Team QUESPA System Submission for the IWSLT 2026 Dialectal and Low-resource Speech Translation Task

John E. Ortega[1], **Rodolfo Zevallos**[2], **Fabrício Carraro**[3],
Stephanny Sánchez[4], **Lewis C. Howe**[5]
[1]Northeastern University, USA, [2]Universitat Pompeu Fabra, Spain,
[3]Barcelona Supercomputing Center, Spain, [4]Universidad Nacional de Ingeniería, Peru
[5]Universtiy of Georgia, USA
contact email: j.ortega@northeastern.edu

Abstract

This paper describes the **QUESPA** team's speech translation (ST) submissions for the Quechua to Spanish (QUE–SPA) track of the IWSLT 2026 Evaluation Campaign on dialectal and low-resource speech translation. The campaign supports a single submission category, namely *unconstrained*. This marks our fourth consecutive participation in the IWSLT shared task, building upon prior systems with substantial improvements. Our 2026 submission comprises three *unconstrained-only* systems. The best-performing system (*contrastive 2*) extends our strongest model from the previous year by leveraging a high-performing pre-trained language model (PLM) for end-to-end speech translation without cascading, augmented with additional *Quechua–Collao* text – now made available on the IWSLT GitHub. Fine-tuning Microsoft's SpeechT5 model in an ST setting, combined with targeted data augmentation, results in a BLEU score of 27.2 on the official evaluation set. Additionally, we evaluate prompt-based machine translation using Gemini, DeepSeek, GPT-5, Claude, and Qwen for the first time. Aside from that, we introduce **SIDON**, an audio enhancement framework designed to improve audio quality. This paper provides a comparative analysis across our current and three previous IWSLT submissions, with a detailed examination of the impact of synthetic data, unconstrained external resources, and audio enhancement techniques on fine-tuning performance. Our results highlight the complementary role of PLM-based ST, LLM prompting, and ASR enhancement in advancing low-resource speech translation.

1 Introduction

In this article, we describe our three systems that were submitted to the IWSLT 2026 Low-Resource Track for Speech Translation (ST) (**?**). The IWSLT task is particularly challenging for low-resource languages (LRLs) due to the lack of data needed to create, or even fine-tune, a pre-trained language model (PLM). While many problems are solvable with privatized large-language model (LLM) prompting approaches, our experiments show that for Quechua–Spanish zero-to-few shot approaches outperform those approaches.

Here, we describe three main approaches that extend previous approaches submitted in the past three iterations of IWSLT (Agostinelli et al., 2025; Ahmad et al., 2024; Agarwal et al., 2023) where the best score achieved for ST until this publishing based on BLEU (Papineni et al., 2002) for the Quechua to Spanish task was: 26.7, submitted by the **QUESPA** team.

Quechuan languages represent a family of Indigenous language spoken by more than 8 million people in the Andean region of South America. They are primarily spoken in Peru, Ecuador, and Bolivia where the colonial high-resource language (HRLs) is Spanish. Quechuan languages are highly agglutinative, with a large inventory suffixes (e.g., tense, person, evidentiality) that is structurally similar to languages like Finnish and Turkish. It is worthwhile to note that previous work (Barbaran and Ticona, 2024; Ahmed et al., 2023; Ortega et al., 2020; Ortega and Pillaipakkamnatt, 2018) has had some success in identifying the inflectional properties of Quechuan languages where the addition of an HRL, namely Finnish, can aid for translation purposes achieving nearly 20 BLEU on liturgical (text-only) tasks. The average number of morphemes per word (synthesis) is approximately two times larger than English, a non-agglutinative language. English typically has around 1.5 morphemes per word, and Quechua has roughly 3 morphemes per word. There are two main regional divisions of Quechua known as Quechua I and Quechua II (Torero, 1964). This data set consists of two main types of Quechuan languages spoken in Ayacucho, Peru (Quechua Chanka ISO:quy) and Cusco, Peru (Quechua Collao ISO:quz) which are

Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026), pages 47–57

typically classified as Quechua II and, thus, considered a "southern" languages. We label the data set with que - the ISO norm for Quechua II varieties.

The **QUESPA** team this year consists of four organizers from five different institutions: Northeastern University (USA), Pompeu Fabra University (Spain), Barcelona Super Computing Center, University of Alberta (Canada) and University of Georgia (USA). Three new organizers have been introduced this year who have linguistic expertise in Quechua, prompting, and automatic speech recognition (ASR). This year, only two of the organizers from the IWSLT 2025 have continued to work on the project. Two of the five total organizers have had experience with the QUE–SPA language pair in the past and submitted have already submitted three times to IWSLT, making this article the fourth submission with an increase of BLEU score for each year's submission. We report the QUESPA consortium submission for the IWSLT 2026 task (**?**) and once again focus on the low-resource task at hand by combining *all* the two dialects *Quechua I and II* into one. However, we specifically make use of the Quechua II variant in Collao (ISO:quz), given the addition of a new corpus to our data set, made more explicit and official this year as an available corpus on Github.[1]

The remainder of this article is structured as follows. Section 2 presents the related work. Since we intend to highlight the addition of our machine translation (MT) comparisons and systems by a new collaborator, we provide an overview specifically of the MT delivery in Section 3.1. Following that, experiments for the for QUE–SPA low-resource track are presented in Section 3 with their results explained in Section 4.

2 Related Work

In this section, we first cover the different approaches used in previous speech processing shared tasks for Quechua (Section 2.1). We then discuss prior work that utilized a similar strategy to our primary submission to the unconstrained track (Section 2.2).

2.1 Quechua Speech Processing

The previous iteration of IWSLT (Agarwal et al., 2023) was the first time that Quechua–Spanish was featured in the low-resource ST track. Due to the small amount of available paired data, the participants focused on exploiting PLMs for speech and/or text in the unconstrained track. The teams all converged on using XLS-R 128 (Babu et al., 2021) as the pre-trained speech encoder, while NLLB 200 (NLLB Team et al., 2022) was the most popular text PLM. However, the teams used the PLMs in very different manners. QUESPA (E. Ortega et al., 2023) separated the PLMs into distinct systems for an ASR+MT cascade, GMU (Mbuya and Anastasopoulos, 2023) performed full fine-tuning on XLS-R for direct ST, and NLE (Gow-Smith et al., 2023) combined the two PLMs via adapter fine-tuning. By using PLMs for both the input and output modalities, NLE and QUESPA obtained the best performances at 15.7 and 15.4 BLEU respectively. For the constrained track, developing a usable system was more difficult to achieve. In this setup, the best performing model was a direct ST system by GMU that achieved 1.46 BLEU. The QUESPA team adopted a near-identical strategy to achieve 1.25 BLEU.

Quechua–Spanish ST was also featured as part of a similar competition in the 2022 edition of AmericasNLP (Ebrahimi et al., 2022). Similar to IWSLT 2023, participants experimented with different ways of leveraging PLMs. XLS-R and NLLB were popular choices, but teams also experimented with DeltaLM (Ma et al., 2021) and Whisper (Radford et al., 2023).

Quechua was most recently part of the 2023 ML-SUPERB Challenge (Shi et al., 2023), which tasked participants on evaluating different self-supervised (SSL) speech encoders on long-tailed languages. Chen et al. (2023a) found that XLS-R 128 outperformed all other SSL encoders on Quechua, further validating its popularity in the other competitions. Additional competitions, such as the most recent iteration of AmericasNLP and the Networking Symposium on Latin America are focused on encouraging innovation and expansion of tools related to the languages of the Americas utilizing cutting-edge methods like Omnilingual MT (Team et al., 2026). Additionally, these tools have been applied in medical settings (Carrillo-Larco et al., 2026), though, to our knowledge, no such studies with interactions in Quechua have been conducted.

2.2 Multilingual Speech Processing

Multilingual training is a common strategy to facilitate cross-linguistic transfer learning, with the goal of boosting performance on LRLs. While this

[1] https://github.com/johneortega/IWSLT2026_Quechua_data

is generally done by pairing HRLs with LRLs, it can also be beneficial in settings where only LRLs are available. Chen et al. (2023b) trained multilingual ASR systems on 102 languages, each in a low-resource setting, and obtained state-of-the-art (SOTA) results on the FLEURS benchmark (Conneau et al., 2023). Radford et al. (2023) and Peng et al. (2023) then combined multilingual ASR and ST at scale, developing SOTA models through supervised training on hundreds of thousands of audio samples. Our strategy for the unconstrained track constitutes a combination of these two methods, enhancing performance on Quechua–Spanish using multilingual ST training with other LRLs.

3 Quechua-Spanish

In this section we present our experiments for the QUE–SPA dataset provided in the low-resource ST track at IWSLT 2026[2], identical to the dataset from IWSLT 2025 with the exception of new additions of a "potential" good-for-use corpus. The audio consists of consists of 1 hour and 40 minutes of *unconstrained* speech along with its corresponding translations in addition to nearly 48 hours of ASR data (with transcriptions) from the Siminichik (Cardenas et al., 2018) corpus. Additionally, an MT dataset is provided from previous neural MT work (Ortega et al., 2020). The audio and corresponding transcriptions along with their translations are mostly made of radio broadcasting from the Andes region of Peru. This dataset has been used in other tasks but not in its entirety (de Giberto et al., 2025; Ebrahimi et al., 2023, 2022; Zevallos et al., 2022a). This year there has been a new addition to the dataset provided by the task which is a machine-translated and post-edited text of the Huqariq corpus (Zevallos et al., 2022b) that was used last year by this team (Ortega et al., 2025) for augmentation of the best performing T5 model (Raffel et al., 2020). On top of that, the Collao-based corpus (iso: quz) used last year for improvements has been introduced as a dataset on the competition website. (Paccotacya-Yanque et al., 2022)

We present the three submissions for *unconstrained* task again as this year's competition has abandoned the *constrained* task:

1. a **primary unconstrained** system consisting of a Mamba ASR model (Zhang et al., 2024) after applying audio noise removal with SIDON (Nakata et al., 2025) then fine-tuned with unconstrained data and cascaded the best performing NLLB MT system from our case study;

2. a **contrastive 1 unconstrained** system consisting of a Whisper Large V3 (Radford et al., 2023) ASR model after applying audio noise removal with SIDON (Nakata et al., 2025) then fine-tuned with the unconstrained data and cascaded with the best performing NLLB MT system from our case study;

3. a **contrastive 2 unconstrained** system consisting of a SpeechT5 model (Ao et al., 2021) after applying audio noise removal with SIDON (Nakata et al., 2025) then fine-tuned for speech translation with two data augmentation techniques and an additional newly introduced corpus based on Quechua Collao (iso: quz) (Paccotacya-Yanque et al., 2022).

We present the experimental settings and results for unconstrained systems starting off with the MT case studies in Section 3.1. Then, we describe the unconstrained task details further in Section 3.2. Primary, Contrastive 1 and Contrastive 2 descriptions are found in Sections Sections 3.3, 3.4 and 3.5, respectively. Afterwards, we offer results and discussion in Section 4.

3.1 Machine Translation

We compare various MT systems based on prompting alone to the original (baseline) system from Ortega et al. (2025) which was trained by fine-tuning the 1.3B parameter version[3] of the NLLB_200 (NLLB Team et al., 2022) and set to a maximum token lengths of 128 for both inputs and outputs. Their model was trained for 10 epochs with a batch size of 8 for both training and evaluation, using 5 beams during generation saving model checkpoints every 10,000 steps and set a random seed of 65 to ensure reproducibility.

This year, the QUESPA team decided to experiment with several prompting approaches. We consider this a new paradigm that is by far the most widely-used paradigm for most NLP tasks and, thus, warranted its use. Last year's baseline a high performing MT system with scores of: **19.5 BLEU** and **23.5 CHRF**. Since the novel Omnilingual MT (Team et al., 2026) system appears to use

[2] https://github.com/johneortega/IWSLT2026_Quechua_data

[3] https://huggingface.co/facebook/nllb-200-1.3B

Model-[Data]	BLEU	CHRF
	Test	Test
2025 NLLB Baseline	**19.5**	**23.5**
GPT 5.4	3.7	25.5
GPT 5.4 Extended	7.6	44.5
Gemini 3 Flash	10.8	48.3
Gemini 3 Flash Reasoning	7.2	38.2
Gemini 3 Pro	8.5	44.0
Claude Haiku 4.5	6.1	41.3
Claude Haiku 4.5 Extended	2.3	26.9
Claude Sonnet 4.6	8.0	46.7
Deep Seek-V3.1	7.7	45.9
Deep Seek-V3.1 DeepThink	7.6	45.5
Deep Seek-V3.1 DeepThink & Smart	7.2	45.7
Deep Seek-V3.2 Expert	7.9	47.5
Deep Seek-V3.2 Expert Deep Think	10.4	49.0
Deep Seek-V3.2 Expert Deep Smart Think	7.5	47.4
Qwen 3.6 Plus	5.0	44.3
Qwen 3.6 Plus Thinking	7.3	47.4
Qwen 3.6 Plus Fast	6.4	42.8

Table 1: Performance of several prompting models on the test set compared to the 2025 QUESPA NLLB Baseline (Ortega et al., 2025).

an identical architecture and corpora in Quechua as the NLLB system, we focus on prompt-based approaches for MT only.

The prompts used for MT are provided in Appendix A–Appendix E. These prompts reflect the different arrangements displayed in Table 1. It should be noted that the prompts were provided in Spanish, exactly as they are listed in the appendices, for each system listed in Table 1, we run our benchmarking code using a Google Colab notebook [4] made available publicly. We report on the best-performing prompt over the **Test** set for comparison to last year's baseline.

Our experiments show that for the QUE–SPA task, prompting does not perform as well as the NLLB baseline. Several takeaways can be noted from our prompting experiments in the following. The GPT-5 (Singh et al., 2025) models tended to use those fields as implicit references, even when the prompt explicitly instructed them not to consider them; therefore, it can be inferred that these models *may* prioritize signals in Spanish over low-resource languages such as Quechua. GPT-5 sometimes filled in missing content, which caused a loss of fidelity because it began mixing variants and adding information from previous sentences, reducing semantic precision. Hallucinations were observed in all prompting models experiment (see Table 1 for the list of models used), as models generated content based on previously learned patterns; when the prompt explicitly instructed it not to hallucinate, its behavior improved, although full compliance was not guaranteed. GPT-5 worked well as a support tool for short words, but it did not seem to be reliable for longer paragraphs or sentences in Quechua. Gemini 3 (Team et al., 2023) models achieved the best results overall; however, there were dialectal biases and monolithic macrolanguage difference, mixing features from different dialects without clearly distinguishing their boundaries, although it demonstrates stronger syntactic handling, likely due to its multilingual training. Some phrases related to sensitive topics such as racism or death caused slower processing and led the system to modify the original context of the main idea in Gemini 3. Claude models[5] performed poorly and were less precise when working with the Collao dialect, as it appears to be more ex-

[4] https://colab.research.google.com/drive/1_BxYPh_3QzaqOzOGZ9mBBxqWVcB0S3tc?usp=sharing

[5] https://platform.claude.com/docs/en/about-claude/models/overview

posed to other variants such as Cusco and Chanca Quechua. DeepSeek (Liu et al., 2024), which relies on cross-lingual transfer and synthetic data, showed robustness but tended toward literal translations, lacking cultural context and a deeper understanding of the Andean worldview. Qwen (Bai et al., 2023), selected due to its widespread use at the time of the benchmark, showed limitations related to the scarcity of high-quality data, making frequent errors in proverbs, poetry, and culturally rich contexts. Overall, these results suggest the need to improve dataset design so that it can be better understood by language models and to refine prompts so that the model clearly identifies which linguistic variety to use.

Despite our hardest efforts, we were unable to achieve BLEU/CHRF scores that surpassed last year's baseline for MT. In future work, we plan on trying more complex prompting techniques like those used by Team et al. (2026), for our final cascade experiments, we continue to use the fine-tuned NLLB MT system from (Agostinelli et al., 2025) for two of our unconstrained systems.

3.2 Unconstrained Setting

As in IWSLT 2025, the organizers provided a total of 48 hours of audio along with their corresponding transcriptions. We also translated the 48 hours of audio provided by the organizers into Spanish. Furthermore, we utilized a portion of the AmericasNLP[6] (ANLP) 2022 speech translation competition corpus, which consists of 19 minutes of Guarani and 29 minutes of Bribri, fully translated into Spanish. Though not a Quechua corpus, these languages have morphological similarities with Quechua, and, as such, we decided to determine see if their inclusion would improve our models. We used the data set provided by the organizers from previous work on Quechua Collao (Paccotacya-Yanque et al., 2022). As mentioned in Ortega et al. (2025), it is characterized as belonging to the Quechua II subfamily. Finally, all the datasets described in this section allowed for further fine-tuning of the previously trained end-to-end speech translation model and our audio files were transformed using SIDON (Nakata et al., 2025), an addition that was not presented in 2025.

The speech enhancement pre-processing stage used SIDON (Nakata et al., 2025) to improve the acoustic quality of the training data prior to model optimization. SIDON effectively reduces background noise, reverberation artifacts, and channel distortions, leading to more robust and consistent input representations for downstream speech models. In our pipeline, all audio samples are normalized and enhanced before being used to train SpeechT5, ConMamba, and Whisper v3. This pre-processing step is particularly critical in low-resource and heterogeneous datasets, where variability in recording conditions can significantly degrade model performance. Prior work has shown that front-end enhancement and large-scale robustness strategies improve both intelligibility and generalization in ASR and TTS systems, especially under noisy or mismatched conditions.

3.3 Primary System

The Primary System for the unconstrained setting consists of a cascaded architecture, where the output of an automatic speech recognition (ASR) model is passed as input to a machine translation (MT) model. For the ASR component, we employ ConMamba (Jiang et al., 2024), a recent extension of the Mamba architecture that integrates convolutional modules into its encoder blocks, inspired by Conformer (Gulati et al., 2020). This hybrid design enhances the model's ability to capture both global and local dependencies. The encoder architecture comprises a sequence of modules: an initial feedforward layer with residual connection, a bidirectional Mamba module (BiMamba) for long-range dependency modeling, a convolutional layer for local context enhancement, and final layer normalization and refinement through another feedforward module (Tang et al., 2024). This combination results in a balanced and efficient encoding mechanism for speech signals.

On the decoder side, we incorporate CrossMamba, a unidirectional variant tailored for sequential processing without native cross-attention. CrossMamba simulates cross-attention by concatenating key and query sequences, retaining only the relevant portion of the output. This mechanism allows for effective integration of encoder context through a structured decoding pipeline: normalization, unidirectional Mamba (UniMamba), a second normalization step, CrossMamba integration, and a final feedforward refinement. We train both ConMamba and Conformer models using publicly available recipes[7], experimenting with small (S)

[6] https://turing.iimas.unam.mx/americasnlp/2022_st.html

[7] https://github.com/xi-j/Mamba-ASR

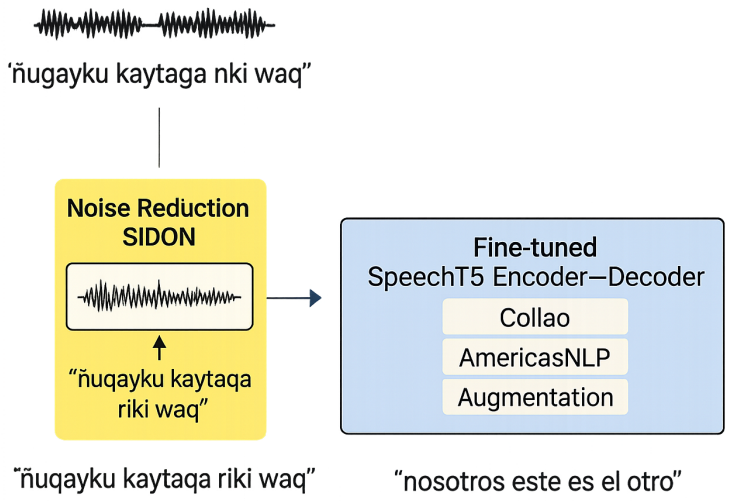

Figure 1: High-level system overview of Contrastive System 2, a fine-tuned SpeechT5 (Ao et al., 2021) model with SIDON (Nakata et al., 2025) applied to the audio file.

and large (L) configurations (144/512 dimensions, 12+4/12+6 layers). Training is performed over 110 epochs using AdamW with a Noam scheduler (30k warm-up steps), and audio is tokenized with a BPE tokenizer trained for each language using SpeechBrain[8]. Once the speech is transcribed, we feed the resulting text into the machine translation model previously described, leveraging its capabilities to produce the final translated output in a cascaded speech translation setup.

3.4 Contrastive 1 System

The Contrastive 1 system is a simple ASR+MT cascade. We develop the ASR module by fine-tuning Whisper Large V3 (Radford et al., 2023) on the entire 48 hours of unconstrained Quechua ASR data in ESPnet (Watanabe et al., 2018). Whisper consists of a Transformer encoder and Transformer decoder (Vaswani et al., 2017). The bidirectional encoder receives mel audio features as input, whereas the decoder is conditioned on a language identity tag and the encoder output. The model is trained for 22K steps with the Adam optimizer (Kingma and Ba, 2015). We use a scheduler that linearly warms up the learning rate to a peak value of 1e-5 for 1500 steps, followed by exponential decay for the remainder of training (Vaswani et al., 2017). ASR inference is performed with greedy decoding, the results of which are then passed to the NLLB-based MT model described in Section 3.1.

3.5 Contrastive 2 System

The Contrastive 2 System for the unconstrained setting consists of a pre-trained model called SpeechT5 (Ao et al., 2022) which was trained on 960 hours of audio from LibriSpeech. SpeechT5 consists of 12 Transformer encoder blocks and 6 Transformer decoder blocks, with a model dimension of 768, an internal dimension (FFN) of 3,072, and 12 attention heads. Additionally, the voice encoder’s pre-net includes 7 blocks of temporal convolutions. Both the pre-net and post-net of the voice decoder used the same configuration as in Shen et al. (2018), except that the number of channels in the post-net is 256. For the text encoder/decoder’s pre/post-net, a shared embedding layer with a dimension of 768 is utilized. For vector quantization, two codebooks with 100 entries each are used for the shared codebook module. The model was trained using the normalized training text from the LibriSpeech language model as unlabeled data, which contains 400 million sentences. Training was optimized using Adam (Kingma and Ba, 2015), with a learning rate that linearly increases during the first 8% of updates up to a maximum of 0.0002.

We fine-tuned SpeechT5[9] for Speech Translation using the SpeechT5 fine-tuning recipe[10] for Speech-Translation with the same hyperparameter settings. We used the 48 hours of audio provided by

[8] https://github.com/speechbrain/speechbrain/tree/develop/recipes/LibriSpeech/Tokenizer

[9] https://github.com/microsoft/SpeechT5

[10] https://github.com/microsoft/SpeechT5/tree/main/SpeechT5

Team **QUESPA** BLEU and CHRF Scores			
Unconstrained 2026			
System	**Description**	**BLEU**	**CHRF**
primary	mamba asr + sidon + nllb mt	15.0	50.7
contrastive 1	whisper-v3 asr + sidon + nllb mt	15.4	52.0
contrastive 2	speechT5 + sidon + anlp + da-tts + nlpaug* + quz	27.2	51.4
Unconstrained 2025			
System	**Description**	**BLEU**	**CHRF**
primary	mamba asr + nllb mt	14.8	51.8
contrastive 1	whisper-v3 asr + nllb mt	15.0	52.4
contrastive 2	speechT5 + anlp + da-tts + nlpaug* + quz	26.7	48.6

Table 2: Team QUESPA results for the Quechua to Spanish low-resource task at IWSLT 2026 compared to IWSLT 2025 (Agostinelli et al., 2025).

the organizers (anlp). We applied a data augmentation technique called *nlpaug* (noise, distortion, duplication)[11] (Ma, 2019), resulting in a total of 96h: 48h original + 48h synthetic data + 15 hours of Quechua Collao (Paccotacya-Yanque et al., 2022) (quz) .

4 Results and Discussion

Results are presented in Table 2. When compared to IWSLT 2025 (Agostinelli et al., 2025; Ortega et al., 2025), it is clear that Speech Translation as a task for Quechua to Spanish translation is best performed using the Speech T5 model which is a non-cascade model. Through the use of SIDON (Nakata et al., 2025) and fine-tuning a Speech T5 model, we were able to increase the score by 0.5 BLEU points to the previous year's results. On the other hand, for machine translation, baseline prompting systems found online do not outperform the NLLB baseline. Other approaches from systems such as Whisper and Mamba were able to gain considerably (on average about 0.4 BLEU) from noise reduction in SIDON also. However, CHRF scores were not improved, they were worsened for the SpeechT5 models despite the BLEU score improvements. Contrastingly, ST systems based on Mamba and Whisper did find gain on CHRF with the use of SIDON.

5 Conclusion and Future Work

Our submission to the IWSLT 2026 () evaluation campaign for low-resource and dialect speech translation includes novelties based on the most state-of-the-art techniques for ASR and ST. The addition of three new characteristics: 1) the Quechua Collao corpus (referred to as quz), 2) SIDON noise reduction and 3) a machine translation case study with guided prompts illustrate the latest possibilities. These three new inclusions have brought to light what MT systems, corpora, and ASR models work best with the language pair when compared to previous year's work.

In subsequent competitions, we plan to include more human annotation and experimentation with the model presented here since the BLEU score achieved (27.2) warrant further investigation. Lastly, other MT systems, such as Omnilingual MT (Team et al., 2026) have shown promising results which we plan to try next year.

[11] https://github.com/makcedward/nlpaug

References

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Yannick Estève, Kevin Duh, Marcello Federico, Souhir Gahbiche, Barry Haddow, Benjamin Hsu, Phu Mon Htut, Hirofumi Inaguma, Dávid Javorský, John Judge, Yasumasa Kano, Tom Ko, Rishu Kumar, Pengwei Li, Xutail Ma, Prashant Mathur, Evgeny Matusov, Paul McNamee, John P. McCrae, Kenton Murray, Maria Nadejde, Satoshi Nakamura, Matteo Negri, Ha Nguyen, Jan Niehues, Xing Niu, Atul Ojha Kr., John E. Ortega, Proyag Pal, Juan Pino, Lonneke van der Plas, Peter Polák, Elijah Rippeth, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Yun Tang, Brian Thompson, Kevin Tran, Marco Turchi, Alex Waibel, Mingxuan Wang, Shinji Watanabe, and Rodolfo Zevallos. 2023. Findings of the IWSLT 2023 Evaluation Campaign. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*. Association for Computational Linguistics.

Victor Agostinelli, Tanel Alumäe, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Fethi Bougares, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, et al. 2025. Findings of the iwslt 2025 evaluation campaign. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 412–481.

Ibrahim Said Ahmad, Antonios Anastasopoulos, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, William Chen, Qianqian Dong, Marcello Federico, Barry Haddow, Dávid Javorský, Mateusz Krubiński, Tsz Kin Lam, Xutai Ma, Prashant Mathur, Evgeny Matusov, Chandresh Maurya, John McCrae, Kenton Murray, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, Atul Kr. Ojha, John Ortega, Sara Papi, Peter Polák, Adam Pospíšil, Pavel Pecina, Elizabeth Salesky, Nivedita Sethiya, Balaram Sarkar, Jiatong Shi, Claytone Sikasote, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Brian Thompson, Alex Waibel, Shinji Watanabe, Patrick Wilken, Petr Zemánek, and Rodolfo Zevallos. 2024. FINDINGS OF THE IWSLT 2024 EVALUATION CAMPAIGN. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 1–11, Bangkok, Thailand (in-person and online). Association for Computational Linguistics.

Nouman Ahmed, Natalia Flechas Manrique, and Antonije Petrović. 2023. Enhancing spanish-quechua machine translation with pre-trained models and diverse data sources: Lct-ehu at americasnlp shared task. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 156–162.

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu Wei. 2022. SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5723–5738, Dublin, Ireland. Association for Computational Linguistics.

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, et al. 2021. Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing. *arXiv preprint arXiv:2110.07205*.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Bruno Barbaran and Wilfredo Ticona. 2024. Implementation of a natural language processing model for spanish-quechua machine translation. In *Proceedings of the Computational Methods in Systems and Software*, pages 464–476. Springer.

Ronald Cardenas, Rodolfo Zevallos, Reynaldo Baquerizo, and Luis Camacho. 2018. Siminchik: A speech corpus for preservation of southern quechua. *ISI-NLP 2*, page 21.

Rodrigo M. Carrillo-Larco, Jesús Lovón-Melgarejo, Manuel Castillo-Cara, and Gusseppe Bravo-Rocca. 2026. PeruMedQA: Benchmarking Large Language Models (LLMs) on Peruvian Medical Exams—Dataset Construction and Evaluation. *Medical Science Educator*.

Chih-Chen Chen, William Chen, Rodolfo Zevallos, and John Ortega. 2023a. Evaluating self-supervised speech representations for indigenous american languages. *arXiv preprint arXiv:2310.03639*.

William Chen, Brian Yan, Jiatong Shi, Yifan Peng, Soumi Maiti, and Shinji Watanabe. 2023b. Improving massively multilingual asr with auxiliary CTC objectives. *arXiv preprint arXiv:2302.12829*.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805.

Ona de Giberto, Robert Pugh, Ali Marashian, Raúl Vázquezo Abteen Ebrahimi, Pavel Denisov, Enora Rice, Edward Gow-Smithγ, Juan C Prietoβ, Melissa Roblesβ, Rubén Manriqueβ, et al. 2025. Findings of the americasnlp 2025 shared tasks on machine translation, creation of educational material, and translation metrics for indigenous languages of the americas. In *The Fifth Workshop on NLP for Indigenous Languages of the Americas*, page 134.

John E. Ortega, Rodolfo Zevallos, and William Chen. 2023. QUESPA submission for the IWSLT 2023 dialect and low-resource speech translation tasks. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 261–268, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Shruti Rijhwani, Enora Rice, Arturo Oncevay, Claudia Baltazar, María Cortés, Cynthia Montaño, John E Ortega, Rolando Coto-Solano, et al. 2023. Findings of the americasnlp 2023 shared task on machine translation into indigenous languages. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 206–219.

Abteen Ebrahimi, Manuel Mager, Adam Wiemerslage, Pavel Denisov, Arturo Oncevay, Danni Liu, Sai Koneru, Enes Yavuz Ugan, Zhaolin Li, Jan Niehues,

Monica Romero, Ivan G Torre, Tanel Alumäe, Jiaming Kong, Sergey Polezhaev, Yury Belousov, Wei-Rui Chen, Peter Sullivan, Ife Adebara, Bashar Talafha, Alcides Alcoba Inciarte, Muhammad Abdul-Mageed, Luis Chiruzzo, Rolando Coto-Solano, Hilaria Cruz, Sofía Flores-Solórzano, Aldo Andrés Alvarez López, Ivan Meza-Ruiz, John E. Ortega, Alexis Palmer, Rodolfo Joel Zevallos Salazar, Kristine Stenzel, Thang Vu, and Katharina Kann. 2022. Findings of the second americasnlp competition on speech-to-text translation. In *Proceedings of the NeurIPS 2022 Competitions Track*, volume 220 of *Proceedings of Machine Learning Research*, pages 217–232. PMLR.

Edward Gow-Smith, Alexandre Berard, Marcely Zanon Boito, and Ioan Calapodescu. 2023. NAVER LABS Europe's multilingual speech translation systems for the IWSLT 2023 low-resource track. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 144–158, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Xilin Jiang, Yinghao Aaron Li, Adrian Nicolas Florea, Cong Han, and Nima Mesgarani. 2024. Speech slytherin: Examining the performance and efficiency of mamba for speech separation, recognition, and synthesis. *arXiv preprint arXiv:2407.09732*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR 2015, Conference Track Proceedings*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Edward Ma. 2019. Nlp augmentation. https://github.com/makcedward/nlpaug.

Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, Alexandre Muzio, Saksham Singhal, Hany Hassan Awadalla, Xia Song, and Furu Wei. 2021. Deltalm: Encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders. *arXiv preprint arXiv:2106.13736*.

Jonathan Mbuya and Antonios Anastasopoulos. 2023. GMU systems for the IWSLT 2023 dialect and low-resource speech translation tasks. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 269–276, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Wataru Nakata, Yuki Saito, Yota Ueda, and Hiroshi Saruwatari. 2025. Sidon: Fast and robust open-source multilingual speech restoration for large-scale dataset cleansing. *arXiv preprint arXiv:2509.17052*.

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.

John E Ortega, Richard Castro Mamani, and Kyunghyun Cho. 2020. Neural machine translation with a polysynthetic low resource language. *Machine Translation*, 34(4):325–346.

John E Ortega and Krishnan Pillaipakkamnatt. 2018. Using morphemes from agglutinative languages like quechua and finnish to aid in low-resource translation. *Technologies for MT of Low Resource Languages (LoResMT 2018)*, page 1.

John E Ortega, Rodolfo Joel Zevallos, William Chen, and Idris Abdulmumin. 2025. Quespa submission for the iwslt 2025 dialectal and low-resource speech translation task. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 260–268.

Rosa YG Paccotacya-Yanque, Candy A Huanca-Anquise, Judith Escalante-Calcina, Wilber R Ramos-Lovón, and Álvaro E Cuno-Parari. 2022. A speech corpus of quechua collao for automatic dimensional emotion recognition. *Scientific Data*, 9(1):778.

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.

Yifan Peng, Jinchuan Tian, Brian Yan, Dan Berrebbi, Xuankai Chang, Xinjian Li, Jiatong Shi, Siddhant Arora, William Chen, Roshan Sharma, Wangyou Zhang, Yui Sudo, Muhammad Shakeel, Jee-Weon Jung, Soumi Maiti, and Shinji Watanabe. 2023. Reproducing whisper-style training using an open-source toolkit and publicly available data. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International*

Conference on Machine Learning, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4779–4783. IEEE.

Jiatong Shi, William Chen, Dan Berrebbi, Hsiu-Hsuan Wang, Wei-Ping Huang, En-Pei Hu, Ho-Lam Chuang, Xuankai Chang, Yuxun Tang, Shang-Wen Li, Abdelrahman Mohamed, Hung-Yi Lee, and Shinji Watanabe. 2023. Findings of the 2023 ml-superb challenge: Pre-training and evaluation over more languages and beyond. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Shengkun Tang, Liqun Ma, Haonan Li, Mingjie Sun, and Zhiqiang Shen. 2024. Bi-mamba: Towards accurate 1-bit state space models. *Preprint*, arXiv:2411.11843.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Omnilingual MT Team, Belen Alastruey, Niyati Bafna, Andrea Caciolai, Kevin Heffernan, Artyom Kozhevnikov, Christophe Ropers, Eduardo Sánchez, Charles-Eric Saint-James, Ioannis Tsiamas, Chierh Cheng, Joe Chuang, Paul-Ambroise Duquenne, Mark Duppenthaler, Nate Ekberg, Cynthia Gao, Pere Lluís Huguet Cabot, João Maria Janeiro, Jean Maillard, Gabriel Mejia Gonzalez, Holger Schwenk, Edan Toledo, Arina Turkatenko, Albert Ventayol-Boada, Rashel Moritz, Alexandre Mourachko, Surya Parimi, Mary Williamson, Shireen Yates, David Dale, and Marta R. Costa-jussà. 2026. Omnilingual mt: Machine translation for 1,600 languages. *Preprint*, arXiv:2603.16309.

Alfredo Torero. 1964. Los dialectos quechuas. *Anales Científicos de la Universidad Agraria*, 2(4):446–478.

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.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. 2018. ESPnet: End-to-end speech processing toolkit. In *Proceedings of Interspeech*, pages 2207–2211.

Rodolfo Zevallos, Nuria Bel, Guillermo Cámbara, Mireia Farrús, and Jordi Luque. 2022a. Data augmentation for low-resource quechua asr improvement. *arXiv preprint arXiv:2207.06872*.

Rodolfo Zevallos, Luis Camacho, and Nelsi Melgarejo. 2022b. Huqariq: A multilingual speech corpus of native languages of peru forspeech recognition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5029–5034.

Xiangyu Zhang, Qiquan Zhang, Hexin Liu, Tianyi Xiao, Xinyuan Qian, Beena Ahmed, Eliathamby Ambikairajah, Haizhou Li, and Julien Epps. 2024. Mamba in speech: Towards an alternative to self-attention. *arXiv preprint arXiv:2405.12609*.

A Prompt 1

Eres un traductor experto en quechua y español latinoamericano. Traduce al español la frase en quechua del campo “que” de cada registro del siguiente array JSON. Ignora cualquier otro campo durante la traducción. Devuelve únicamente un array JSON con los campos “segment” (copiado exacto del original),“spa_tc”, “que” (copiado exacto del original), “traduccion” (tu traducción en español natural y fluida), “bleu” (puntaje BLEU de 0 a 100 comparando tu traducción contra el campo “spa_tc” del registro original) y “chrf” (puntaje chrF de 0 a 100 con la misma referencia). Si hay nombres propios de personas o lugares, consérvelos tal cual. No incluyas explicaciones, comentarios ni texto fuera del JSON.

B Prompt 2

Eres un traductor experto en quechua y español latinoamericano. Traduce al español la frase en quechua del campo “que” de cada registro del siguiente array JSON. Ignora cualquier otro campo durante la traducción. Devuelve los mismos campos y agrega “traducción” (tu traducción en español natural y fluida), “bleu” (puntaje BLEU de 0 a 100 comparando tu traducción contra el campo "spa_tc"

del registro original) y "chrf" (puntaje chrF de 0 a 100 comparando tu traducción contra el campo "spa_tc" del registro original). Si hay nombres propios de personas o lugares, consérvalos tal cual. No incluyas explicaciones, comentarios ni texto fuera del JSON. Genera un json con los resultados

C Prompt 3

Eres un traductor experto en quechua y español latinoamericano. Traduce al español la frase en quechua del campo “que” de cada registro del siguiente array JSON. Ignora cualquier otro campo durante la traducción. Devuelve los campos que, spa, spa_tc y agrega “traduccion” (tu traducción en español natural y fluida). Si hay nombres propios de personas o lugares, consérvalos tal cual. No incluyas explicaciones, comentarios ni texto fuera del JSON. Genera un archivo json.

D Prompt 4

Eres un traductor experto en quechua y español latinoamericano. Traduce al español la frase en quechua del campo "que" de cada registro del archivo JSON. Solamente considera el campo que para la traduccion, no tomes spa ni spa_tc como referencia. Devuelve los campos que, spa, spa_tc y agrega “traduccion” (tu traducción en español natural y fluida). Si hay nombres propios de personas o lugares, consérvalos tal cual. No incluyas explicaciones, comentarios ni texto fuera del JSON. Genera un archivo json.

E Prompt 5

Eres un traductor experto en quechua y español latinoamericano. Traduce al español la frase en quechua que aparece en el campo que a continuación.

Reglas:

- Usa solamente el texto en quechua.
- No uses otros campos como referencia.
- Conserva nombres propios de personas o lugares tal cual.
- Devuelve solo la traducción en español natural y fluida.
- Dame un archivo json final con el campo index y traducción.
- No alucines.