# WillaykuqAI: Plataforma Inteligente para la Traducción y Difusión Multimodal de Noticias en Quechua mediante Inteligencia Artificial

Jesús G. Campos Cardenas[1], Marco A. Barrera Ninamango[1], Thomy Villanueva Quinteros[1]
[1]*Universidad Nacional de Ingeniería (UNI), Perú*
{j.campos.c@uni.pe, marco.barrera.n@uni.pe, tvillanuevaq@uni.pe}

***Resumen-****WillaykuqAI es una plataforma basada en inteligencia artificial diseñada para producir, traducir y difundir noticias en lengua quechua a partir de contenidos periodísticos originalmente escritos en español. El sistema integra modelos avanzados de traducción automática neuronal, destacando el uso de NLLB-200 para la traducción Español-Quechua, así como síntesis de voz mediante un modelo VITS para generar versiones habladas en lenguas de bajos recursos. La plataforma implementa un pipeline automatizado que abarca la recolección de noticias, normalización textual, traducción, generación de audio y publicación web multimodal. Los resultados experimentales, evaluados mediante las métricas BLEU, chrF y COMET, muestran que NLLB-200 supera consistentemente a modelos alternativos como MADLAD y Qwen, consolidándose como una solución eficaz para la inclusión informativa y el fortalecimiento de la presencia digital del quechua en el Perú.*

***Palabras clave:*** Quechua, traducción automática neuronal, NLLB-200, VITS, lenguas de bajos recursos, procesamiento de lenguaje natural, síntesis de voz.

## I. Introducción

Las lenguas indígenas enfrentan una exclusión progresiva en los entornos digitales, lo que limita el acceso de sus hablantes a información, educación y servicios públicos. Según la UNESCO, aproximadamente el 40 % de las lenguas del mundo se encuentran en peligro de desaparición, afectando de manera desproporcionada a las lenguas originarias. En el Perú, el quechua constituye una de las lenguas indígenas con mayor número de hablantes, pero su presencia en plataformas informativas digitales continúa siendo limitada.

La traducción automática neuronal (Neural Machine Translation, NMT) basada en arquitecturas Transformer ha demostrado avances significativos frente a enfoques estadísticos o basados en reglas [1]. Sin embargo, su desempeño depende de grandes volúmenes de datos paralelos, lo que representa una limitación crítica para lenguas de bajos recursos como el quechua, caracterizadas además por una morfología aglutinante y una considerable variación dialectal.

En este contexto, WillaykuqAI propone una solución integral que combina traducción automática Español-Quechua y síntesis de voz, permitiendo no solo la lectura sino también la escucha de noticias en lengua quechua. El objetivo principal del proyecto es desarrollar y evaluar una plataforma informativa multimodal que contribuya a la inclusión digital, a la reducción de la brecha informativa y al fortalecimiento de la soberanía lingüística de las comunidades quechua-hablantes del Perú.

## II. Estado del Arte

La traducción automática neuronal constituye el paradigma dominante en la traducción automática moderna, siendo la arquitectura Transformer el modelo base más utilizado [1]. Esta arquitectura reemplaza mecanismos recurrentes por self-attention, permitiendo un procesamiento paralelo eficiente y un mejor modelado de dependencias de largo alcance. No obstante, su entrenamiento efectivo requiere grandes corpus paralelos, lo que limita su aplicabilidad directa en lenguas de bajos recursos.

Para mitigar esta limitación, investigaciones recientes han demostrado que el aprendizaje por transferencia y los modelos multilingües preentrenados representan estrategias efectivas [7], [15]. En este contexto, el proyecto No Language Left Behind (NLLB) introduce un modelo multilingüe entrenado en más de 200 lenguas, optimizado explícitamente para lenguas de bajo recurso mediante minería masiva de datos, transferencia cruzada y evaluación humana responsable [2].

Benchmarks como FLORES-101 y FLORES-200 evidencian que NLLB supera a sistemas multilingües previos tanto en métricas automáticas como en correlación con evaluaciones humanas. En la región latinoamericana, las competiciones AmericasNLP muestran que el fine-tuning de modelos multilingües como NLLB, mBART y M2M-100 supera consistentemente a modelos entrenados desde cero en traducción Español-Quechua, especialmente cuando se emplean métricas basadas en caracteres como chrF++ [6], [7].

En cuanto a la evaluación automática, métricas tradicionales como BLEU presentan limitaciones en lenguas morfológicamente ricas, por lo que se recomienda el uso complementario de métricas como chrF y COMET, las cuales muestran mayor correlación con juicios humanos [3], [4], [5].

Respecto a la síntesis de voz, la generación de habla en lenguas indígenas se considera un problema de muy bajo recurso. Modelos end-to-end como VITS han demostrado producir audio más natural y estable que enfoques tradicionales como Tacotron 2, incluso con datasets relativamente pequeños, lo que los convierte en una opción adecuada para aplicaciones informativas en lenguas indígenas [9], [10].

# III. Diseño del Dataset

## A. Dataset Texto-Texto (Español-Quechua)

El dataset para la traducción automática se construye a partir de múltiples fuentes académicas y abiertas, priorizando la diversidad lingüística y la calidad semántica. Se emplean recursos paralelos ampliamente utilizados en la literatura, como OPUS y datasets abiertos, los cuales han demostrado su efectividad en traducción automática de lenguas de bajos recursos [6], [7], [8], [12].

**Cuadro 1:** DATASET TEXTO-TEXTO (TRADUCCIÓN ESPAÑOL–QUECHUA)

| Fuente | Tipo | Variante | Pares |
|---|---|---|---|
| JW300 (OPUS) | Paralelo | quz_Latn | ~150 000 |
| Bible / OPUS | Paralelo | quz_Latn / quy_Latn | ~40 000 |
| Hugging Face | Paralelo | quy_Latn | ~110 000 |
| Textos institucionales | Paralelo | quy_Latn | ~4 800 |
| **Total** | | | **~304 800** |

El corpus es sometido a limpieza, normalización ortográfica básica, alineación frase a frase y validación lingüística automática y manual. Se utiliza tokenización subword compatible con SentencePiece y etiquetas de idioma requeridas por NLLB-200.

## B. Dataset Texto-Audio (Síntesis de Voz en Quechua)

Para el módulo de síntesis de voz se utiliza un dataset anotado texto–audio en lengua quechua, adecuado para entrenamiento de modelos VITS, siguiendo prácticas reportadas en trabajos recientes de TTS end-to-end [9], [10].

**Cuadro 2:** DATASET TEXTO-AUDIO (TTS QUECHUA)

| Característica | Valor |
|---|---|
| Audios | 12 420 |
| Duración total | ~15.24 horas |
| Formato | WAV |
| Unidad | Oraciones |
| Locutores | Múltiples |
| Etiquetas emocionales | Sí |
| Variante | Quechua Collao |

# IV. Metodología y Desarrollo

El sistema WillaykuqAI implementa un pipeline automatizado compuesto por recolección de noticias, normalización textual, traducción automática, síntesis de voz y publicación web.

La traducción automática se basa en el modelo NLLB-200, el cual adopta la arquitectura Transformer encoder-decoder [1] y la extiende mediante entrenamiento multilingüe y componentes Mixture-of-Experts, permitiendo escalar a más de 200 lenguas [2], [15]. El modelo es ajustado mediante fine-tuning supervisado sobre el corpus Español-Quechua consolidado, estrategia que ha demostrado ser efectiva en escenarios de bajo recurso [6], [7], [12].

La síntesis de voz se implementa utilizando VITS (Coqui TTS), un modelo end-to-end que combina inferencia variacional y entrenamiento adversarial para generar audio natural directamente desde texto [9]. La elección de VITS se fundamenta en su desempeño superior en escenarios con datos limitados frente a modelos TTS tradicionales [10].

# V. Métricas de Evaluación

La evaluación del sistema de traducción automática se realiza mediante un conjunto de métricas automáticas complementarias, cada una diseñada para capturar aspectos específicos de la calidad de traducción en lenguas de bajos recursos.

## A. BLEU (Bilingual Evaluation Understudy)

BLEU es una métrica ampliamente utilizada que mide la similitud entre la traducción generada y una o varias traducciones de referencia mediante la coincidencia de n-gramas. Se calcula como la media geométrica ponderada de las precisiones de n-gramas (típicamente de 1 a 4), modificada por un factor de penalización por brevedad para evitar traducciones artificialmente cortas [3]. BLEU produce valores entre 0 y 1, donde 1 indica coincidencia perfecta. Sin embargo, presenta limitaciones conocidas en lenguas morfológicamente ricas como el quechua, donde variaciones flexivas válidas pueden ser penalizadas injustamente.

## B. chrF (Character n-gram F-score)

chrF es una métrica basada en n-gramas de caracteres en lugar de palabras, lo que la hace especialmente adecuada para lenguas aglutinantes y morfológicamente complejas [4]. Calcula el F-score (media armónica de precisión y recall) sobre n-gramas de caracteres, típicamente de 1 a 6 caracteres. A diferencia de BLEU, chrF no requiere tokenización perfecta y es más robusta ante variaciones morfológicas, lo que resulta fundamental para evaluar traducción al quechua. Los valores de chrF se reportan típicamente multiplicados por 100, con rangos entre 0 y 100.

## C. COMET (Crosslingual Optimized Metric for Evaluation of Translation)

COMET es una métrica neuronal basada en modelos de lenguaje preentrenados multilingües que estima la calidad de traducción mediante embeddings contextuales [5]. A diferencia de métricas léxicas como BLEU y chrF, COMET evalúa la adecuación semántica comparando representaciones vectoriales del texto fuente, la traducción generada y la referencia. Esta métrica ha demostrado una correlación significativamente superior con evaluaciones humanas, especialmente en escenarios multilingües y para lenguas de bajos recursos. Los valores de COMET

típicamente oscilan entre -1 y 1, donde valores cercanos a 1 indican alta calidad semántica.

# VI. Resultados Experimentales

La evaluación comparativa de los modelos de traducción Español-Quechua se presenta en el Cuadro 3, utilizando las tres métricas descritas anteriormente.

**Cuadro 3:** RESULTADOS DE EVALUACIÓN – TRADUCCIÓN ESPAÑOL–QUECHUA

| Modelo | BLEU | chrF | COMET |
|---|---|---|---|
| NLLB-200 | 0.66 | 27.56 | 0.6453 |
| MADLAD-400-3B | 0.18 | 17.19 | 0.3734 |
| Qwen2.5-1.5B-Instruct | 0.01 | 6.13 | 0.5242 |

Los resultados experimentales revelan que NLLB-200 supera consistentemente a los modelos comparados en todas las métricas evaluadas. El puntaje BLEU de 0.66 indica una alta coincidencia léxica con las traducciones de referencia, superando significativamente a MADLAD-400-3B (0.18) y Qwen2.5-1.5B-Instruct (0.01).

El valor de chrF de 27.56 para NLLB-200 evidencia una mejor adaptación morfológica al quechua, capturando apropiadamente las estructuras aglutinantes características de la lengua. Esta métrica demuestra ser particularmente informativa para el quechua, donde MADLAD obtiene 17.19 y Qwen apenas 6.13.

El puntaje COMET de 0.6453 para NLLB-200 confirma una mejora sustancial en la adecuación semántica de las traducciones generadas. Es notable que Qwen2.5-1.5B-Instruct, a pesar de su bajo desempeño léxico (BLEU: 0.01, chrF: 6.13), mantiene un COMET relativamente aceptable de 0.5242, lo que sugiere cierta coherencia semántica global sin correspondencia léxica precisa. MADLAD-400-3B presenta el COMET más bajo (0.3734), indicando limitaciones tanto léxicas como semánticas.

# VII. Conclusiones

Este trabajo demuestra que la integración de modelos multilingües preentrenados como NLLB-200 con síntesis de voz basada en VITS constituye una solución técnicamente viable y efectiva para la difusión informativa en lenguas indígenas de bajos recursos como el quechua.

Los resultados experimentales validan la superioridad de NLLB-200 en traducción Español–Quechua, alcanzando métricas sustancialmente superiores a alternativas multilingües recientes. El desempeño consistente en BLEU, chrF y COMET confirma que el modelo no solo mantiene fidelidad léxica, sino que también preserva adecuación morfológica y coherencia semántica, aspectos críticos en lenguas aglutinantes.

WillaykuqAI representa un avance concreto hacia la inclusión digital de comunidades quechua-hablantes, habilitando acceso multimodal (texto y audio) a información periodística en su lengua materna. La plataforma contribuye al fortalecimiento de la presencia digital del quechua en el ecosistema informativo peruano, reduciendo la brecha informativa que afecta históricamente a las comunidades indígenas.

Desde una perspectiva tecnológica, el proyecto demuestra que el fine-tuning supervisado de modelos masivos multilingües constituye una estrategia efectiva para lenguas de bajos recursos, superando enfoques alternativos que requieren mayores volúmenes de datos o infraestructura computacional. La modularidad del pipeline implementado facilita su adaptación a otras lenguas indígenas y contextos informativos similares.

Finalmente, WillaykuqAI establece una base metodológica y técnica replicable para futuras iniciativas de revitalización lingüística mediante tecnologías de inteligencia artificial, promoviendo la soberanía lingüística y el derecho a la información en lenguas originarias.

# VIII. Limitaciones y Trabajo Futuro

A pesar de los resultados prometedores, el presente trabajo enfrenta limitaciones significativas que deben ser consideradas para una interpretación adecuada de sus alcances.

La escasez de datos paralelos de alta calidad constituye la limitación fundamental. Aunque se consolidaron aproximadamente 304 800 pares de oraciones Español-Quechua, una porción considerable proviene de dominios religiosos (JW300, Bible), lo que introduce un sesgo temático y estilístico que puede afectar la generalización del modelo a contextos periodísticos contemporáneos. La validez ecológica de las traducciones generadas en escenarios informativos reales requiere verificación adicional.

La cobertura dialectal representa otra limitación crítica. El sistema se enfoca principalmente en la variante Quechua Collao (quy_Latn), con presencia menor de Quechua Cusco (quz_Latn). El quechua presenta una diversidad dialectal considerable en el Perú, y la aplicabilidad del modelo a otras variantes (Quechua Ayacucho, Quechua Ancash, entre otros) permanece sin evaluar. Esta restricción limita el alcance geográfico y demográfico del impacto potencial de la plataforma.

La ausencia de evaluaciones humanas formales constituye una limitación metodológica importante. Aunque las métricas automáticas empleadas (BLEU, chrF, COMET) presentan correlación demostrada con juicios humanos, no sustituyen evaluaciones directas de hablantes nativos respecto a fluidez, adecuación cultural y comprensibilidad. La participación de comunidades quechua-hablantes en la validación del sistema es esencial para garantizar su utilidad real.

Respecto a la síntesis de voz, el dataset de 15.24 horas, aunque adecuado para entrenamiento inicial de VITS, resulta limitado para capturar la variabilidad prosódica y emocional necesaria en contextos periodísticos diversos. La calidad perceptual del audio generado requiere evaluación formal con hablantes nativos.

Futuras extensiones del trabajo deberán abordar estas limitaciones mediante: (1) recolección y anotación de corpus paralelos en dominios periodísticos y cotidianos, involucrando hablantes nativos; (2) expansión del sistema a múltiples varian-

tes dialectales del quechua; (3) implementación de protocolos de evaluación humana rigurosos con participación comunitaria; (4) ampliación del dataset de voz con mayor diversidad de hablantes y contextos comunicativos; (5) integración de mecanismos de retroalimentación comunitaria para mejora continua del sistema.

Adicionalmente, se recomienda explorar arquitecturas más recientes de traducción automática, técnicas de aumentación de datos específicas para lenguas aglutinantes, y estrategias de adaptación al dominio que permitan reducir el sesgo temático observado en los datos de entrenamiento.

# IX. Referencias

# Referencias


[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *Proc. 31st Int. Conf. Neural Information Processing Systems (NeurIPS)*, Long Beach, CA, USA, 2017, pp. 6000-6010.

[2] M. R. Costa-jussà et al., "No Language Left Behind: Scaling Human-Centered Machine Translation," *arXiv preprint arXiv:2207.04672*, 2022.

[3] M. Post, "A Call for Clarity in Reporting BLEU Scores," in *Proc. 3rd Conf. Machine Translation (WMT)*, Brussels, Belgium, 2018, pp. 186-191.

[4] M. Popović, "chrF: Character n-gram F-score for Automatic MT Evaluation," in *Proc. 10th Workshop on Statistical Machine Translation (WMT)*, Lisbon, Portugal, 2015, pp. 392-395.

[5] R. Rei, C. Stewart, A. C. Farinha, and A. Lavie, "COMET: A Neural Framework for MT Evaluation," in *Proc. 2020 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, Online, 2020, pp. 2685-2702.

[6] A. Chaudhary et al., "AmericasNLP 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas," in *Proc. 1st Workshop on Natural Language Processing for Indigenous Languages of the Americas*, Online, 2021, pp. 202-217.

[7] J. Ortega, R. Castro Mamani, and K. Puede Choquehuanca, "Neural Machine Translation with Universal Visual Representation for Low-Resource Languages," in *Proc. 2020 Conf. North American Chapter of the Association for Computational Linguistics (NAACL)*, Online, 2020, pp. 2662-2672.

[8] J. Tiedemann and S. Thottingal, "OPUS-MT - Building Open Translation Services for the World," in *Proc. 22nd Annual Conf. European Association for Machine Translation (EAMT)*, Lisbon, Portugal, 2020, pp. 479-480.

[9] J. Kim, J. Kong, and J. Son, "Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech," in *Proc. 38th Int. Conf. Machine Learning (ICML)*, Virtual Event, 2021, pp. 5530-5540.

[10] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech," in *Proc. 9th Int. Conf. Learning Representations (ICLR)*, Virtual Event, 2021.

[11] E. M. B. Nagoudi, A. Elmadany, and M. Abdul-Mageed, "IndT5: A Text-to-Text Transformer for 10 Indigenous Languages," *arXiv preprint arXiv:2104.14096*, 2021.

[12] J. Oncevay, B. Haddow, and A. Birch, "Bridging Linguistic Typology and Multilingual Machine Translation with Multi-View Language Representations," *Machine Translation*, vol. 34, no. 2-3, pp. 139-158, 2020.

[13] J. Cross, C. Ott, and P. Koehn, "Improving Massively Multilingual Neural Machine Translation and Zero-Shot Translation," in *Proc. 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, Online, 2021, pp. 148-164.

[14] S. Mauger, "Digital Technology and Indigenous Languages: Theoretical and Practical Approaches," *Language Documentation & Conservation*, vol. 19, pp. 1-28, 2025.

[15] J. Tiedemann, "Parallel Data, Tools and Interfaces in OPUS," in *Proc. 11th Int. Conf. Language Resources and Evaluation (LREC)*, Miyazaki, Japan, 2018, pp. 2214-2218.