Robustness of Fine-Tuned Models for Machine Translation with Varying Noise Levels: Insights for Asturian, Aragonese and Aranese

[1]Martin Bär, [1]Elisa Forcada Rodríguez and **[2]María García-Abadillo Velasco**

[1]Erasmus Mundus Master in Language and Communication Technologies (LCT)
[2]Master in Language Analysis and Processing (LAP)
University of the Basque Country (UPV/EHU)

{mbr001, eforcada001, mgarciaabadill001}@ikasle.ehu.eus

Abstract

We present the LCT-LAP proposal for the shared task on *Translation into Low-Resource Languages of Spain* at WMT24 within the constrained submission category. Our work harnesses encoder-decoder models pretrained on higher-resource Iberian languages to facilitate MT model training for Asturian, Aranese and Aragonese. Furthermore, we explore the robustness of these models when fine-tuned on datasets with varying levels of alignment noise. We fine-tuned a Spanish-Galician model using Asturian data filtered by BLEU score thresholds of 5, 15, 30 and 60, identifying BLEU 15 as the most effective. This threshold was then applied to the Aranese and Aragonese datasets. Our findings indicate that filtering the corpora reduces computational costs and improves performance compared to using nearly raw data or data filtered with language identification. However, it still falls short of the performance achieved by the rule-based system Apertium in Aranese and Aragonese.

1 Introduction

Spain is home to a rich linguistic landscape, yet this diversity is accompanied by disparities in terms of speaker numbers and language resources. Languages with co-official status, such as Basque, Catalan and Galician, were previously considered to have limited resources but are now included in numerous popular LLMs. Consequently, research in this field has shifted towards cases where data scarcity is even more pronounced, such as Asturian, Aragonese and Aranese. These languages are the focus of a shared task at the Conference on Machine Translation 2024. The objectives of this task include investigating transferability among Romance languages and determining the most effective methods for utilizing pretrained models in translations between Spanish and low-resource Romance languages.

The methodology employed involved the following steps:

1. Implementing automated methods for curating data. The constrained submission framework enables researchers to utilize corpora that may be notably noisy. Our work aims to propose solutions to this challenge.

2. Creating synthetic data for the monolingual PILAR Galiano-Jiménez et al. (2024b) corpora.

3. Harnessing models trained on other, resource-richer (Iberian) Romance languages with the presumption that this facilitates cross-lingual transfer. The model fine-tuned for Asturian was originally trained on Galician, while the models fine-tuned for Aranese and Aragonese were originally trained on Catalan.

The official metrics for the shared task are BLEU and chrF. The metrics employed in this study are BLEU and chrF++, as they are relatively straightforward to calculate and there is currently no robust neural-based metric for our target languages.

2 Background

2.1 Spanish Linguistic Landscape

Although the official language in Spain is Spanish, it coexists with other co-official and minority languages. The predominance of Spanish over the other languages and dialects is associated with historical reasons: since the Middle Ages, Spain had undergone a process of Castilianisation, which became very important in the 14th century, when the dominance of the Kingdom of Castile in the centre of the Iberian Peninsula led to the expansion of the use of Castilian. This continued until the 20th century, with the consequent marginalisation of the other vernacular languages (Martínez, 1982). The

Proceedings of the Ninth Conference on Machine Translation, pages 918–924

co-official languages, Basque, Catalan and Galician, were considered to have limited resources in the past. This picture has changed, as efforts from both research and industry have contributed significantly to integrating them into the field of Language Technology. However, there are also non-official Ibero-Romance languages that are considered as having limited resources:

- Asturian: spoken in Asturias, the northeastern part of Leon, Zamora and the north of Portugal (ARIAS, 2002).
- Aragonese: spoken in the north of the province of Huesca and in the extreme northwest of Zaragoza (Marco Villanueva, 2012).
- Aranese: a variant of Occitan, spoken in the province of Aran (Rey and Canalís, 2006).

In this context, initiatives such as PILAR (Pan-Iberian Language Archival Resource) work to enrich and expand the resource availability of these languages (see Section 3).

2.2 Other related works

The interest in low-resource languages has recently increased, leading to a considerable amount of research on the subject (Ranathunga et al., 2023). Several studies on machine translation for low-resource languages can be found, such as the article by Karakanta et al. (2018), which works with non-parallel corpora, or Kumar et al. (2021), which focuses on recasting systems from high-resource languages to low-resource languages.

As far as Iberian languages are concerned, there are other investigations, such as the one published by Oliver et al. (2023), which explores techniques for training NMT systems applied to high- and low-resource Iberian languages or the work by Ko et al. (2021), which adapts high-resource NMT models to translate low-resource languages related to Spanish.

With respect to WMT, since its first edition in 2016, there have been three shared tasks related to the field: In 2020, a task was proposed on unsupervised and very low-resource languages, focusing on Upper Sorbian (Fraser, 2020). The following year, a workshop on multilingual low-resource translation for Indo-European languages was presented, focusing on North Germanic languages such as Icelandic and Romance languages such as Occitan (Libovický and Fraser, 2021). Finally, in 2022, a task related to unsupervised MT and very low-resource supervised MT was suggested, with Upper and Lower Sorbian languages (Weller-Di Marco and Fraser, 2022).

3 Data

This research falls into the constrained submission category, as all data used was obtained from the mentioned sources in the shared task: the Open Parallel Corpora, also known as OPUS (Tiedemann, 2009), and the Pan-Iberian Language Archival Resource, shortened as PILAR (Galiano-Jiménez et al., 2024b).

The Spanish development set is part of the FLORES+ Evaluation Benchmark (NLLB Team et al., 2022). The Asturian, Aragonese, and Aranese counterparts of FLORES+ are published alongside PILAR.

Finally, both the BLEU reference translations for the OPUS data and the synthetic Spanish counterparts for the PILAR data were generated with Apertium (Forcada and Tyers, 2016).

3.1 OPUS corpora

OPUS is a public multilingual collection of parallel corpora that gathers open-source documents available on the Internet (Tiedemann and Thottingal, 2020) and supports 744 languages. The constraint submission is limited to all data in OPUS, thereby enabling researchers to create synthetic translations from other languages into Asturian (ast), Aragonese (arg), Aranese (arn), or Spanish (es). However, the data utilized in this work exclusively employs the corpora for the combinations es <> ast/arg/arn.

Given that the collected corpora were not consistently well-aligned, we implemented a filtering pipeline, as detailed in Section 3.2, to produce a smaller but cleaner dataset. The effectiveness of this approach is reflected in the "BLEU 15" column of Table 1.

3.2 OPUS Data filtering

Around 8 million aligned sentences were collected from OPUS for the three target languages, although this number was significantly reduced when applying a filtering pipeline. Three main steps were followed to filter out invalid sentences:

- **Basic filtering:** removing unnecessary white spaces, empty lines, and characters not sup-

ported by the file encoding for all target languages.

- **Idiomata Cognitor:** filtering out all sentence pairs whose target language was not labeled as Asturian, Aragonese or Aranese and whose source language was not labeled as Spanish by Idiomata Cognitor (Galiano-Jiménez et al., 2024a), a high-precision classifier trained using Bayesian methods and capable of identifying 10 Romance languages.

- **BLEU threshold filtering**[1]: we first translated the Spanish counterparts of the Asturian/Aragonese/Aranese datasets into the respective target languages using Apertium. Next, we calculated BLEU scores for the original Asturian/Aragonese/Aranese sentences as references and their translations as hypotheses. Then, we filtered the datasets to various BLEU thresholds, assuming that alignments are more likely to be correct if the sentence pairs have high BLEU scores[2]. For Asturian, this was done using four different BLEU thresholds: 60, 30, 15 and 5 BLEU. For Aragonese and Aranese, we only used one threshold.

3.3 PILAR Corpora

PILAR is a recently created corpus of texts from different languages spoken in the Iberian Peninsula, including Asturian, Aragonese, Aranese, Balearic and Valencian.

For our purposes, the monolingual data from Asturian, Aragonese and Aranese, and the Aranese counterpart from the Catalan-Aranese parallel corpora was backtranslated into Spanish (see Table 2) using Apertium: backtranslation can be understood as providing monolingual training data with a synthetic sentence source obtained by automatically translating the target sentence into the source language (Sennrich et al., 2015).

[1]We used Bleualign as a reference (Sennrich and Volk, 2010). However, we did not calculate the BLEU score between the hypothesis and reference sentences for both languages, nor did we compute the subsequent harmonic mean, given the fact that Apertium web tool does not support translations from Asturian into Spanish. Instead, we limited our calculations to the BLEU score of the original Asturian/Aragonese/Aranese sentences as references and the translation of its Spanish counterpart obtained with Apertium as hypotheses.

[2]BLEU evaluates translations by comparing n-grams between the model output and a reference, favouring those that are closest in terms of word and order. This may favour sentences in both the source and target languages that are easier to translate for Apertium.

	Asturian		**Aragonese**		**Aranese**	
	BLEU 15	Raw	BLEU 15	Raw	BLEU 15	Raw
GNOME	18,435	68,668	2,004	5,529	0	77
KDE4	4,515	26,023	-	-	667	49,593
NLLB	585,683	6,470,015	-	-	65,797	925,448
QED	125	421	18	222	45	282
Tatoeba	58	159	3	13	5	189
TED2020	40	116	-	-	-	-
WikiMatrix	-	-	13,639	33,724	7,398	35,805
wikimedia	27,776	45,506	2,908	4,457	629	1,980
XLEnt	0	274,257	3	16,822	0	99,472
	636,632	6,884,903	**18,575**	60,767	**74,502**	1,112,879

Table 1: Number of raw sentence pairs obtained from the OPUS repository and the final number of sentences after filtering them with a BLEU score threshold of 15.

	Asturian	**Aragonese**	**Aranese**
crawled	14,776	60,028	7,358
literary	24,093	24,675	229,886
paragraphs	-	-	86,568
sentences	-	-	64,141
Total	38,869	84,703	387,953

Table 2: Number of monolingual sentences from PILAR that were backtranslated with Apertium.

4 Methodology

The methodology of this work involved fine-tuning two pretrained models (see section 4.1) on backtranslated PILAR and filtered OPUS data (see section 3 and section 3.2). The total number of sentences for each language is presented in Table 3.

The experimental setup utilized a Tesla V100-PCIE-32GB GPU running with NVIDIA driver version 535.104.12 and CUDA version 12.2, alongside the HuggingFace Transformers library for model loading and fine-tuning.

All models underwent training for at least 1 epoch (Table 3 shows when each model converged). The best model selection was based on the BLEU score derived from the development set. Additionally, zero-shot translation without fine-tuning was conducted as a baseline for comparing results.

4.1 Models

Two models from `Helsinki-NLP` (Tiedemann and Thottingal, 2020) were used for our experiment:

- `opus-mt-es-gl`: a transformer-align model from Spanish into Galician that achieved a BLEU 67.6 and a chr-F score of 80 in the Tatoeba test. Given the close linguistic relationship between Asturian and Galician-Portuguese, we aimed to explore transfer learning when fine-tuning on Asturian data.

	Model	Data	Sentences	Epochs	Steps	BLEU	chrF++
AST	**apertium**	**-**	**-**	**-**	**-**	17.1	**50.69**
	es-gl-noft-ast	-	-	-	-	5.75	38.66
	es-gl-ft-basic	basic clean	6,884,903	0.87	55k	17.07	49.89
	es-gl-ft-idiomata	idiomata cognitor	4,521,302	1.76	36k	17.32	50.24
	es-gl-ft-bleu	bleu_60	440,794	2.61	14k	17.61	50.39
	es-gl-ft-bleu	bleu_30	582,883	3.95	22k	17.79	50.48
	es-gl-ft-bleu	bleu_5	743,846	3.44	25k	17.84	50.57
	es-gl-ft-bleu	bleu_15	636,632	2.61	18k	17.85	50.46
	es-gl-ft-backtr	**bleu_15 + PILAR**	**675,501**	**3.22**	**22k**	**17.90**	50.58
ARG	**apertium**	**-**	**-**	**-**	**-**	**66.05**	**82.23**
	es-ca-noft-arg	-	-	-	-	8.38	46.23
	es-ca-ft-arg	idiomata cognitor	27,335	4.67	1k	32.87	64.79
	es-ca-ft-arg	basic clean	60,767	2.91	1k	33.17	65
	es-ca-ft-arg	bleu_15	18,575	47.95	7k	41.39	70.38
	es-ca-ft-arg	bleu_15 + PILAR	103,278	7.43	8k	41.53	70.84
ARN	**apertium**	**-**	**-**	**-**	**-**	**38.02**	**60.01**
	es-ca-noft-arn	-	-	-	-	5.75	38.66
	es-ca-ft-arn	idiomata cognitor	383,575	4.67	14k	9.61	40.67
	es-ca-ft-arn	basic clean	1,112,879	2.65	14k	9.70	40.74
	es-ca-ft-arn	bleu_15	74,502	6.86	9k	10.19	41.88
	es-ca-ft-arn	bleu_15 + PILAR	462,455	0.83	8k	29.04	54.85

Table 3: BLEU and chrF++ scores on the FLORES+ devset comparing baselines (`apertium` and models with `noft` in their names) and fine-tuned models (`-ft-`) across varying levels of alignment noise. Baselines always occupy the first two rows for each language. Subsequent models are listed in ascending order of BLEU scores. Best performing architectures are highlighted in bold.

- `opus-mt-es-ca`: a transformer-align from Spanish into Catalan with a BLEU score of 68.9 and a chr-F score of 0.832 in the Tatoeba test. We aimed to explore transfer learning when fine-tuning Catalan for Aranese and Aragonese.

5 Results

As Table 3 shows, the results of our experiments were compared with two baselines: Apertium, a rule-base system that supports translations in the same languages as those investigated in this work, and the respectively selected model for our experiments with zero-shot translations (i.e. without fine-tuning).

Overall, the results for Aragonese and Aranese show the same trend: the highest performance was achieved by fine-tuning on data filtered with a BLEU threshold of 15, combined with the back-translated PILAR corpora. While the backtranslated data yielded improvements of 18.85 BLEU for Aranese, this improvement was only 0.14 BLEU for Aragonese. Interestingly, fine-tuning on data that had only undergone basic cleaning outperformed our approach of filtering out sentences in other languages. The zero-shot translation approach yielded the lowest results by a significant margin. Despite these efforts, our results still fall short of the baseline Apertium by approximately 9 points in Aranese and nearly 25 points in Aragonese.

Our best result for Asturian is the only one comparable to the baseline Apertium. Our fine-tuned model, which uses a BLEU score threshold of 15 and the PILAR corpora, outperforms the baseline by 0.8 BLEU points. However, it falls short of the baseline by 0.11 chrF++ points.

See the following sections for a more detailed description of each language's results.

5.1 Asturian

Our results show that setting a threshold of 15 BLEU for OPUS-aligned corpora yields the best performance in Asturian. It slightly outperforms thresholds of 5 and 30 BLEU and achieves an improvement of almost 0.25 over the cleanest filtered set of corpora with a threshold of 60 BLEU.

Note that an Asturian tokenizer was trained and implemented; however, its performance did not exceed a BLEU score of 17.6 and it was consequently omitted from Table 3. Consequently, no tokenizer was trained for Aranese and Aragonese.

Integrating backtranslated Asturian PILAR results in almost a 1-point BLEU score improvement compared to the slightly preprocessed raw OPUS data (basic clean in Table 3), and a slight improvement of 0.05 compared to the filtered OPUS data with 15 BLEU threshold without PILAR data.

Regarding the baselines, our best method (data filtered with a 15 BLEU threshold and backtranslated PILAR) achieves similar performance as Apertium, with a 0.8 BLEU score improvement and a 0.11 lower chrF++ score. The zero-shot translation results are by far the worst, with scores approximately 12 points below the best results.

5.2 Aragonese

As detailed in section 7, only the best filtering threshold for OPUS data in Asturian was also applied to Aragonese.

Our best result is again the result of fine-tuning the bleu_15 + PILAR corpora on a model initially fine-tuned for Spanish-Catalan translation. It outperforms the model finetuned on almost raw data by 8.36 BLEU points. Comparing these results to a model only trained on the bleu_15 data, reveals that using the backtranslated data only yielded an improvement of 0.14 BLEU. However, these results lag behind the Apertium baseline, which obtains scores with approximately 25 points difference in BLEU and around 12 points in chrF++.

The zero-shot baseline model (`es-ca-noft-arg`) achieved a similarly low score as the zero-shot models in the other languages and performed significantly worse than the other approaches. It lags behind the best result from Apertium by approximately 58 BLEU points and around 36 chrF++ points.

5.3 Aranese

As detailed in section 7, only the best filtering threshold for OPUS data in Asturian was applied to Aranese.

Showing the same trend as the results for the other two languages, the approach using bleu_15 + PILAR corpora is the most effective. It achieves an improvement of about 20 BLEU points and approximately 14 chrF++ points over the other data sets, which only underwent basic cleaning or language filtering with Idiomata Cognitor. In contrast to our results for Aragonese, the Aranese backtranslated data helped to increase performance tremendously (+18.8 BLEU). As expected, the zero-shot baseline performs the worst, with even greater score disparities.

Despite this significant improvement compared to our other techniques, the BLEU filtering approach fails to outperform the Apertium baseline. Apertium performs significantly better, with approximately 9 points difference in BLEU and over 5 points in chrF++.

6 Conclusions

This work was conducted within the constrained category of the shared task *Translation into Low-Resource Languages of Spain* at WMT24. It introduces a pipeline for filtering low-quality alignments in parallel corpora and subsequently fine-tuning translation models to assess the noise robustness of Neural Machine Translation. The paper details the data collection and curation processes for the three target languages selected for this task (Asturian, Aragonese and Aranese), with a particular focus on fine-tuning models for Spanish to Asturian under varying levels of noise and generalizing the results to the other two language pairs.

The initial phase involved curating the OPUS corpora for the Asturian-Spanish pair. This pipeline included **1)** cleaning unsupported characters and blank spaces, **2)** filtering out sentence pairs that were not in Spanish or one of the target languages using Idiomata Cognitor, and **3)** generating translations with Apertium to determine alignment quality of the sentence pairs and establishing four different BLEU thresholds for filtering. After observing that a BLEU threshold of 15 yielded the best performance, we incorporated backtranslated PILAR data into the filtered OPUS corpora. Part of step **3** was omitted for Aranese and Aragonese due to computational constraints.

Despite these filtering approaches resulting in the loss of some significant portions of the available corpora, we observed that the fine-tuned models effectively leveraged prior knowledge from the chosen related languages (Galician and Catalan).

- Our best performing fine-tuned model for Asturian outperformed the baseline Spanish-Galician model by 12.15 BLEU points.
- Our best performing fine-tuned model for Aragonese outperformed the baseline Spanish-Catalan model by 33.15 BLEU points.
- Our best performing fine-tuned model for Aranese outperformed the baseline Spanish-Catalan model by 23.29 BLEU points.

The results for our best fine-tuned Asturian model were relatively strong, achieving competitive scores compared to Apertium. Although the same approach was applied to Aranese and Aragonese, it did not surpass the Apertium baseline by a significant margin.

Overall, we demonstrated that 1) filtering out low-quality translations from a noisy parallel dataset improves fine-tuning results and yields faster training times, and 2) results for Asturian can reach baseline levels with a smaller, cleaner and more computationally efficient corpus, suggesting that the selected models can handle noise only to a certain degree. However, we cannot assert that this approach is effective for Aranese and Aragonese, as the results for these languages fall short of the rule-based baseline.

7 Limitations

The scope of this work is mainly limited by computational resources. The HiTZ Basque Center for Language Technology kindly allowed the authors access their resources, but understandably, priority was given to projects more closely related to their main research focus at the time. This led us to 1) implement our own BLEU score filter and dispense with newer, more accurate sentence alignment algorithms, 2) generalize the best BLEU score threshold in Asturian to the other two languages, Aranese and Aragonese.

One potential improvement to our approach would be the application of curriculum learning, where initial fine-tuning is performed on large synthetic data, followed by further fine-tuning on high-quality parallel data.

8 Further Work

Future work could address the limitations discussed in Section 7 by 1) exploring the outcomes of fine-tuning a language model on corpora cleaned using not just one, but various sentence alignment algorithms such us Bertalign (Liu and Zhu, 2022) or Vecalign (Thompson and Koehn, 2019), and 2) investigating whether Aranese and Aragonese tolerate different noise thresholds compared to Asturian. Additionally, future research might:

- estimate the amount of KWh required to fine-tune different amounts of corpora,
- examine whether data augmentation through backtranslation of additional OPUS corpora could enhance performance, as this is permitted in the constrained category,
- explore whether a tokenizer trained on a larger corpus and specialized in Asturian, Aranese, and Aragonese could improve results.

Ethics Statement

The authors of this study adhered to the principles outlined in the ACL Ethics Policy and the ACM Code of Ethics. Our goal is for this research to benefit society by exploring language transferability among three low-resource languages, thus advancing machine translation techniques for underrepresented languages. All data used in this study was sourced from institutional and legal channels, explicitly approved and aligned with the original guidelines of the constrained submission for the shared task on *Translation into Low-Resource Languages of Spain* at WMT24. Throughout the research process, we prioritized transparency and fairness. We conducted honest and reliable evaluations and clearly communicated the limitations of our methods.

Acknowledgements

We would like to express our gratitude to Gorka Labaka and Nora Aranberri for their feedback and support. Furthermore, we would like to thank the HiTZ Basque Center for Language Technology for providing us with access to their computational resources, which will facilitate further research. This work was co-funded by the Erasmus Mundus Masters Programme in Language and Communication Technologies, EU grant no. 2019-1508.

References

X. LL. GARCÍA ARIAS. 2002. Breve reseña sobre la lengua asturiana. *Informe sobre la llingua asturiana*, page 15.

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.

Alexander Fraser. 2020. Findings of the wmt 2020 shared tasks in unsupervised mt and very low resource supervised mt. In *Proceedings of the Fifth Conference on Machine Translation*, pages 765–771.

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.

Alina Karakanta, Jon Dehdari, and Josef van Genabith. 2018. Neural machine translation for low-resource languages without parallel corpora. *Machine Translation*, 32:167–189.

Wei-Jen Ko, Ahmed El-Kishky, Adithya Renduchintala, Vishrav Chaudhary, Naman Goyal, Francisco Guzmán, Pascale Fung, Philipp Koehn, and Mona Diab. 2021. Adapting high-resource nmt models to translate low-resource related languages without parallel data. *arXiv preprint arXiv:2105.15071*.

Sachin Kumar, Antonios Anastasopoulos, Shuly Wintner, and Yulia Tsvetkov. 2021. Machine translation into low-resource language varieties. *arXiv preprint arXiv:2106.06797*.

Jindřich Libovickỳ and Alexander Fraser. 2021. Findings of the wmt 2021 shared tasks in unsupervised mt and very low resource supervised mt. In *Proceedings of the Sixth Conference on Machine Translation*, pages 726–732.

Lei Liu and Min Zhu. 2022. Bertalign: Improved word embedding-based sentence alignment for Chinese–English parallel corpora of literary texts. *Digital Scholarship in the Humanities*, 38(2):621–634.

Cristian Marco Villanueva. 2012. Lengua aragonesa: Historia y situación actual.

Jesús Neira Martínez. 1982. La desaparición del romance navarro y el proceso de castellanización. *Revista española de lingüística*, 12(2):267–280.

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, Mercè Vàzquez, Marta Coll-Florit, Sergi Álvarez, Víctor Suárez, Claudi Aventín-Boya, Cristina Valdés, Mar Font, and Alejandro Pardos. 2023. Tan-ibe: Neural machine translation for the romance languages of the iberian peninsula. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 495–496.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Comput. Surv.*, 55(11).

Cecilio Lapresta Rey and Ángel Huguet Canalís. 2006. Identidad colectiva y lengua en contextos pluriculturales y plurilingües. el caso del valle de arán (lleida. españa). *Revista internacional de sociología*, 64(45):83–115.

R Sennrich, B Haddow, and A Birch. 2015. Improving neural machine translation models with monolingual data. arxiv 2015. *arXiv preprint arXiv:1511.06709*.

Rico Sennrich and Martin Volk. 2010. MT-based sentence alignment for OCR-generated parallel texts. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Research Papers*, Denver, Colorado, USA. Association for Machine Translation in the Americas.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. 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 1342–1348, Hong Kong, China. Association for Computational Linguistics.

Jörg Tiedemann. 2009. *News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces*, volume V, pages 237–248.

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

Marion Weller-Di Marco and Alexander Fraser. 2022. Findings of the wmt 2022 shared tasks in unsupervised mt and very low resource supervised mt. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 801–805.