Parameter-Efficient Fine-Tuning and Layer Analysis of Self-Supervised Speech Models: A Multilingual Study with ML-SUPERB

Carlos CUEVAS VILLARMIN, Oliver JACK,
Javier Alejandro LOPETEGUI GONZALEZ, Théo MOLFESSIS
ENS Paris-Saclay
April 9, 2025

Abstract

Self-supervised speech learning (SSL), as well as many other language related tasks in Deep Learning, has been extensively studied for high resources languages such as English, in detriment of the rest. However, recent works aim to overcome this limitation giving attention to a broader range of languages, including very low-resource ones. In particular, ML-SUPERB [8] extended the SUPERB benchmark (mainly in English) to 143 languages across. They evaluated models' performance for Automatic Speech Recognition (ASR) and Language Identification (LID). Moreover, they considered monolingual and multilingual scenarios for training and evaluation. In this work, we study and reproduce their pipeline, particularly for the case of the HuBERT-base [6] model. Moreover, following the idea from ML-SUPERB 2.0 [9] we explore additional PEFT techniques such as LoRA and Q-LoRA. Finally, we present a detailed summary of the obtained results, as well as performing analysis on layer importance and cross-lingual capabilities. Our work is publicly available at `https://github.com/olijacklu/ML-SUPERB-Project/tree/main`.

1. Introduction

Self-Supervised Learning (SSL) has significantly advanced speech processing, leveraging large unlabeled datasets to learn robust speech representations without explicit annotations [1, 6]. Such models have improved performance across various downstream tasks, including Automatic Speech Recognition (ASR) and Language Identification (LID) [11].

Traditionally, research and benchmarks in speech SSL have concentrated primarily on high-resource languages like English, limiting insights into multilingual and low-resource performance. Although the Speech processing Universal PERformance Benchmark (SUPERB) provided standardized evaluations, it remained largely focused on English tasks [10, 11]. To address this limitation, ML-SUPERB extended SUPERB to include 143 languages, covering both high-resource languages and endangered ones such as Totonac [8]. ML-SUPERB evaluated SSL models using frozen representations combined with lightweight downstream models on ASR and LID tasks, highlighting nuanced interactions between multilingual training data and model performance.

Despite its advances, ML-SUPERB has key limitations: it freezes SSL upstream models without considering potential improvements through fine-tuning and employs a fixed downstream architecture, neglecting variations that can influence SSL model rankings [9]. Recent studies, such as ML-SUPERB 2.0, highlight the benefits of partial fine-tuning strategies and more flexible downstream architectures for enhanced multilingual ASR performance [9].

In this work, we reproduce and extend ML-SUPERB, specifically focusing on the HuBERT-base model. Inspired by ML-SUPERB 2.0, we explore Parameter-Efficient Fine-Tuning (PEFT) methods such as LoRA and Q-LoRA. Additionally, we investigate cross-lingual transfer performance, assessing the impact of training in one language (e.g., Spanish) on a closely related language (e.g., Portuguese), and conduct a layer-wise analysis of SSL representations to understand language-specific distribution of learned weights.

2. Methodology

In this section we present a detailed description of the methodology used in our work. Section 2.1 contains the description of the data used in our experiments. Afterwards, we present in section 2.2 the three different fine-tuning strategies we experimented with. Finally, in section 2.3 we explain the evaluation procedure we followed, based on the ML-SUPERB [8] approach.

2.1. Data

We use the ML-SUPERB [8] dataset [1], which includes audios and transcripts in 143 languages from 14 sources. Our experiments primarily focus on six: Multilingual Librispeech (mls) [7], Spoken Wikipedia corpus (swc) [2], ALFFA corpus [5], M-AILab multilingual corpora, Mexican endangered languages [3], and Nordic Language Technology ASR corpora [4].

The dataset has 271 `(lang, source)` pairs, each with three 10-minute subsets for training, development (validation), and testing. A 1-hour training set exists but is unused as we focus on the 10-minute scenario. Each subset includes `.wav` files and transcripts.

For the monolingual track, we use seven of the nine ML-SUPERB languages, excluding Japanese and Chinese Mandarin due to differing settings. We select one `(lang, source)` pair per language, reducing from fourteen to seven. The same subset is used for multilingual experiments, covering `eng, fra, deu, rus, swa, swe, xty`. Data statistics are in Table 1.

Dataset	Hours	Selected Langs (7)
10-minute	1.17	∼10min × 7 (lang, data)
Dev	1.17	∼10min × 7 (lang, data)
Test	1.17	∼10min × 7 (lang, data)

Table 1. Statistics of the data used for training, development, and testing in our monolingual and multilingual experiments

Additionally, to evaluate the cross-lingual ASR capabilities, we used the Spanish, French and Portuguese data from the Multilingual Librispeech data, focusing once again on the 10 minutes subset.

2.2. Model Description and Fine-Tuning Methods

Following ML-SUPERB [8], our model consists of an SSL-pretrained feature extractor followed by a downstream transformer-based model. The extracted features are aggregated using a learnable weighted average, passed through a 1D convolutional layer, and fed to a downstream model with two transformer layers (attention dimension: 256, feedforward dimension: 1024, 8 attention heads). Due to computational constraints, we use this configuration without further exploration of alternatives proposed in ML-SUPERB 2.0 [9] for the downstream architecture.

Training follows the reference framework, using Adam optimizer with an initial learning rate of $1e-4$ and $1e-6$ weight decay. The batch size is set to 8 with gradient accumulation of 4. We train for 15,000 iterations in the monolingual track and 30,000 in the multilingual track, adjusting for hardware limitations.

We explore different fine-tuning approaches:

SSL model (frozen) + fine-tuning downstream architecture. The SSL model remains frozen while training only the downstream architecture, assuming the extracted features are robust enough for specific tasks just as in ML-SUPERB [8].

SSL model LoRA + downstream model fine-tuning. Following ML-SUPERB 2.0 [9], we apply LoRA for parameter-efficient fine-tuning (rank = 8, scaling factor = 32), modifying the upstream model with minimal computational overhead. The number of trainable parameters increases with respect to the previous approach but not significantly considering our hyperparameter choices.

SSL model Q-LoRA + downstream model fine-tuning. As an extension to the previous benchmarks, we consider the use of Q-LoRA, which integrates 4-bit quantization (via `BitsAndBytes`) with LoRA to further reduce memory usage while maintaining model performance. While it keeps constant the number of trainable parameters with respect to the LoRA approach, it significantly reduces the memory usage.

2.3. Tasks and Evaluation Framework

We make experiments on both the monolingual and multilingual tracks in a similar setting to ML-SUPERB [8]. For the monolingual case we work with the ASR task, while for the multilingual we additionally consider the LID task, as well as the joint ASR+LID task.

For evaluation, we are using the Character Error Rate (CER) in the case of the ASR task and accuracy for LID. Finally, we also compute the SUPERB [8] metric across tasks to have an idea of the overall performance. For the loss functions, we use CTC loss for ASR and the Cross-entropy loss for LID.

3. Experiments: Results and Discussion

In this section, we present the results obtained along with a thorough analysis. Section 3.1 reports the main results, including reproducibility, monolingual, and multilingual performance, based on the methods described in the previous section. In Section 3.2, we provide experiments aimed at analyzing the interpretability of the results. Finally, Section 3.3 includes an ablation study with models trained on a single language (Spanish, French, or Portuguese) and evaluated on the others, in order to assess the models' cross-lingual predictive capabilities based on the inherent similarity between the languages.

3.1. Fine-Tuning Results

The results presented in the tables reveal several interesting patterns across different models and fine-tuning approaches. In the monolingual ASR setup (Table 2), the HuBERT-base model achieves consistent performance across different fine-tuning methods, with average CER scores hovering

[1] ML-SUPERB data

Model	FT – Method	English	French	German	Russian	Swahili	Swedish	Mixtec	Avg	SUPERB$_s$
HuBERT-base	Standard	29.50	46.78	41.00	36.75	39.44	36.35	71.24	43.01/42.80	1015.85
	LoRA	29.46	48.09	41.54	36.83	37.23	38.48	70.99	43.23/-	1007.38
	QLoRA	29.56	47.68	41.76	36.78	38.03	34.80	71.17	42.83/-	1005.25
XLSR-128	QLoRA	51.61	48.83	48.47	40.49	41.12	44.16	70.30	49.28/-	937.08

Table 2. Monolingual ASR 10-minute set results. Each column represents the language used to fine-tune, as well as the CER metric. The average column presents our results / [9] results, if available.

around 43%. Meanwhile, the performance varies substantially across languages, with English showing the best results (CER around 29%) and Mixtec being the most challenging (CER above 70%). This is most likely due to the data quality and quantity that the upstream model was trained on. For the multilingual tasks (Table 3), we observe strong LID performance (accuracy of 94-99%) across all models, while ASR performance in the multilingual setting shows slightly higher CER compared to the monolingual average, suggesting some trade-offs when handling multiple languages simultaneously. The combined ASR+LID task shows relatively strong performance, particularly in terms of LID accuracy. The LID accuracies are higher than those reported in [8], most likely due to the reduced number of datasets that we evaluated our experiments on. Furthermore, the SUPERB scores calculated across our various experimental setups consistently outperformed those reported in the original paper.

Surprisingly, some of our findings contradict expectations based on previous research. While the original paper reports superior performance for the XLSR-128 model compared to HuBERT-base, our experiments show the opposite trend, with XLSR-128 yielding higher CER (49.28% vs. 43.01%) in the monolingual setting. Additionally, the LoRA and QLoRA fine-tuning methods, which modify the upstream model weights, unexpectedly perform worse than the standard frozen approach in several cases. These discrepancies can likely be linked back to the limited number of training iterations (15,000) compared to the recommended 25,000 warm-up steps suggested in the paper [9]. The low rank value (8) used in our LoRA implementations may have also constrained the model's capacity to adapt effectively to the downstream tasks.

Model	FT – Method	ASR (CER)	LID (Acc.)	ASR+LID		SUPERB$_s$
				CER	ACC	
HuBERT-base	Standard	42.67 / 39.80	97.00 / 61.20	44.94 / 39.20	100.00 / 71.50	1015.85
	LoRA	43.53	99.00	46.59	99.00	1007.38
	QLoRA	43.57	97.00	44.61	98.00	1005.25
XLSR-128	QLoRA	45.14	94.00	45.07	100.00	937.08

Table 3. Multilingual ASR, LID, and ASR+LID 10-minute set average results over target languages. Frozen row includes our results and those from [8].

3.2. Interpretability

3.2.1. Layer Weight Distribution

To go beyond purely measuring the models and the respective fine-tuning techniques solely based on performance metrics, an interesting extension is to analyse the learned weights associated to each individual layer in the upstream model. As proposed in the original ML-SUPERB paper, this weight distribution for a specific SSL model can be easily visualised using a heatmap, broken down by language.

The heatmap for the HuBERT-base model reveals distinct patterns in how different languages adapt to various layers of the model, you can see it in the Figure 1. In general, for this specific model and setup, there's a clear preference across languages for the early layers (particularly layer 0) and the later layers (8-11), with relatively less weight assigned to the middle layers (3-5). This suggests that the layer importance is divided between the low-level acoustic features from early layers and higher-level semantic representations from later layers, both contributing significantly to the model's performance. Breaking things down by language-specific patterns, Swahili (”swa”) and Mixtec (”xty”) show stronger weights for layer 0 compared to other languages, while English (”eng”) displays a unique pattern with substantial weights in the final layers (9-11). The relative uniformity of the middle layers (2-5) across most languages suggests that these encode more language-agnostic features that are less discriminative for specific languages.

In contrast to the HuBERT-base results, according to the original paper, the XLSR-128 model displays a remarkably different pattern with a strong preference for middle layers. In this context, it is worth mentioning that both models have very different architectures, as XLSR-128 has 24 layers compared to HuBERT-base's 12 layers. A possible explanation for this clear behavioral difference could be XLSR-128's deeper architecture, which may allow for more gradual feature abstraction, with the most useful representations emerging in middle layers rather than at the extremes. Meanwhile, both models show that they are well adapted to multilingual settings since they show consistent weights across the different languages per model.

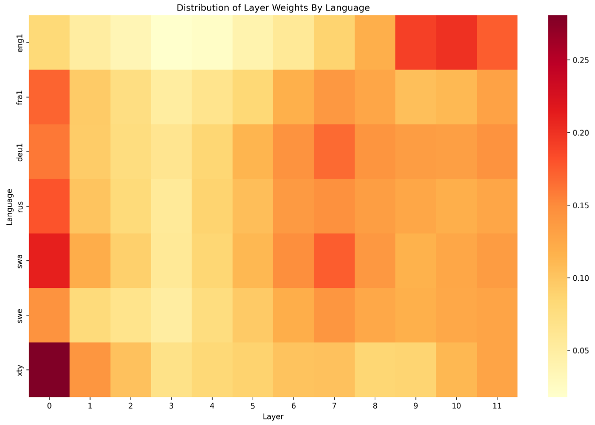


Figure 1. Layer importance heatmap for HuBERT-base model.

3.2.2. Language Identification Visualisation across Layers

For the specific task of language identification in the multimodal scenario, we consider the T-SNE projection of the logits obtained from different layers of the fine-tuned models (LID and ASR+LID): the last layer of the upstream model (HuBERT-base), and each of the layers added for language identification (two transformer layers and a linear layer). Figure 2 shows this evolution for the mentioned fine-tuned models.

It is noticeable that in both cases transformer layers logits catch up meaningful descriptions for language identification, being able to identify well-separated clusters for each given language which will allow to do the classification task without the last linear layer. On the other hand, in both LID (Figure 2a) and ASR+LID (Figure 2b) models the upstream model output (leftmost t-SNE plot) reveals that the HuBERT-base representations already show moderate language-specific clustering, even before any task-specific layers are applied. However, the clusters are less compact and more overlapping compared to later layers. This suggests that raw upstream representations are not yet optimal for language identification, but they provide a strong initialisation for subsequent fine-tuning. The presence of some structure without supervision supports the idea that self-supervised speech models learn latent linguistic properties.

3.3. Cross-Lingual Transfer Analysis

Cross-lingual zero-shot experiments can reveal interesting patterns in how well representations from one language transfer to others when the upstream model (here HuBERT-base) is kept frozen. For this we decided to train the combined model on one of three languages (Spanish, French and Portuguese), before testing it on the remaining two languages.

Looking at Figure 3, when trained on Spanish, the model achieved a lower mean CER score for Portuguese than for French, with the prior showing noticeably better transfer. This aligns with linguistic expectations, as Spanish and Portuguese are more closely related Romance languages than French. Similarly, when trained on Portuguese, the model performed better on Spanish than on French. Figure 4 provides qualitative examples comparing the prediction performance in French and Spanish, using the model fine-tuned on Portuguese.

Overall, these results demonstrate that HuBERT-base learns cross-lingual representations that are somewhat transferable between closely related languages even without explicit multilingual training, with Spanish and Portuguese showing the strongest mutual transfer due to their linguistic similarity.

4. Conclusion

In this work, we have explored different techniques to leverage the SSL pre-training on speech models on different tasks such as ASR and LID across different languages. We assessed the monolingual and multilingual capacities of these approaches. We extended previously applied PEFT techniques, with the use of the QLoRA approach. The best overall results were obtained for the originally proposed approach, i.e., freezing the pretrained model, then training the lightweight downstream model. It suggests that further analysis of hyperparameters selection for LoRA and QLoRA should be done in order to obtain the expected improvement in performance. Our layer analysis revealed a repetitive importance distribution in HuBERT-base with early layers capturing essential acoustic features and later layers encoding higher-level semantic representations, contrasting with XLSR-128's preference for middle layers. T-SNE visualisations revealed that while HuBERT-base representations already exhibit language clustering, transformer layers in our fine-tuned models progressively refine these into well-separated clusters, enhancing the linguistic properties inherently captured during self-supervised pretraining. The cross-lingual zero-shot experiments demonstrated effective transfer between Spanish and Portuguese, suggesting that HuBERT-base learns representations that capture linguistic similarities even when trained on monolingual data.

Future work should explore longer training iterations, higher LoRA ranks, and expanded language diversity to better understand the relationship between parameter-efficient fine-tuning methods and cross-lingual transfer capabilities in self-supervised speech models.

References

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *NeurIPS*, 2020. 1

[2] Timo Baumann, Arne Köhn, and Felix Hennig. The spoken wikipedia corpus collection: Harvesting, alignment and an application to hyperlistening. *Language Resources and Evaluation*, 53:303–329, 2019. 2

[3] Dan Berrebbi, Jiatong Shi, Brian Yan, Osbel López-Francisco, Jonathan D Amith, and Shinji Watanabe. Combining spectral and self-supervised features for low resource speech recognition and translation. *arXiv preprint arXiv:2204.02470*, 2022. 2

[4] Koenraad De Smedt, Georg Rehm, and Hans Uszkoreit. *The Norwegian language in the digital age*. Springer, 2012. 2

[5] Nic J De Vries, Marelie H Davel, Jaco Badenhorst, Willem D Basson, Febe De Wet, Etienne Barnard, and Alta De Waal. A smartphone-based asr data collection tool for under-resourced languages. *Speech communication*, 56:119–131, 2014. 2

[6] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. 1

[7] Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. Mls: A large-scale multilingual dataset for speech research. In *Interspeech 2020*. ISCA, 2020. 2

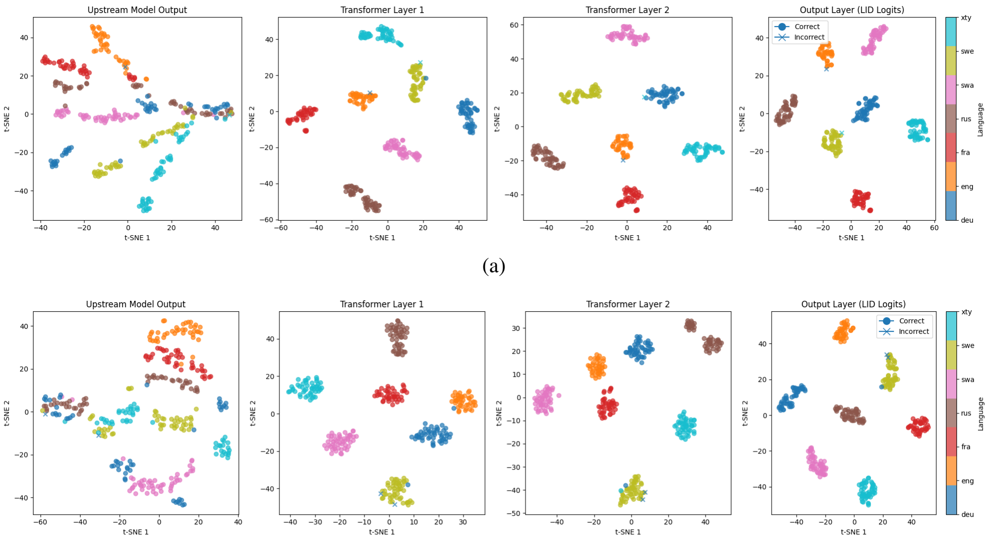
[8] Jiatong Shi, Dan Berrebbi, et al. Ml-superb: Multilingual speech universal performance benchmark. In *INTERSPEECH*, 2023. 1, 2, 3

[9] Jiatong Shi, Shih-Heng Wang, et al. Ml-superb 2.0: Benchmarking multilingual speech models across modeling constraints, languages, and datasets, 2024. 1, 2, 3

[10] Hsuan-Shao Tsai et al. Superb-sg: Enhanced speech processing universal performance benchmark for semantic and generative capabilities. In *ACL*, 2022. 1

[11] Shu-wen Yang et al. Superb: Speech processing universal performance benchmark. In *INTERSPEECH*, 2021. 1

A. Appendix



(a)

(b)

Figure 2. **t-SNE visualizations of language representations across different model layers**.2a for fine-tuned model for LID and 2b for the ASR-LID case. Each point represents a sample colored by target language. Circles indicate correct classifications and crosses indicate errors.

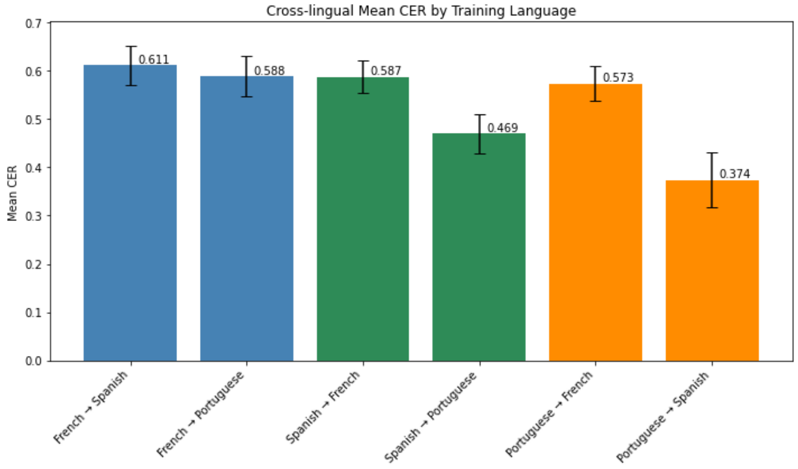
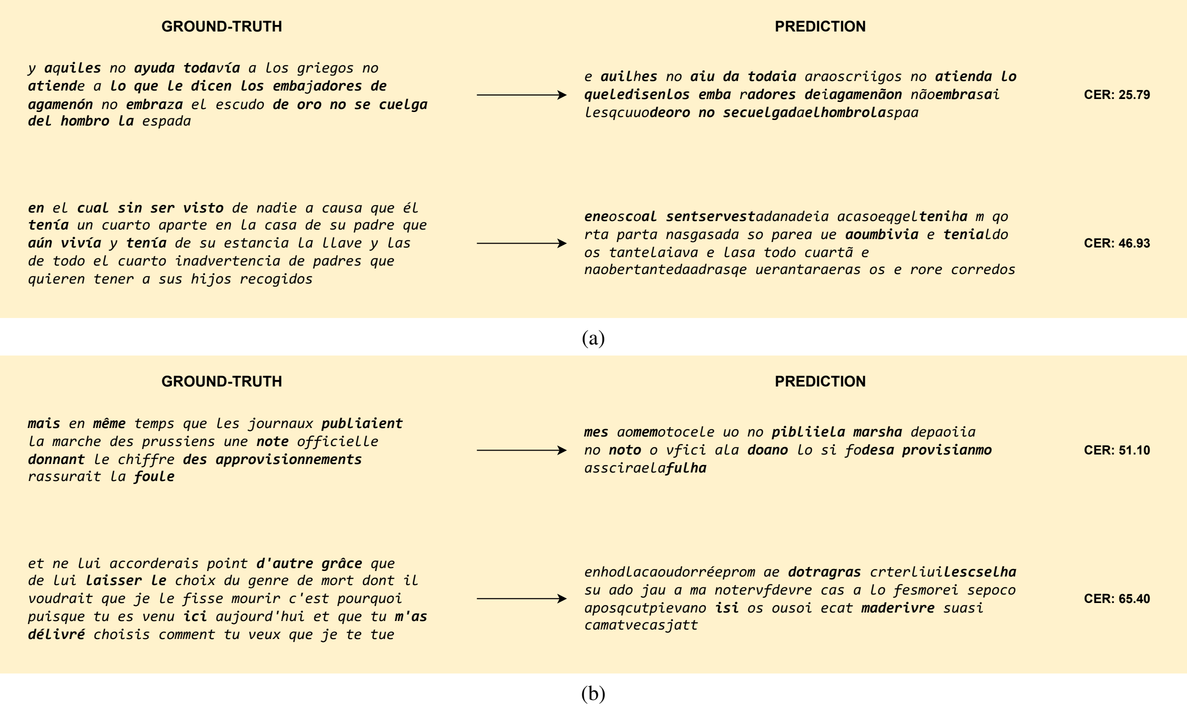


Figure 3. Cross-lingual mean CER with standard deviation across models trained on French (blue), Spanish (green), and Portuguese (orange).



GROUND-TRUTH		PREDICTION	
y **aquiles** no **ayuda toda**vía a los griegos no **atiend**e a **lo que le dicen los embajadores de agamenón** no **embraza** el escudo **de oro no se cuelga del hombro la** espada	→	e **auilhes** no **aiu da todaia** araoscriigos no **atienda lo queledisenlos emba** r**adores de**i**agamenãon** não**embras**ai lesqcuuo**deoro no secuelgad**a**elhombrol**aspaa	CER: 25.79
en el **cual sin ser visto** de nadie a causa que él **tenía** un cuarto aparte en la casa de su padre que **aún vivía** y **tenía** de su estancia la llave y las de todo el cuarto inadvertencia de padres que quieren tener a sus hijos recogidos	→	**ene**os**coal sentservest**adanadeia acasoeqgel**teniha** m qo rta parta nasgasada so parea ue **aoumbivia** e **tenia**ldo os tantelaiava e lasa todo cuartã e naobertantedaadrasqe uerantaraeras os e rore corredos	CER: 46.93

(a)

GROUND-TRUTH		PREDICTION	
mais en **même** temps que les journaux **publiaient** la marche des prussiens une **note** officielle **donnant** le chiffre **des approvisionnements** rassurait la **foule**	→	**mes** aome**m**otocele uo no **pibliiela marsha** depaoiia no **noto** o vfici ala **doano** lo si fo**desa provisianmo** assciraela**fulha**	CER: 51.10
et ne lui accorderais point **d'autre grâce** que de lui **laisser le** choix du genre de mort dont il voudrait que je le fisse mourir c'est pourquoi puisque tu es venu **ici** aujourd'hui et que tu **m'as délivré** choisis comment tu veux que je te tue	→	enhodlacaoudorréeprom ae **dotragras** crterliui**lescselha** su ado jau a ma notervfdevre cas a lo fesmorei sepoco aposqcutpievano **isi** os ousoi ecat **maderivre** suasi camatvecasjatt	CER: 65.40

(b)

Figure 4. Example outputs from cross-lingual evaluation with a model trained on Portuguese and evaluated in Spanish 4a and French 4b.