view, notably the quality of NMT and LLM translations for either direct use or data augmentation.

5 Discussion

As previously indicated, given the low performance of all models in Asturian in preliminary experiments, we used a filtered subset of 3,000 sentences from CCMatrix as development set. However, to ensure consistency across languages, we relied on the best-performing model on the original PILAR dev set as the criterion for model selection for submission, leading to the exclusion of models that performed better on our custom dev set.

Model	Source	Official Dev	Custom Dev
Apertium	-	17.1	79.8
Open Submission	-	**18.6**	78.4
Marian	CCMatrix	17.2	**87.4**

Table 7: BLEU scores in Spanish-Asturian on the official WMT development set and on our custom development set from CCMatrix.

For reference, Table 7 presents the results of the top-performing model on our dev set: a Marian model trained exclusively on CCMatrix data without any synthetic data. While this model shows lower performance than the one chosen for the official submission and is comparable to the baseline obtained with Apertium, it performed notably bet-

ter on our own development set. Considering the large differences in scores between the PILAR and custom dev sets, it would be interesting to examine in detail the differences between the two datasets in future work.

Among our best submissions to both tracks, only one dataset was tagged, namely the forward-translations based on Wikimedia in ES-ARG. We performed several additional experiments on the use of tags to discriminate between types of data, with the most salient results shown in Table 6. Tags on forward-translations were beneficial for Asturian and Aranese, but for Aragonese their use resulted in a substantial decrease of almost 20 BLEU points on the dev set. This variation might be due to the differing amounts of data available: Asturian and Aranese featured 500K and 364K sentence pairs without tags, respectively, while Aragonese only counted with 119K such pairs. Whereas tags have been shown to be a successful means to discriminate between parallel and other types of data, their use might thus be detrimental when tagged data largely dominate the other types of data.

6 Conclusions

We described our submission to the WMT 2024 shared task on translation into low-resource languages of Spain. We followed a multi-pronged approach based on data filtering and augmentation, with multiple types of models trained on different combinations of data with or without tagging. Although we improved over the baselines in general, the gains were minor overall on the development sets provided for the task. Nonetheless, our experiments showed the benefits of training dedicated NMT models, which proved optimal in most cases over fine-tuning pre-trained translation models. We also demonstrated the potential of zero-shot LLM-based translation for translation of the selected low-resource languages into Spanish, an interesting path for future research as standalone translation or as a source of data augmentation.

Acknowledgments

We wish to thank the anonymous WMT reviewers for their helpful comments. This work was partially supported by the Department of Economic Development and Competitiveness of the Basque Government (Spri Group) via funding for project ADAPT-IA (KK-2023/00035).

References

AI@Meta. 2024. Llama 3 model card.

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, OndÅ™ej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina EspaÃ±a-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (wmt21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63, Florence, Italy. Association for Computational Linguistics.

Marta R. Costa-jussà. 2017. Why Catalan-Spanish neural machine translation? analysis, comparison and combination with standard rule and phrase-based technologies. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 55–62, Valencia, Spain. 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, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Jimmy Ba Diederik P. Kingma. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, CA, USA.

Thierry Etchegoyhen, Eva Martínez Garcia, Andoni Azpeitia, Gorka Labaka, Iñaki Alegria, Itziar Cortes Etxabe, Amaia Jauregi Carrera, Igor Ellakuria Santos, Maite Martin, and Eusebi Calonge. 2018. Neural Machine Translation of Basque. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pages 139–148.

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

Samuel Frontull and Georg Moser. 2024. Rule-based, neural and LLM back-translation: Comparative insights from a variant of Ladin. In *Proceedings of the The Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 128–138, Bangkok, Thailand. 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.

Harritxu Gete and Thierry Etchegoyhen. 2022. Making the most of comparable corpora in neural machine translation: a case study. *Language Resources and Evaluation*, 56(3):943–971.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Zhenhao Li and Lucia Specia. 2019. Improving neural machine translation robustness via data augmentation: Beyond back-translation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 328–336, Hong Kong, China. 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.

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.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021a. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021b. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

A Training Hyperparameters

The Marian models were transformer-base models. Optimization was performed with Adam (Diederik P. Kingma, 2015), with $\alpha = 0.0003$, $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 10^{-9}$. We used a working memory of 20GB and automatically chose the largest mini-batch that fit the specified memory. The learning rate was set to increase linearly for the first 16,000 training steps and decrease afterward proportionally to the inverse square root of the corresponding step. The validation data was evaluated every 5000 steps.

For fine-tuning the NLLB model, optimization was performed using Adafactor (Shazeer and Stern, 2018), with a learning rate of 0.0001, a clipping threshold of 1.0, and weight decay set to 0.001. The training used a batch size of 32 and a maximum sequence length of 128 tokens.

Each model was trained on a Nvidia L40 with 48GB of VRAM. Early stopping was applied with a patience of 10 epochs to prevent overfitting.

B Generation Parameters

For inference with Marian, we set a beam size of 6 and a normalization factor of 0.6.

For the NLLB model, implemented on the transformer library, the maximum input length was configured to 200 tokens, with a beam size of 4.

For Llama3, we set a maximum of 256 new tokens, enabled sampling with a temperature of 0.1, and set top-p to 0.9. We used the following prompt to direct the model to generate translations in the specified target language without additional commentary: "*Traduce a [Español|Aragonés|Aranés|Asturiano] la siguiente frase. No añadas ningún otro comentario.*" .

C Catalan-Spanish MT Model

We considered two main options to translate Catalan into Spanish, as a means to create additional Aranese-Spanish data via pivot translation: the pretrained multilingual NLLB model or an in-house Marian model trained on parallel corpora from OPUS (namely: dogc, gnome, opensubs, tatoeba, ubuntu, globalvoices, wikimatrix, ted and paracrawl). The latter achieved significantly better results, as shown in Table 8 on a test set of 2,000 sentence pairs randomly sampled from OPUS data.

Translation Model	**BLEU Score**
NLLB Model	55.4
Marian Model	**70.7**

Table 8: BLEU scores for Catalan to Spanish translation.