TIM-UNIGE Translation into Low-Resource Languages of Spain for WMT24

Jonathan Mutal and **Lucía Ormaechea**
TIM, University of Geneva
40 Boulevard du Pont-d'Arve – Geneva, Switzerland
Jonathan.Mutal@unige.ch, Lucia.OrmaecheaGrijalba@unige.ch

Abstract

We present the results of our constrained submission to the WMT 2024 shared task, which focuses on translating from Spanish into two low-resource languages of Spain: Aranese (spa-arn) and Aragonese (spa-arg). Our system integrates real and synthetic data generated by large language models (e.g., BLOOMZ) and rule-based Apertium translation systems. Built upon the pre-trained NLLB system, our translation model utilizes a multistage approach, progressively refining the initial model through the sequential use of different datasets, starting with large-scale synthetic or crawled data and advancing to smaller, high-quality parallel corpora. This approach resulted in BLEU scores of 30.1 for Spanish to Aranese and 61.9 for Spanish to Aragonese.

1 Introduction

This work presents the results of our constrained submission for the *Translation into Low-Resource Languages of Spain* shared task at WMT24.[1] The task involves translating from Spanish into two low-resource languages spoken in the northeast of the Iberian Peninsula: Aragonese (spa-arg) and Aranese (spa-arn).

Despite the existence of monolingual corpora for these languages, parallel data from Spanish to Aragonese is extremely scarce, amounting to only about 60,000 parallel sentences in OPUS (Tiedemann, 2016). In the case of Aranese, fewer than a thousand parallel sentences are available (FLORES+, Guzmán et al., 2019). In addition to that, these Romance languages are notable for their graphemic instability. Although proposals for orthographic standardization (Estudio de Filología Aragonesa, 2010) and official recognition (Boletín Oficial del Estado, 2006) have been introduced, the absence of a commonly accepted writing system has hindered the development of machine translation (MT) systems into Aragonese and Aranese (Forcada, 2020).

A few previous works have explored MT for these language combinations. For instance, Apertium MT systems (Forcada et al., 2011) provided translations for the above-mentioned pairs using a rule-based approach, achieving better results than neural-based MT systems (Oliver, 2020). Similarly, Cortés et al. (2012) complemented Apertium with an additional orthographic module, and proposed a bidirectional spa-arg MT system. More recently, a multilingual MT model (No Language Left Behind, NLLB Team et al., 2022) included under-resourced Iberian languages like Asturian in its training set. However, it did not cover Aranese or Aragonese.

Given the characteristics of this low-resource scenario, we addressed the translation from Spanish into Aragonese and Aranese using a multilingual multistage approach. The multilingual aspect involved leveraging data from linguistically related languages (such as Occitan for Aranese translation), and employing multilingual pre-trained models (specifically, NLLB[2]) to facilitate generalization across different languages. The multistage approach was designed to consecutively enhance translation performance in the target languages using increasingly specific fine-tuning data sets.

Additionally, we applied data augmentation techniques to increase the volume of relevant data in our training set. This involved: *i*) resorting to LLMs within the constraint of one thousand million parameters (in particular, BLOOMZ[3]) to synthetically create more data in the target languages, and *ii*) producing aligned data through Apertium systems on

[1] The source code for the experiments discussed in this article is available at https://github.com/jonathanmutal/WMT-24-Submission.

[2] Particularly, the following model: https://huggingface.co/facebook/nllb-200-distilled-600M.

[3] Specifically: https://huggingface.co/bigscience/bloomz-560m.

Proceedings of the Ninth Conference on Machine Translation (WMT), pages 862–870
November 15–16, 2024.

the basis of real and synthetic monolingual data from both sides of the languages pairs.

The structure of this paper is as follows: Section 2 describes the methods employed to gather parallel and monolingual data for our experiments. Section 3 introduces the multistage fine-tuning approach. In Section 4, we discuss the experiments conducted on both language combinations and the results obtained. Lastly, Section 5 summarizes our findings and suggests directions for future research.

2 Data

To train our MT `spa-arn` and `spa-arg` models, we first compiled parallel data from OPUS and FLORES+ (FLORES+$_{\text{DEV}}$) bilingual corpora. We then compiled monolingual data from two sources: *i*) we sub-sampled 19 million sentences from Wikimedia and NLLB datasets available in OPUS for Spanish, and *ii*) we collected all monolingual corpora for Aragonese, Aranese and Occitan from OPUS and PILAR (Galiano-Jiménez et al., 2024) when available. Table 1 details the number of segments in the bilingual corpora, and Table 2 reports the segments counts for each monolingual corpus. The notation "k" denotes thousands, and "M" signifies millions. A ✗ indicates the absence of available data.

Corpus	`spa-arn`	`spa-oci`	`spa-arg`
OPUS	✗	1.11M*	60k
FLORES+$_{\text{DEV}}$	997	997**	997

Table 1: Number of parallel segments for the available bilingual dataset. *CCMATRIX was not utilized. **These sentences were not used in any experiment.

Corpus	`spa`	`arn`	`oci`	`arg`
OPUS	19M	✗	739k	213k
PILAR	✗	322k*	✗	84k

Table 2: Number of monolingual segments for each available dataset. *Monolingual paragraphs were not utilized.

2.1 Synthetic Monolingual Data

We generated synthetic monolingual data in Aranese using BLOOMZ (Muennighoff et al., 2023). To do so, we fine-tuned BLOOMZ with the monolingual data (i.e., PILAR) using a causal language modeling objective, which involves predicting the next token in a sequence. We used a learning rate of 5×10^{-5} with an early stopping mechanism based on accuracy with a patient value of 5. As for the validation data, we randomly picked $1,000$ segments extracted from the same data distribution.

To generate new sentences in the target language, we took the beginnings of sentences in FLORES+$_{\text{DEV}}$. Then, the model completed segments from varying numbers of input words (ranging from 1 to 60 words) and generated up to a maximum of 65 tokens. We produced $59,820$ (997×60) sentences in Aranese using multinomial sampling. All other generation hyperparameters were set to their default values.[4]

2.2 Synthetic Parallel Data

Using the monolingual and synthetic data described above, we produced parallel data through Apertium systems (see Table 3). The following strategies were employed to synthetically create parallel sets:

- **Forward translation** (Burlot and Yvon, 2018). We generated synthetic Aranese, Occitan and Aragonese from monolingual Spanish (see Table 2).
- **Backtranslation** (Sennrich et al., 2016). We backtranslated the segments from monolingual Occitan, Aranese, and Aragonese. We also backtranslated synthetic segments in Aranese produced by BLOOMZ (see Section 2.1).

Strategy	***Corpus***	`spa-arn`	`spa-oci`	`spa-arg`
FT	OPUS	20M	20M	20M
BT	OPUS	✗	1.8M	273k
	PILAR	322k	✗	✗
	BLOOMZ	59k	✗	✗
Total		20.3M	21.8M	20.2M

Table 3: Training data synthetically generated using forward translation (FT) and backtranslation (BT).

3 Approach

Our approach, termed "multistage fine-tuning" involves sequentially refining a model using multiple datasets arranged in a specific order – a method proven to improve performance in machine translation for low-resource language pairs (Dabre et al., 2019).

[4]See documentation: `https://huggingface.co/docs/transformers/en/main_classes/text_generation`.

System	*Stage*	*Data*	BLEU↑	ChrF↑	TER↓
Apertium	-	-	28.8	49.4	72.3
MarianNMT	1	OPUS+PILAR (38M)	25.0	47.1	76.4
Helsinki-NLP	1	OPUS+PILAR (38M)	22.3	45.6	81.9
NLLB	1	OPUS+PILAR (38M)	29.0	49.4	72.3
NLLB	2.i	PILAR	28.2	48.8	73.0
NLLB	2.ii	PILAR+BLOOMZ	28.9	49.2	72.5
NLLB	3.i	FLORES+$_{\text{DEV}}$	*30.0	*49.7	*71.8
NLLB	3.ii	FLORES+$_{\text{DEV}}$	***30.1**	***49.8**	***71.5**

Table 4: BLEU, ChrF and TER calculated on the test data for `spa-arn`. Scores with * are significantly better than the baseline Apertium with $p < 0.01$, calculated using paired approximate randomization with $10,000$ trials.

In this work, the models were initially trained using large-scale synthetic or crawled data aiming to match or surpass the performance of the open-source Apertium MT systems. Following this, the models underwent further fine-tuning with smaller, high-quality parallel corpora to improve their performance.

Performance comparisons for the initial models were conducted among three systems: *i*) a model built from scratch using MarianNMT (Junczys-Dowmunt et al., 2018); *ii*) a fine-tuned Helsinki-NLP model with ≈72M parameters; and *iii*) a fine-tuned large language model, NLLB, trained on 200 different languages with a larger number of parameters (600M). This enabled us to identify the best performing model for the first stage.

4 Experiments and Results

All our systems are Encoder-Decoder models based on the Transformer architecture (Vaswani et al., 2017). The models were trained until convergence, with training progress monitored using BLEU score each $5,000$ steps and an early stopping patience value of 10 using FLORES+$_{\text{DEV}}$ as validation data. The details of the training procedure and the results obtained for validation are detailed in Appendix A and B.

In the following sections, we describe the evaluation setup as well as the experiments and results obtained for each language pair.

4.1 Evaluation Setup

We evaluated our models using the FLORES+ test data ($1,012$ sentences). We calculated accuracy-based metrics BLEU (Papineni et al., 2002) and ChrF (Popović, 2015), and also computed an error-based metric, i.e., Translation Error Rate (TER, Snover et al., 2006). All metrics were calculated using the Sacrebleu implementation (Post, 2018).[5] We used paired approximate randomization with $10,000$ trials to calculate the level of significance of the results.

We compared the performance of our models with Apertium MT systems, which are strong baselines for these language pairs.

4.2 Spanish-Aranese

For this specific language pair, we had almost no parallel sentences, but we did have a larger corpus of parallel sentences from a linguistically close language, Occitan (see Tables 1 and 2). To leverage the non-negligible quantity of data in this language, we built an MT model using all available data in Occitan and Aranese. In previous experiments, we observed that fine-tuning NLLB with multilingual data (i.e., Spanish-Aranese and Spanish-Occitan) outperformed its bilingual version (i.e., Spanish-Aranese). We also observed that using special tokens to differentiate the two languages is beneficial, and thus used them whenever possible. Appendix A.1 and A.2 show the results of these experiments.

Consequently, in the first stage, the models leveraged all available multilingual data from the OPUS and PILAR (including also synthetic data produced by forward and backtranslation), comprising roughly 42M sentences in Occitan and Aranese. We excluded sentences longer than 100 tokens, resulting in a total of 38M segments. We deliberately omitted synthetic data from BLOOMZ and the validation set to mitigate the risk of overfitting and ensure generalization in the first stage.

In the second stage, the NLLB model, identified

[5]The signatures are:
nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp
nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no
nrefs:1|case:lc|tok:tercom|norm:no|punct:yes|asian:no.

System	*Stage*	*Data*	BLEU↑	ChrF↑	TER↓
Apertium	-	-	61.1	79.3	27.2
MarianNMT	1	OPUS (15M)	58.2	77.8	29.9
Helsinki-NLP	1	OPUS (15M)	57.5	77.2	30.5
NLLB	1	OPUS (15M)	*60.5	*79.0	*27.7
NLLB-Post-Edition	2	FLORES+$_{\text{DEV}}$	61.0	*78.9	27.2
NLLB-Translation	2	FLORES+$_{\text{DEV}}$	***61.9**	**79.5**	***26.8**

Table 5: BLEU, ChrF and TER calculated on the test data for spa-arg. Scores with * are significantly better or worse than the baseline Apertium with $p < 0.01$, calculated using paired approximate randomization with $10,000$ trials.

as the most performing from the first stage, was fine-tuned using two different data combinations: PILAR data (2.i), and a combination of PILAR and synthetic Aranese data from BLOOMZ (2.ii). To mitigate the risk of overfitting in 2.ii, we fine-tuned the model using a fixed number of steps to reach a slightly higher validation BLEU score than the model trained in 2.i. During the third and final stage, the two models from the previous stage underwent $7,500$ additional training steps on the FLORES+$_{\text{DEV}}$ (3.i and 3.ii).

Results Results from the first stage showed that NLLB slightly outperformed the Apertium spa-arn system by 0.2 BLEU points, although this improvement was not statistically significant. MarianNMT and Helsinki-NLP performed worse than Apertium, which appears to agree to the findings in Oliver (2020). Interestingly, MarianNMT outperformed Helsinki-NLP, which might indicate that knowledge acquired during pre-training does not help to the task at hand. The underlying reasons for this discrepancy should be explored in future research.

The most performing model, NLLB (stage 3.ii), which was trained through a three-stage process, surpassed all previous models, improving the Apertium systems by 1.3 BLEU points and 0.4 ChrF points, and reduced the TER by 0.8 points.

The results indicate that the multistage approach enhance model performance. They also underscore the importance of a high-capacity model pre-trained on a diverse set of languages to improve translation from Spanish to Aranese. Additionally, the findings suggest that integrating synthetic data generated by BLOOMZ is beneficial in the third stage of fine-tuning (NLLB 3.i vs. NLLB 3.ii).[6]

4.3 Spanish-Aragonese

The model training for spa-arg was conducted in two stages. In the first stage, we used all OPUS-based synthetic data from Spanish to Aragonese to fine-tune NLLB.[7] This initial corpus amounted to roughly 20M parallel sentences, but we later filtered out the source or target sentences exceeding 100 tokens, which resulted in 15M pairs. With this set, we achieved comparable performance to the Apertium MT system in the validation data.

In the second stage, the model was fine-tuned with a lower learning rate, and utilized the FLORES+$_{\text{DEV}}$ in two different approaches:

- **Translation**, using as source the original sentences in Spanish.
- **Post-Edition (PE)**, using the Aragonese generated by the Apertium rule-based system as the source to train a post-edition model (apertium_arg-arg).

Results The experiments indicate that the performance is superior for translation tasks compared to post-edition tasks. Specifically, our optimal system, NLLB-Translation, surpassed the Apertium baseline by 0.8 BLEU points and reduced the translation error rate by 0.4 points.

Regarding the PE model, we assumed that a system trained using apertium_arg-arg could only help correct the mistakes made by such rule-based approach and thus improve its performance. Surprisingly, the resulting model (NLLB-Post-Edition) did not outperform the rule-based system, and instead degraded its results (see Table 5). One possible explanation for this is that the NLLB model from stage 1 was trained on spa-arg translation

[6]We also fine-tuned the resulting NLLB model from the first stage with FLORES+$_{\text{DEV}}$ data using $7,500$ steps. It underperformed the systems from the third stage.

[7]In previous experiments, we observed that PILAR was not helpful for the spa-arg task, so we decided to exclude it from the training set in our final models.

data rather than post-edition data. Further experiments need to be conducted in order to better understand the behavior of the PE model.

On another note, the results obtained for the fine-tuned Helsinki-NLP model revealed that the knowledge gained during pre-training does not appear to improve the results on the task. As can be observed, the model trained from scratch (MarianNMT) slightly outperforms the small-scale fine-tuned one (Helsinki-NLP), verified by the paired bootstrap statistical test.

5 Conclusions

Our experiments demonstrate the potential of combining synthetic data with multilingual pre-trained models to improve translation from Spanish into Iberian low-resource languages like Aranese and Aragonese. By leveraging data from linguistically related languages and employing a multistage approach, the `spa-arn` model achieved a BLEU score of 30.1, while the `spa-arg` model (NLLB-Translation) achieved 61.9 BLEU points. Our findings also indicate that the NLLB model, which benefited from a large number of pre-trained languages and high model capacity, delivered the best performances.

While these results are promising, we have identified several avenues for future research. One key area is to explore the impact of the ratio of real vs. synthetic data for training, as it can help evaluate how changes in data composition influence automatic metrics. Additionally, we plan to investigate the integration of external resources, such as dictionaries (Institut d'Estudis Aranesi, 2019) and orthographic standards (Academia Aragonesa de la Lengua, 2023), to determine whether these can further enhance the performance of our models.

Acknowledgments

We would like to thank the reviewers for their valuable comments and suggestions.

References

Academia Aragonesa de la Lengua. 2023. *Ortografía de l'aragonés*. Academia Aragonesa de la Lengua.

Boletín Oficial del Estado. 2006. Ley Orgánica 6/2006, de 19 de julio, de reforma del Estatuto de Autonomía de Cataluña. Art. 6.5.

Franck Burlot and François Yvon. 2018. Using monolingual data in neural machine translation: a systematic study. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 144–155, Brussels, Belgium. Association for Computational Linguistics.

Juan Pablo Martínez Cortés, Jim O'Regan, and Francis Tyers. 2012. Free/open source shallow-transfer based machine translation for Spanish and Aragonese. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2153–2157, Istanbul, Turkey. European Language Resources Association (ELRA).

Raj Dabre, Atsushi Fujita, and Chenhui Chu. 2019. Exploiting multilingualism through multistage finetuning for low-resource neural machine translation. 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 1410–1416, Hong Kong, China. Association for Computational Linguistics.

Estudio de Filología Aragonesa. 2010. *Propuesta ortografica provisional de l'Academia de l'Aragonés*. Edicions Dichitals de l'Academia de l'Aragonés, Zaragoza, Spain.

Mikel Forcada. 2020. Building machine translation systems for minor languages: challenges and effects. In *Revista de Llengua i Dret, Journal of Language and Law*, volume 73, pages 1–20.

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

Aarón Galiano-Jiménez, Felipe Sánchez-Martínez, and Juan Antonio Pérez-Ortiz. 2024. Pan-iberian language archival resource.

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.

Institut d'Estudis Aranesi. 2019. *Diccionari der aranés*. Acadèmia Aranesa dera Lengua Occitana.

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.

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.

NLLB Team, 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.

Antoni Oliver. 2020. Traducción automática para las lenguas románicas de la península ibérica. *Studia Romanica et Anglica Zagrabiensia*, 65:367–375.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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

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

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.

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. 2016. OPUS – Parallel Corpora for Everyone. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation: Projects/Products*. Baltic Journal of Modern Computing.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Training Setup NLLB and Helsinki-NLP

We employed the Adam optimizer with a batch size of 16. We used 50 warm-up steps, and the number of beams was set to 5. The maximum sequence length was set to 100, and the remaining hyperparameters were left unchanged[8], except for the learning rates which are reported in the following sections. All experiments were conducted using the Transformers library (Wolf et al., 2020) and the University of Geneva HPC clusters, Baobab and Yggdrasil. We used a fixed seed (111) for reproducibility purposes.

A.1 Helsinki-NLP Results

Given the absence of Aragonese or Aranese as targets in any of the existing OPUS-based Helsinki-NLP MT models, we decided to fine-tune them using different target languages. More specifically, our goal was to determine which of the available Romance languages (namely, Galician, Catalan, French, Italian and Romanian) would be most relevant for the `spa-arg` and `spa-arn` tasks.

After conducting an initial round of experiments, we observed that a geographically close language, Italian (i.e., `Helsinki-NLP/opus-mt-es-it`), most aided the translation into Aragonese on the validation set. Similarly, Catalan (i.e., `Helsinki-NLP/opus-mt-es-ca`) proved to be the most helpful target language for Aranese translation. For this language combination, we also conducted experiments to evaluate the potential gain from the use of two dedicated special tokens for Aranese and Occitan. Specifically, we used `<arn>` for Aranese and `<ca>` for Occitan.

LR	BLEU	ChrF
1×10^{-5}	59.1	78.6
2×10^{-5}	61.9	80.1
3×10^{-5}	62.1	80.3
4×10^{-5}	62.1	80.3
5×10^{-5}	**62.2**	**80.3**

Table 6: Results of Helsinki-NLP `spa-arg` models on validation data with different learning rates.

Once we selected the most relevant model for each language pair, we used different learning rates to fine-tune them for our task at hand. Table 6 reports the BLEU and ChrF results for `spa-arg` translation. Table 7 shows the results for the two versions of our `spa-arn` models: one that uses a single special token (`<ca>`) and another one that distinguishes between the two languages with distinct special tokens (`<ca>|<arn>`). All experiments were conducted using the Trainer class.[9]

	LR	BLEU	ChrF
Helsinki-NLP$_{\text{<ca>}}$	1×10^{-5}	26.0	52.7
	2×10^{-5}	26.5	53.2
	3×10^{-5}	24.8	52.2
	4×10^{-5}	25.8	52.8
Helsinki-NLP$_{\text{<ca>\|<arn>}}$	1×10^{-5}	**29.7**	**54.9**
	2×10^{-5}	28.6	54.3
	3×10^{-5}	28.8	54.2
	4×10^{-5}	29.0	54.9

Table 7: Results of Helsinki-NLP `spa-arn` models on validation data with different learning rates and different special token configurations.

A.2 NLLB Results

To generate Aranese, we used the Occitan special token (`oci_Latn`) in the target, which is presumably the closest language to Aranese covered by NLLB. Similarly to the Helsinki-NLP models, we used the Italian special token (`ita_Latn`) for Aragonese.

	LR	BLEU	ChrF
NLLB-Bi$_{\text{<oci>}}$	9×10^{-6}	37.7	59.9
	1×10^{-5}	37.7	59.9
	3×10^{-5}	37.6	59.8
NLLB-Multi$_{\text{<oci>}}$	9×10^{-6}	29.5	55.0
	1×10^{-5}	28.3	54.3
	3×10^{-5}	26.5	53.2
NLLB-Multi$_{\text{<oci>\|<cat>}}$	9×10^{-6}	37.8	60.0
	1×10^{-5}	**38.1**	**60.1**
	3×10^{-5}	37.9	60.0

Table 8: Results of NLLB `spa-arn` bilingual (NLLB-Bi$_{\text{<oci>}}$) and multilingual models (NLLB-Multi$_{\text{<oci>}}$ and NLLB-Multi$_{\text{<oci>|<cat>}}$) on validation data with different learning rates and special token configurations.

For Aranese translation, we carried out experiments to evaluate the gain of using a dedicated special token for Aranese and Occitan. In particular, we compared the performance of a multilingual model trained with Aranese and Occitan using the same token (`oci_Latn`), NLLB-Multi$_{\text{<oci>}}$, and another model using two special tokens: one for Aranese (`oci_Latn`) and a different one for

[8] Refer to: https://huggingface.co/docs/autotrain/en/seq2seq_params.

[9] Refer to: https://huggingface.co/docs/transformers/main_classes/trainer.

Occitan (cat_Latn), NLLB-Multi$_{<oci>|<cat>}$. We also assessed the performance of a bilingual model trained only with Spanish-Aranese data for comparison purposes (NLLB-Bi$_{<oci>}$). Table 8 shows the results on the validation data for the three approaches, indicating that the use of special tokens to differentiate the language is beneficial, and so is including Occitan in the training set.

Data	PILAR		PILAR+BLOOMZ	
LR	**BLEU**	**ChrF**	**BLEU**	**ChrF**
1×10^{-8}	35.2	57.5	38.1	60.1
5×10^{-8}	36.0	58.4	38.0	60.0
1×10^{-6}	37.7	59.9	39.2	60.5
9×10^{-6}	37.4	59.6	**39.9**	**60.9**

Table 9: Results of NLLB on stage two with PILAR and BLOOMZ on validation data with different learning rates.

Table 9 shows the results of NLLB on stage two and Table 10 shows the results of NLLB on spa-arg.

LR	**BLEU**	**ChrF**
5×10^{-7}	64.2	81.4
1×10^{-6}	63.6	81.1
3×10^{-6}	65.2	81.9
9×10^{-6}	65.2	81.9
1×10^{-5}	**65.4**	**82.0**
3×10^{-5}	65.3	81.9

Table 10: Results of NLLB-Baseline spa-arg on validation data with different learning rates.

B MarianNMT Setup and Results

LR	**BLEU**	**ChrF**
3×10^{-5}	26.6	53.2
5×10^{-5}	29.6	54.9
3.5×10^{-4}	**30.5**	**55.5**
3×10^{-3}	n.a.n	n.a.n

Table 11: Results of MarianNMT spa-arn models on validation data with different learning rates.

We used the default hyperparameters from the Marian toolkit (Junczys-Dowmunt et al., 2018) to train the models.[10] We conducted all experiments employing three random seeds and averaging the results measured by the automatic metrics. This approach is intended to reduce the variability of results inherent to individual models randomly initialized.

LR	**BLEU**	**ChrF**
3×10^{-5}	**55.7**	**78.6**
5×10^{-5}	53.3	77.7
3.5×10^{-4}	50.9	76.8
3×10^{-3}	n.a.n	n.a.n

Table 12: Results of MarianNMT spa-arg models on validation data with different learning rates.

Tables 11 and 12 present the results for spa-arn and spa-arg across different learning rates. The notation "n.a.n" indicates that the model diverged at that particular learning rate.

[10]Refer to: https://marian-nmt.github.io/docs/cmd/marian/.