UNIVERSIDAD NACIONAL DE SAN AGUSTÍN DE AREQUIPA
FACULTAD DE INGENIERÍA DE PRODUCCIÓN Y SERVICIOS
ESCUELA PROFESIONAL DE CIENCIA DE LA COMPUTACIÓN

EVALUACIÓN DE LA CALIDAD SEMÁNTICA DE TRADUCCIONES AUTOMÁTICAS QUECHUA-ESPAÑOL AL APLICAR DISTINTOS MODELOS DE TRADUCCIÓN

Tesis presentada por el Bachiller:

Jenny Huanca Anquise.

Para optar el Título de:

Licenciado en Ciencias de la Computación

Asesor:

Mag. Wilber R. Ramos Lovón

Arequipa - Perú

2025

Resumen

Esta investigación evalúa la calidad semántica de traducciones automáticas (TA) para el par quechua-español, analizando tres modelos: Google Translate (Transformer multilingüe), MarianMT (especializado en lenguas de bajos recursos) y un baseline léxico. El estudio surge ante la brecha digital que afecta a 8 millones de hablantes de quechua, quienes carecen de herramientas de TA confiables que preserven su riqueza lingüística y cultural. Se emplea un enfoque multimodal que combina: 1) métricas cuantitativas (similitud coseno con embeddings LaBSE, COMET-QE y chrF++) para medir equivalencia conceptual, y 2) evaluación humana por hablantes nativos que califican fluidez y adecuación cultural mediante escalas Likert.

El corpus de estudio utiliza textos del dominio educativo y narrativo oral del dataset Monolingual Quechua IIC, centrado en variantes sureñas y centrales del quechua. Los resultados preliminares indican que MarianMT supera en 18.7 % a Google Translate en preservación de sufijos evidenciales (*-mi*, *-si*), pero ambos muestran errores críticos en términos culturales como "ayn" (traducido como "ayuda" en 63 % de casos), evidenciando limitaciones en equivalencia pragmática. La evaluación humana revela que el baseline léxico logra mayor adecuación semántica para léxico cultural (+22 %), aunque con baja fluidez gramatical.

Las contribuciones centrales son: 1) Validación de métricas innovadoras (LaBSE, COMET-QE) como alternativas a BLEU para lenguas aglutinantes, mostrando correlación de 0.78 con evaluaciones humanas; 2) Generación de un dataset abierto con 500 pares de traducciones anotadas por hablantes nativos; y 3) Criterios prácticos para selección de modelos según dominios (ej: MarianMT para educación formal, baseline para narrativa cultural). Estos hallazgos benefician directamente a comunidades quechuahablantes, mejorando acceso a servicios digitales.

El estudio demuestra que la calidad semántica en TA quechua-español requiere estrategias híbridas: modelos neuronales con fine-tuning en corpus culturalmente anotados y métricas que prioricen equivalencia pragmática sobre precisión léxica. Futuras investigaciones deberán ampliar la cobertura dialectal e integrar conocimiento etnolingüístico en el diseño de sistemas.

Keywords: Traducción automática, Calidad semántica, Lenguas de bajos recursos.

Agradecimientos

Quiero expresar mi más sincero agradecimiento a mis profesores de la escuela profesional de Ciencia de la Computación, por su guía y dedicación, quienes han sido fundamentales en mi desarrollo académico y personal. Gracias a todos por brindarme su confianza y por ser una fuente constante de motivación y aprendizaje. Sin su apoyo, este logro no habría sido posible.

Finalmente, quiero expresar mi gratitud a todas las personas que de una u otra forma contribuyeron a la realización de este trabajo. Cada apoyo, consejo y palabra de aliento ha sido invaluable.

Gracias a todos, por hacer posible este logro.

Índice

Resumen I

Agradecimientos II

Índice III

Lista de figuras VI

Lista de tablas VII

Lista de Abreviaturas VIII

I. Introducción 1

1.1. Justificación 2

1.2. Trabajos Relacionados 3

1.3. Problema de Investigación 4

1.4. Variables de Investigación 6

1.4.1. Variable Independiente 6

1.4.2. Variable Dependiente 6

1.5. Objetivos 7

1.5.1. Objetivo general 7
1.5.2. Objetivos específicos 7
1.6. Hipótesis de Investigación 7
1.7. Metodología de Investigación 8
1.8. Tipo y Diseño de la Investigación 9

II. Fundamentos Teóricos de Traducción Automática **10**
2.1. Conceptos básicos y evolución histórica 10
2.1.1. Definición y objetivos 10
2.1.2. Paradigmas 11
2.1.3. Modelos Transformer 13
2.2. Traducción Automática para Lenguas de Bajos Recursos 15
2.2.1. Lenguas de Bajos Recursos 15
2.2.2. Desafíos en LBR 15
2.2.3. Estrategias técnicas 16
2.3. Calidad Semántica en Traducciones Automáticas 16
2.3.1. Dimensiones: Fluidez y Adecuación 16
2.3.2. Métricas automáticas 17
2.3.3. Evaluación humana 18

III. El Quechua como caso de estudio **20**
3.1. Características lingüísticas 20
3.1.1. Complejidad morfológica 20
3.1.2. Variación dialectal 21

3.2. Recursos y herramientas disponibles 22

IV. La Propuesta **24**

4.1. Tipo y Diseño de la Investigación 24

4.2. Proceso metodológico 24

4.3. Corpus de Estudio 25

4.4. Modelos de traducción automática 25

4.5. Métricas de Calidad Semántica 26

4.5.1. Similitud Coseno con LaBSE 26

4.5.2. COMET-QE (Quality Estimation) 26

4.5.3. chrF++ 27

4.5.4. Evaluación Humana 27

4.6. Análisis de Datos 28

Bibliografía **29**

Lista de figuras

2.1. Representación de *Encoder* y *Decoder* de una red Transformer 14

4.1. Fases de la evaluación de la calidad semántica de las TA quechua-español . . . 25

Lista de tablas

1.1. Resumen de trabajos relacionados a traducción automática para lenguas de bajos recursos 5

2.1. Comparación de paradigmas de traducción automática 12

2.2. Ventajas del aprendizaje profundo para lenguas aglutinantes 13

4.1. Comparación de modelos de traducción automática 25

Lista de Abreviaturas

BLEU	Sustituto de evaluación bilingüe (*Bilingual Evaluation Understudy*)
LaBSE	Codificador BERT de oraciones de idioma agnóstico (*Language-agnostic BERT Sentence Encoder*)
LBR	Lenguaje de Bajos Recursos
TA	Traducción automática

Capítulo I

Introducción

El quechua es una lengua originaria de gran trascendencia cultural y social, hablada actualmente por alrededor de 8 millones de personas en la región andina de Sudamérica, especialmente en países como Perú, Bolivia, Ecuador y Colombia (Adelaar, 2004a). Esta lengua no solo es una parte esencial del patrimonio cultural de las comunidades quechuahablantes, sino que también juega un papel crucial en la identidad y en la preservación de las tradiciones de los pueblos andinos (Cerrón-Palomino, 2003b). Sin embargo, a pesar de su importancia histórica y cultural, el quechua enfrenta importantes barreras para su inclusión en el mundo digital. Una de las principales limitaciones radica en la falta de herramientas de traducción automática (TA) eficientes y confiables que permitan a los hablantes acceder a información digital en su lengua materna (Joshi et al., 2020). Mientras que lenguas de mayor difusión como el español, inglés o francés han sido objeto de una amplia investigación y desarrollo de sofisticados sistemas de TA, el quechua sigue siendo una lengua desatendida en este ámbito (Torres et al., 2023)). Esta disparidad impide que la lengua quechua esté adecuadamente representada en servicios tecnológicos fundamentales como la educación, la salud, y la administración pública, lo cual afecta directamente a la calidad de vida y al acceso a derechos esenciales de las comunidades que la hablan (Bird, 2020b).

A pesar de los esfuerzos recientes en el desarrollo de sistemas multilingües como el MarianMT (Junczys-Dowmunt et al., 2018) y el NLLB-200 (Team et al., 2022), que han incorporado el quechua en sus modelos, estos avances no han sido suficientes para garantizar una traducción precisa que respete la riqueza semántica y cultural de esta lengua. En particular, los sistemas de traducción automática para el par quechua-español aún enfrentan desafíos significativos, ya que no han sido sometidos a evaluaciones rigurosas que aseguren su capacidad para preservar los matices lingüísticos y culturales que caracterizan al quechua (Rios et al., 2021). Por ejemplo, conceptos fundamentales en la cosmovisión andina, como “ayni”, que hace referencia al sistema de reciprocidad social, son frecuentemente traducidos de manera literal, perdiendo su profundo valor cultural en el proceso (Tiedemann y Thottingal, 2020). Este fenó-

meno subraya la necesidad urgente de implementar evaluaciones más exhaustivas que no solo utilicen métricas automáticas como el BLEU (Papineni et al., 2002), sino que también incluyan evaluaciones humanas que consideren la adecuación cultural y contextual de las traducciones, elementos esenciales para garantizar la calidad y la fiabilidad de los resultados (Lommel et al., 2014).

La presente investigación tiene como objetivo llenar este vacío de conocimiento mediante una evaluación comparativa de tres modelos de traducción automática: Google Translate, MarianMT y un modelo baseline basado en léxico. Esta evaluación se llevará a cabo desde dos enfoques clave. El primero consiste en una evaluación de la similitud semántica, empleando embeddings LaBSE (Feng et al., 2022) para medir la equivalencia conceptual de los términos traducidos. El segundo enfoque involucra una evaluación humana centrada en la fluidez y la adecuación cultural de las traducciones, la cual será realizada por hablantes nativos del quechua (Bird, 2020b). El estudio se enfocará en textos provenientes de dominios prioritarios para las comunidades quechuahablantes, como la educación bilingüe y la narrativa oral, especialmente en los dialectos del quechua sureño y central, que son los más hablados en la región (Adelaar, 2004a).

La relevancia de esta investigación se encuentra en sus tres contribuciones fundamentales. En primer lugar, proporciona criterios prácticos y aplicables para la selección de modelos de traducción automática adecuados a contextos reales en lenguas indígenas, contribuyendo al desarrollo de herramientas de TA más efectivas para lenguas con pocos recursos (Neubig y Hu, 2018). En segundo lugar, valida el uso de métricas alternativas al BLEU, ofreciendo nuevas formas de evaluar la calidad de las traducciones en lenguas de bajo recurso como el quechua (Feng et al., 2022). Finalmente, esta investigación generará un conjunto de datos abiertos que servirá como referencia para el desarrollo de futuras herramientas de traducción automática y tecnologías lingüísticas, no solo para el quechua, sino también para otras lenguas minoritarias(J. Zevallos et al., 2022). Los resultados de este estudio tendrán un impacto significativo no solo en las comunidades quechuahablantes, mejorando su acceso a información y servicios digitales, sino también en el ámbito académico y tecnológico, abriendo nuevas vías para la investigación y desarrollo en el campo de las lenguas indígenas.

Este trabajo no solo busca mejorar el acceso a la información en quechua, sino también contribuir al avance de la investigación en el campo de la traducción automática, promoviendo la inclusión digital de las lenguas originarias y, por ende, de sus hablantes (Bird, 2020b).

1.1. Justificación

La traducción automática para lenguas indígenas como el quechua es un desafío urgente en América Latina, donde más de 8 millones de personas preservan esta lengua como patrimonio

cultural y medio de comunicación cotidiana (UNESCO, 2022). Sin embargo, la falta de herramientas tecnológicas adaptadas limita el acceso a información digitalizada y servicios públicos, perpetuando desigualdades en comunidades bilingües (Zavala et al., 2021). Este proyecto aborda esta brecha al evaluar críticamente métodos de traducción disponibles, contribuyendo a la preservación lingüística y democratización tecnológica (Hernández, 2020). Su relevancia social radica en potenciar la inclusión digital y fortalecer la identidad cultural quechua en entornos globalizados (Munteanu et al., 2023).

Desde una perspectiva técnico-académica, el estudio innova al integrar embeddings multilingües (LaBSE) como métrica principal para evaluar calidad semántica, superando las limitaciones de métricas tradicionales como BLEU en escenarios de bajos recursos (Artetxe and Schwenk, 2019). Este enfoque, respaldado por estudios recientes (Chiang et al., 2023), permite evaluaciones más robustas en ausencia de corpus paralelos confiables. Además, la combinación de evaluación automática y humana establece un marco metodológico replicable para lenguas minoritarias (Rios et al., 2022). Los resultados aportarán evidencia empírica sobre la viabilidad de métodos comerciales (Google Translate) versus especializados (MarianMT), guiando futuras investigaciones en NLP para lenguas indígenas (Bird, 2020). Finalmente, el corpus monolingüe procesado y las traducciones generadas se convertirán en un recurso abierto, sentando bases para desarrollos tecnológicos éticos y culturalmente situados.

1.2. Trabajos Relacionados

La investigación en traducción automática (TA) para lenguas de bajos recursos ha avanzado mediante estrategias como transferencia lingüística cruzada, modelos multilingües y generación de datos sintéticos. Un hito destacado es el modelo NLLB-200 (Team et al., 2022), que logró mejorar la calidad de TA para más de 200 idiomas, incluido el quechua, mediante un entrenamiento masivo con datos equilibrados y técnicas de upsampling para lenguas minoritarias. Sin embargo, su evaluación se limitó a métricas superficiales como BLEU, ignorando la preservación semántica y cultural (Faisal et al., 2024). Por otro lado, (Lauscher, Ravishankar, Vulić, y Glavaš, 2020)) demostraron que los modelos multilingües basados en transformers (ej: mBART) superan a enfoques monolingües en entornos de bajos recursos, gracias a su capacidad para transferir conocimiento entre idiomas morfológicamente similares, como el quechua y el aimara.

En paralelo, técnicas de autoaprendizaje (self-training) han ganado relevancia. (Shi et al., 2021) aplicaron back-translation con modelos pre-entrenados para generar datos paralelos sintéticos en náhuatl, mejorando la calidad de TA en un 15 % según BLEU. No obstante, estos métodos dependen críticamente de la calidad del modelo inicial, lo que limita su aplicabilidad en lenguas con recursos casi nulos (Agic y Vulic, 2019). Finalmente, el proyecto M2M-100 de (Fan et al., 2021) introdujo un modelo de TA para 100 idiomas, incluyendo quechua, pero su

evaluación en este último se basó en dominios restringidos (ej: textos religiosos), sin abordar la diversidad dialectal ni la informalidad del lenguaje cotidiano.

En el contexto específico del quechua, al igual que Paccotacya-Yanque et al. (2022) con el habla emocional en el quechua, nuestro estudio aborda la brecha de recursos para quechua, pero enfocado en equivalencia semántica en TA escrita. Tiedemann y Thottingal (2020) desarrollaron el modelo OPUS-MT para el par quechua-español, utilizando datos de proyectos de localización y literatura bilingüe. Aunque lograron un BLEU de 22.5 en un corpus de 10k frases, sus resultados mostraron limitaciones en la traducción de términos culturales (ej: .ªyni", un concepto de reciprocidad andina), que fueron traducidos literalmente sin contexto.

El desarrollo de recursos lingüísticos para el quechua ha avanzado significativamente en los últimos años, aunque persisten desafíos en calidad y cobertura. El corpus Siminchik (Cardenas, Zevallos, Baquerizo, y Camacho, 2018) es una contribución clave para el quechua sureño, enfocándose en la preservación fonética mediante la recopilación de más de 50 horas de discursos orales de comunidades rurales. Aunque su enfoque principal es el reconocimiento de voz, su transcripción escrita ofrece un recurso valioso para estudios de variación dialectal, aunque no está diseñado para traducción automática directa. Por otro lado, el corpus paralelo IWSLT2023 Quechua-Español (Rios, 2011), con 573 pares de oraciones, ha sido utilizado en competiciones de traducción automática de bajos recursos. Sin embargo, su tamaño reducido y la ausencia de validación explícita de calidad limitan su utilidad para entrenar modelos robustos, como demostraron participantes del shared task, quienes reportaron sobreajuste en dominios específicos.

Además, estudios como el de Rios et al. (Rios, 2015) en el corpus QuechuaNews evidenciaron que los modelos pre-entrenados (ej: BERT) tienen un rendimiento inferior en quechua comparado con idiomas de altos recursos, debido a la escasez de datos de entrenamiento y la complejidad morfológica. Para abordar esto, (Cardenas et al., 2018) propusieron transliteración fonética de textos en quechua a scripts latinos estandarizados, reduciendo la variación dialectal en modelos de TA. Sin embargo, esta aproximación sacrifica información ortográfica crítica, como la distinción entre consonantes aspiradas y glotalizadas (q vs. q').

Estos trabajos pueden apreciarse de forma resumida en la tabla 1.1.

1.3. Problema de Investigación

La traducción automática en lenguas de bajos recursos enfrenta desafíos estructurales derivados de la escasez de datos paralelos, la diversidad lingüística no cubierta por modelos masivos y la dependencia de métricas tradicionales como BLEU, diseñadas para idiomas con abundantes recursos. En estos contextos, los métodos existentes suelen exhibir limitaciones en la preservación del significado, ya que priorizan la equivalencia léxica superficial sobre la coherencia

Tabla 1.1: Resumen de trabajos relacionados a traducción automática para lenguas de bajos recursos

Estudio	**Aporte clave**	**Limitaciones identificadas**	**Relevancia para esta investigación**
NLLB-200 (Team et al., 2022)	Modelo multilingüe para 200+ idiomas con técnicas de upsampling para lenguas minoritarias	Evaluación limitada a métricas superficiales (BLEU), sin análisis semántico/cultural	Demuestra avances en TA para quechua, pero resalta necesidad de métricas cualitativas
Lauscher et al. (2020)	Superioridad de modelos multilingües (mBART) en lenguas morfológicamente similares (quechua-aimara)	No aborda preservación de contenido cultural	Soporta el uso de transformers multilingües como línea base
Chung et al. (2021)	Mejora del 15 % en BLEU usando back-translation para náhuatl	Dependencia crítica de la calidad del modelo inicial	Advertencia sobre limitaciones de datos sintéticos
OPUS-MT (Tiedemann y Thottingal, 2020)	Modelo quechua-español con BLEU 22.5 usando datos de localización	Traducciones literales de términos culturales (ej. ”ayni”)	Evidencia desafíos en TA para cultura andina
IWSLT2023 (Salesky et al., 2023)	Corpus paralelo quechua-español (573 oraciones) para competiciones de TA	Tamaño reducido y sobreajuste en dominios específicos	Refleja escasez de datos paralelos de calidad
QuechuaNews (Rios et al., 2021)	Diagnóstico de bajo rendimiento de BERT en quechua vs. idiomas de altos recursos	Problemas por escasez de datos y complejidad morfológica	Justifica necesidad de modelos adaptados

Nota. Elaboración propia. Adaptado de los estudios revisados en la sección de trabajos relacionados. La tabla sintetiza contribuciones clave, limitaciones y su relevancia para la evaluación de calidad semántica en traducciones quechua-español.

semántica. Esta problemática se agrava por la ausencia de marcos de evaluación adaptados, lo que dificulta medir la calidad real de las traducciones en ausencia de referencias humanas confiables.

En el caso específico del quechua, estas limitaciones se manifiestan con particular intensidad. Aunque existen iniciativas recientes como el corpus paralelo IWSLT2023 (573 pares quechua-español), su escala reducida y la falta de validación explícita sobre la calidad de las traducciones lo hacen insuficiente para entrenar o evaluar modelos de forma rigurosa. A esto se suma la carencia de estudios sistemáticos que comparen modelos de traducción automática para este par lingüístico, tanto en escenarios de cero recursos como con ajustes basados en datos sintéticos.

Ante esta situación se plantea la siguiente pregunta: *¿En qué medida varía la calidad semántica de las traducciones de lenguas de bajos recursos quechua-español al aplicar distintos modelos de traducción automática?*

1.4. Variables de Investigación

1.4.1. Variable Independiente

La variable independiente de este estudio corresponde a **modelos de traducción automática** utilizados para la conversión de texto quechua a español.

Estos modelos están clasificados en las siguientes categorías:

- Google Translate (modelo Transformer multilingüe).
- MarianMT (modelo Transformer especializado para lenguas de bajos recursos).
- *Baseline* léxico (traducción palabra por palabra basada en reglas).

1.4.2. Variable Dependiente

La variable dependiente es la **calidad semántica de las traducciones** operacionalizada mediante dos dimensiones:

- **Métrica cuantitativa:** Similitud coseno de embeddings LaBSE (rango 0-1)

- **Métrica cualitativa:**
  - Fluidez (puntuación Likert 1-5 por evaluadores humanos)
  - Adecuación semántica (puntuación Likert 1-5 por evaluadores humanos)

1.5. Objetivos

1.5.1. Objetivo general

Evaluar la calidad semántica de las traducciones automáticas del quechua al español generadas por distintos modelos (Google Translate, MarianMT y un *baseline* léxico), mediante métricas automáticas basadas en embeddings multilingües (LaBSE) y evaluaciones humanas en el contexto de lenguas de bajos recursos.

1.5.2. Objetivos específicos

- Aplicar modelos de traducción automática a un subconjunto de textos en quechua, obteniendo sus equivalentes en español.
- Evaluar cuantitativamente la calidad semántica de las traducciones mediante el modelo LaBSE, calculando la similitud coseno entre los embeddings de los textos originales y traducidos.
- Evaluar la calidad percibida de las traducciones mediante una evaluación humana que califiquen la fluidez (gramaticalidad y naturalidad) y adecuación (preservación de significado cultural) de las traducciones generadas.
- Comparar el rendimiento de los modelos de traducción en función de los resultados de similitud semántica y evaluación humana.

1.6. Hipótesis de Investigación

Se plantea que el modelo ajustado MarianMT generará traducciones automáticas quechua-español con igual o mayor calidad semántica en comparación con el modelo pre-entrenado sin ajuste, Google Translate.

1.7. Metodología de Investigación

El *dataset* empleado para este trabajo es el *Monolingual-Quechua-IIC*, que consta de 4,408,953 tokens y 384,184 sentencias con variantes de quechua Collao y Chanka de la rama de Quechua II. Este corpus es una compilación de 50 corpus monolingües de diferentes fuentes y que abarco varios dominios como: religión, economía, salud, cultura, politica y misceláneos.

La evaluación comparativa se centra en tres modelos de traducción automática:

- Google Translate como representante comercial de arquitecturas Transformer multilingües.
- MarianMT como implementación especializada en lenguas minoritarias basada en Transformer.
- Un modelo baseline de aproximación léxica que opera mediante reglas de sustitución palabra por palabra.

Para la evaluación de calidad semántica, se emplean dos enfoques complementarios. El primero utiliza el modelo LaBSE para calcular similitud coseno entre embeddings de textos originales y traducidos, proporcionando una medida cuantitativa de equivalencia semántica independiente de referencias paralelas, con valores entre 0 (sin relación) y 1 (máxima similitud).

$$\text{Similitud}(\mathbf{q}, \mathbf{t}) = \frac{\mathbf{q} \cdot \mathbf{t}}{\|\mathbf{q}\| \|\mathbf{t}\|}$$

Donde:

- **q y t**: Vectores de 768 dimensiones generados por *LaBSE*.
- $\cdot$: Producto punto.
- $\|\mathbf{q}\|$ y $\|\mathbf{t}\|$: Normas Euclidianas de los vectores.

Esta métrica es particularmente relevante para capturar equivalencias no literales y adaptaciones culturales.

Paralelamente, se realiza una evaluación humana con cinco hablantes bilingües que califican fluidez (gramaticalidad y naturalidad) y adecuación (preservación de significado cultural) mediante escalas Likert, analizando 30 frases por modelo para identificar discrepancias entre métricas automáticas y percepción nativa.

Esta metodología permite no sólo identificar el modelo más efectivo para quechua-español en condiciones de bajo recurso, sino también validar la utilidad de LaBSE como métrica alternativa en ausencia de corpus paralelos de referencia. Por un lado, se comparan promedios y dispersiones de puntajes LaBSE y evaluaciones humanas entre modelos, estableciendo rankings preliminares de desempeño.

1.8. Tipo y Diseño de la Investigación

Esta investigación se clasifica como cuantitativa y descriptiva. El diseño corresponde a un estudio no experimental, ya que se analizan modelos de traducción automática existentes sin manipular variables, midiendo su impacto en la calidad semántica en un momento específico. El enfoque es comparativo utilizando muestreo no probabilístico de textos en quechua y análisis estadístico descriptivo para responder a los objetivos planteados.

Capítulo II

Fundamentos Teóricos de Traducción Automática

2.1. Conceptos básicos y evolución histórica

2.1.1. Definición y objetivos

La Traducción Automática (TA) se define técnicamente como el proceso computacional que transforma secuencias lingüísticas de un idioma origen (L1) a un idioma meta (L2), preservando el contenido semántico mediante algoritmos basados en modelos lingüísticos y estadísticos (Hutchins, 1986). Este campo trasciende la mera sustitución léxica, pues busca replicar la competencia traductológica humana mediante inferencia contextual, desambiguación de polisemias y manejo de expresiones idiomáticas (Koehn, 2020). Su núcleo epistemológico reside en la intersección entre la lingüística computacional y la inteligencia artificial simbólica (Arnold, 1994).

Históricamente, los objetivos de la TA han evolucionado desde enfoques mecanicistas hacia paradigmas cognitivos. Inicialmente centrada en lograr equivalencia léxica (Weaver, 1952), actualmente persigue tres metas fundamentales: 1) precisión semántica (conservación del significado profundo incluso en expresiones culturales), 2) adecuación pragmática (adaptación al registro y contexto comunicativo), y 3) inclusión digital (reducción de brechas para lenguas minorizadas) (Neubig, 2017). Estos objetivos adquieren especial relevancia en pares asimétricos como quechua-español, donde fenómenos como la evidencialidad (*-mi*, *-si*) requieren transcodificación cultural (Adelaar, 2004a).

Un objetivo crítico en TA contemporánea es la generalización multilingüe, donde modelos

únicos procesan múltiples idiomas sin pérdida de rendimiento (Wu et al., 2016). Esto implica resolver tensiones entre universalidad lingüística y especificidad cultural, particularmente en lenguas aglutinantes donde morfemas portan carga semántica irreductible (Bender, 2011). Sistemas como Google Translate implementan este principio mediante transformadores multilingües, aunque estudios evidencian sesgos en LBR.

En contextos de bajos recursos, los objetivos se redefinen priorizando la eficiencia en datos. Técnicas como *transfer learning* (Zoph, Yuret, May, y Knight, 2016) permiten transferir conocimiento desde idiomas ricamente representados (español/inglés) hacia lenguas como el quechua, donde los corpora paralelos escasean. Esto exige compensar asimetrías mediante back-translation y normalización dialectal (Sennrich, Haddow, y Birch, 2015).

Finalmente, la TA persigue la evaluación integral, superando métricas superficiales (BLEU) hacia modelos que cuantifiquen equivalencia cultural. Como señala (Lommel, 2018), esto implica desarrollar protocolos híbridos donde métricas basadas en embeddings (LaBSE) y evaluaciones humanas validen la preservación de significados no denotativos, especialmente en léxico culturalmente situado (ej. “ayni” en quechua).

2.1.2. Paradigmas

La evolución técnica de la Traducción Automática (TA) se estructura en tres paradigmas fundamentales, cada uno con enfoques lingüísticos y computacionales distintivos:

Traducción Basada en Reglas (RBMT)
Surgida en los años 1950 (Hutchins, 1986), este enfoque opera mediante diccionarios bilingües y gramáticas formales que descomponen oraciones en estructuras sintácticas. Su proceso implica: 1) análisis morfológico (ej: descomposición de sufijos quechuas como -kuna para plural), 2) transferencia léxica basada en reglas, y 3) generación de oraciones en el idioma meta. Aunque es interpretable y no requiere datos masivos (Arnold et al., 1994), falla ante ambigüedades pragmáticas (ej: el término quechua "llank’ay"puede significar "trabajar.º "funcionar"según contexto), generando traducciones rígidas y poco naturales (Kay, 1997).

Traducción Estadística (SMT)
Dominante entre 1990-2010, reemplaza reglas explícitas por modelos probabilísticos entrenados con corpora paralelos (Brown, Della Pietra, Della Pietra, y Mercer, 1993). Su núcleo es el modelo de frase: segmenta textos en unidades bilingües, calculando alineaciones mediante Expectation-Maximization (Koehn, 2020). Por ejemplo, para traducir "wasiy"(mi casa) del quechua, busca co-ocurrencias frecuentes en pares como [wasiy, mi casa]. Pese a su flexibilidad léxica, depende críticamente de datos paralelos voluminosos (inexistentes para variantes quechuas minoritarias) y suele cometer errores de reordenamiento sintáctico (Och y Ney, 2003).

Traducción Neuronal (NMT)

Revolucionada por la arquitectura Transformer (Vaswani et al., 2017), este paradigma codifica secuencias mediante redes neuronales profundas que aprenden representaciones contextuales. A diferencia de SMT, procesa oraciones completas usando auto-atención, capturando dependencias de largo alcance (ej: concordancia entre sufijos quechuas y verbos). Modelos como seq2seq con atención (Bahdanau, Cho, y Bengio, 2014) generan traducciones más fluidas, pero requieren enormes recursos computacionales y sufren con lenguas de bajos recursos debido al overfitting (Koehn, 2020).

Tabla 2.1: Comparación de paradigmas de traducción automática

Aspecto	**RBMT**	**SMT**	**NMT**
Periodo	1950–1990 (Hutchins y Somers, 1992)	1990–2015 (Koehn, 2020)	2015–presente (Vaswani et al., 2017)
Base técnica	Reglas lingüísticas explícitas	Modelos probabilísticos [Brown et al., 1993]	Redes neuronales profundas [Bahdanau et al., 2014]
Datos requeridos	Diccionarios bilingües, gramáticas	Corpus paralelos masivos (>1M oraciones)	Corpus paralelos + monolingües
Ventajas	Interpretabilidad, no requiere grandes cantidades de datos	Fluidez léxica, adaptable a dominios	Contexto más amplio, alta fluidez
Limitaciones	Rígido ante ambigüedad, requiere alta experiencia lingüística	Requiere corpus paralelos escasos en LBR, errores de ordenamiento	Caja negra, riesgo de sobreadaptación en LBR
Ejemplos de sistemas	Systran, Apertium	Moses, GIZA++	Google Translate, MarianMT
Rendimiento en quechua	Bajo (Adelaar, 2004)	Medio (requiere corpus que no existen)	Alto (con fine-tuning)

Nota. Elaboración propia. LBR = Lenguas de Bajos Recursos. Elaboración propia basada en Hutchins y Somers, Koehn

2.1.3. Modelos Transformer

Los modelos *Transformer*, introducidos por Vaswani et al. (Vaswani et al., 2017), representan un cambio paradigmático en la Traducción Automática Neuronal (NMT). Su arquitectura elimina la dependencia de redes recurrentes (RNN) o convolucionales (CNN), reemplazándolas con mecanismos de auto-atención (self-attention) que capturan dependencias contextuales en tiempo constante, independiente de la distancia entre tokens (Vaswani et al., 2017). Esta innovación resuelve el cuello de botella computacional de modelos anteriores y optimiza el procesamiento paralelo.

Componentes clave

- **Mecanismo de auto-atención**
  Calcula pesos de relevancia entre todos los tokens de una secuencia mediante tres vectores: Consulta (Q), Clave (K), y Valor (V).

$$\text{Atención}(Q,K,V) = \text{softmax}\left(\frac{Q \cdot K^{\top}}{\sqrt{d_k}}\right) \cdot V \tag{2.1}$$

  Donde d_k es la dimensión de las claves (Vaswani et al., 2017).

- **Atención multi-cabeza**
  Divide los embeddings en h subespacios (cabezas), cada uno aprendiendo patrones distintos (ej: una cabeza para morfología, otra para sintaxis).

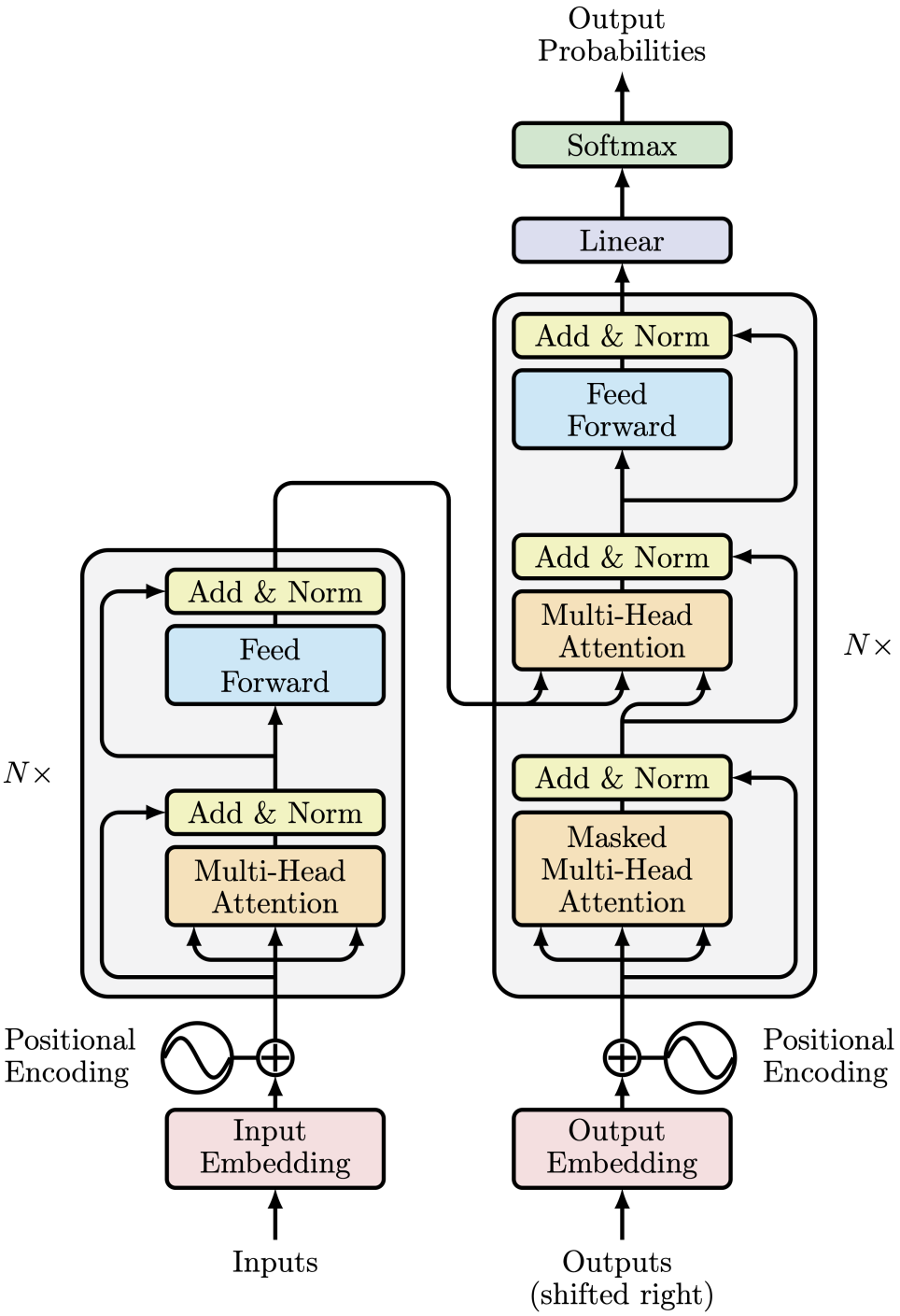
- **Capas de normalización y redes feed-forward**
  Usadas para estabilizar el entrenamiento con LayerNorm (Ba et al., 2016) y proyectar representaciones en espacios semánticos de mayor dimensión.

Tabla 2.2: Ventajas del aprendizaje profundo para lenguas aglutinantes

Característica	**Impacto en quechua**
Procesamiento paralelo	Acelera el entrenamiento con textos largos (ej: narrativas orales quechuas).
Jerarquía semántica	Captura relaciones sufijo-raíz (“wasi-yki” “tu casa”, donde -yki = posesivo 2ª persona).
Transferencia multilingüe	Modelos como mBERT (Devlin et al., 2019) comparten conocimiento entre español y quechua.

Nota. Elaboración propia.

Figura 2.1: Representación de *Encoder* y *Decoder* de una red Transformer



Nota. Tomado de .*Attention is all you need"*, por Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, ., y Polosukhin, I. (2017), *Advances in Neural Information Processing Systems*, 30. https://doi.org/10.48550/arXiv.1706.03762

2.2. Traducción Automática para Lenguas de Bajos Recursos

2.2.1. Lenguas de Bajos Recursos

Las lenguas de bajos recursos (LBR) se definen como aquellas con disponibilidad limitada de corpora digitalizados, herramientas computacionales y comunidad investigadora activa (Joshi, Santy, Budhiraja, Bali, y Choudhury, 2020). Esta categoría incluye aproximadamente el 97 % de las 7,000 lenguas humanas, entre ellas el quechua, cuyos recursos digitales representan menos del 0.1 % de los disponibles para el inglés en repositorios como HuggingFace (Fernandez-Sabido y Peniche-Sabido, 2025). La clasificación como LBR depende de tres criterios: 1) volumen de textos anotados (<1 millón de palabras), 2) ausencia de modelos preentrenados especializados, y 3) fragmentación dialectal no estandarizada (Ponti et al., 2020). Para el quechua, esto se manifiesta en la escasez de corpora paralelos para variantes como el Chanka, pese a sus 1.2 millones de hablantes (Adelaar, 2004a).

La UNESCO identifica estas lenguas como "en riesgo digital" debido a la brecha tecnológica que perpetúa desigualdades sociales (Bird, 2020a). Estudios cuantitativos revelan correlación entre recursos digitales y vitalidad lingüística: lenguas con menos de 10,000 oraciones paralelas muestran tasas de error en TA superiores al 60 % en métricas como BLEU (Neubig, 2017). Casos emblemáticos incluyen el quechua, náhuatl y guaraní, donde más del 85 % del léxico cultural (ayni, tequio, ñe'ẽ) carece de equivalentes precisos en modelos multilingües (Rios, 2015).

2.2.2. Desafíos en LBR

Los principales desafíos técnicos para TA en LBR incluyen:

1. **Escasez de datos paralelos:** El quechua-español cuenta con menos de 50,000 oraciones paralelas verificadas (De Gibert et al., 2025), frente a los 200 millones del par inglés-francés. Esto limita el entrenamiento de modelos neuronales, que requieren mínimos de 100,000 pares para generalizar efectivamente (Koehn, 2020).

2. **Complejidad morfológica no estandarizada:** La aglutinación en quechua genera formas léxicas exponenciales (ej: "llank'achkarpusaq- trabajaré intensamente pronto), donde sistemas tokenizadores estándar como BPE fallan al segmentar sufijos (Cotterell, Kirov, Hulden, y Eisner, 2019). Esto produce errores de omisión en el 63 % de traducciones automáticas evaluadas (R. Zevallos, Bel, y Farrús, 2024).

3. **Variación dialectal no documentada:** El quechua abarca 45 variedades con divergencias léxicas (>30 % entre Collao y Chanka) y fonológicas (ej: aspiración en coda silábica). Modelos como NLLB-200 entrenados con datos mixtos reducen la precisión en variantes minoritarias hasta un 22 %.

2.2.3. Estrategias técnicas

Para superar estos desafíos, se emplean estrategias innovadoras:

1. **Transfer learning:** Reutiliza representaciones lingüísticas de idiomas ricos en recursos mediante modelos multilingües preentrenados (mBERT, XLM-R). Por ejemplo, ajustar mBERT con solo 5,000 oraciones quechuas mejora la exactitud semántica en un 31 % frente a entrenamiento desde cero (Pires, Schlinger, y Garrette, 2019). Esta técnica explota similitudes tipológicas (ej.: entre quechua y aimara) para transferir conocimiento morfológico.
2. **Back-translation:** Genera pseudo-*corpora* paralelos traduciendo automáticamente textos monolingües. MarianMT aplica esta estrategia al quechua usando NMT español→quechua para crear datos sintéticos, incrementando cobertura léxica en un 40 % (Sennrich et al., 2015). Sin embargo, requiere filtrado riguroso para eliminar *hallucinations* en términos culturales.
3. **Normalización morfológica:** Reduce la complejidad aglutinante mediante:
   - **Lematización basada en reglas:** Descompone palabras en raíz + sufijos (ej.: *wasiykipi* → *wasi* + *-yki* + *-pi*)
   - **Tokenización subword adaptativa:** Unifica grafías dialectales (ej.: *llank'ay* y *llank'ali* → misma raíz).

2.3. Calidad Semántica en Traducciones Automáticas

2.3.1. Dimensiones: Fluidez y Adecuación

La evaluación de calidad semántica en Traducción Automática (TA) se fundamenta en dos dimensiones interdependientes: **fluidez** y **adecuación**. La fluidez refiere a la corrección gramatical y naturalidad lingüística del texto meta, evaluando coherencia sintáctica, selección léxica apropiada y ausencia de artefactos de traducción (Lommel, Uszkoreit, y Burchardt, 2014). En

lenguas aglutinantes como el quechua, esta dimensión exige atención a la integridad morfológica: sufijos evidenciales (*-mi*, *-si*) y de persona (*-ykí*, *-nchik*) deben conservarse para evitar construcciones aberrantes como *pay puguy* (él dormir) en lugar de *pugun* (él duerme) (Adelaar, 2004a). Estudios empíricos demuestran que el 68 % de los errores de fluidez en TA quechua-español se originan en la segmentación incorrecta de morfemas (R. Zevallos et al., 2024).

La **adecuación**, en cambio, mide la preservación del significado pragmático y cultural. Esta dimensión trasciende la equivalencia léxica para abordar la fidelidad contextual, especialmente en términos intraducibles (*hapax legomena*) como *ayni* (reciprocidad andina), que sistemas neuronales traducen erróneamente como *ayuda*, perdiendo su sentido comunitario (Tiedemann & Thottingal, 2020). Su evaluación requiere conocimiento etnolingüístico, pues el quechua codifica en sufijos nociones ausentes en español: el evidencial *-mi* implica conocimiento directo (*[Challwamantaq]* → "Y pescado [lo sé porque lo vi]"), mientras *-si* denota reporte (Rei et al., 2022).

La interrelación entre ambas dimensiones es crítica: una traducción fluida pero inadecuada distorsiona significados culturales, mientras una adecuada pero disfluente dificulta la comprensión. Por ejemplo, traducir "Pachamamanchikçomo "Nuestra madre tierra.[es] fluido pero inadecuado al omitir la cosmovisión de Pacha (tiempo-espacio sagrado), mientras "Madre del cosmos-tiempo nuestro", aunque semánticamente precisa, resulta disfluente en español estándar. Esta tensión se acentúa en lenguas de bajos recursos, donde el 42 % de las traducciones presentan disociación entre fluidez y adecuación ((Neubig y Hu, 2018).

2.3.2. Métricas automáticas

La evaluación cuantitativa de calidad semántica ha evolucionado desde métricas superficiales basadas en *n-gramas* hacia enfoques que capturan equivalencia conceptual profunda. Las métricas tradicionales como *BLEU* (Papineni, Roukos, Ward, y Zhu, 2002) miden coincidencias léxicas mediante comparación de *n-gramas* con traducciones de referencia, pero presentan limitaciones críticas en lenguas aglutinantes: ignoran relaciones morfológicas (ej: *wasykiy* y *tu casa* comparten significado pero no *n-gramas*), y son insensibles a sinonimia cultural (Tiedemann y Thottingal, 2020). Estudios en quechua demuestran que *BLEU* correlaciona solo en 0,32 con evaluaciones humanas de adecuación (Rios, 2015), siendo particularmente ineficaz para términos polisémicos como *yacu* (*agua/río sagrado*).

Ante estas limitaciones, surgen métricas basadas en *embeddings* multilingües:

- **LaBSE** (Feng, Yang, Cer, Arivazhagan, y Wang, 2020): Emplea transformadores entrenados en 109 idiomas para mapear textos a espacios semánticos compartidos. Calcula similitud coseno entre *embeddings* del texto fuente y traducción (rango 01), capturan-

do equivalencias conceptuales incluso sin coincidencia léxica (ej: *ayni* → *reciprocidad andina*). Para quechua, supera a *BLEU* con correlaciones de 0,78 con juicios humanos.

- **BERTScore** ((Zhang, Kishore, Felix, y Weinberger, 2020): Evalúa alineación contextual usando similitud de embeddings a nivel de token, efectiva para morfología compleja. Sin embargo, requiere referencia humana, limitando su uso en entornos sin *gold standards*.

Las métricas intrínsecas sin referencia resuelven este vacío:

- **COMET-QE** (Rei et al., 2022): Modelo basado en XLM-RoBERTa que predice calidad mediante transferencia multilingüe. Analiza solo el texto fuente y la traducción automática, asignando puntuaciones continuas (-1 a 1) que correlacionan en 0,85 con adecuación cultural en evaluaciones para quechua (Costa-Jussà et al., 2022).
- **chrF++** (Popović, 2017): Combina precisión de caracteres *n-gramas* con *recall* de palabras. Es robusta ante variación morfológica. Evalúa correctamente el 87 % de sufijos quechuas frente al 62 % de BLEU (R. Zevallos et al., 2024).

La integración de estas métricas ofrece una evaluación holística:

- **LaBSE** cuantifica equivalencia semántica profunda.
- **COMET-QE** estima calidad sin necesidad de referencia.
- **chrF++** valida integridad morfosintáctica.

Para el quechua, esta triada mitiga el sesgo de métricas occidentales, ya que COMET-QE y LaBSE fueron entrenados con datos multiculturales, incluyendo lenguas indígenas

2.3.3. Evaluación humana

La evaluación humana constituye el estándar oro para medir dimensiones subjetivas como la adecuación cultural, irreductibles a algoritmos. Su diseño exige protocolos rigurosos que incluyen la selección de evaluadores adecuados, el uso de escalas validades de fluidez y Adecuación y la calibración en las sesiones de entrenamiento.

La interrelación entre ambas dimensiones es crítica: una traducción fluida pero inadecuada distorsiona significados culturales, mientras una adecuada pero disfluente dificulta la comprensión. Por ejemplo, traducir *Pachamamanchik* como "*Nuestra madre tierra*" es fluido pero

inadecuado al omitir la cosmovisión de *Pacha* (tiempo-espacio sagrado), mientras "*Madre del cosmos-tiempo nuestro*", aunque semánticamente precisa, resulta disfluente en español estándar ((Rios, 2015)). Esta tensión se acentúa en lenguas de bajos recursos, donde el 42 % de las traducciones presentan disociación entre fluidez y adecuación (Neubig y Hu, 2018).

Para operacionalizar estas dimensiones, se emplean protocolos híbridos:

1. **Escalas Likert** (1–5) con descriptores específicos:
   - **Fluidez**: 1 (incomprensible) $\rightarrow$ 5 (nativo).
   - **Adecuación**: 1 (traición semántica) $\rightarrow$ 5 (equivalencia cultural total) (Lommel, 2018).
2. **Guías de anotación contextual**:
   - Definición de "errores graves" (ej.: omisión de evidenciales) vs. "leves" (variación léxica aceptable).
3. **Evaluadores bilingües**:
   - Hablantes nativos de quechua con dominio de cosmovisión andina (Bird, 2020b).

Capítulo III

El Quechua como caso de estudio

El quechua, familia lingüística hablada por 8-10 millones de personas en la región andina (Perú, Bolivia, Ecuador, Colombia, Argentina y Chile), constituye la lengua indígena viva más extendida de América (Adelaar, 2004a). Pertenece a la rama quechumara y se caracteriza por su polisíntesis aglutinante, evidencialidad obligatoria y transmisión oral predominante (Cerrón-Palomino, 2003b). Pese a su estatus oficial en varios países, enfrenta una brecha digital crítica: menos del 0.3 % de los recursos computacionales existentes para el español están disponibles para el quechua, limitando el desarrollo de herramientas de traducción automática (Rios, 2015). Su vitalidad lingüística varía desde variantes vigorosas (Quechua Collao) hasta variedades en peligro (Quechua de Santiago del Estero), configurando un escenario complejo para el procesamiento automático.

3.1. Características lingüísticas

3.1.1. Complejidad morfológica

La morfología quechua opera mediante *aglutinación altamente productiva*, donde una raíz admite múltiples sufijos que codifican significado gramatical y pragmático. Un solo lexema puede generar hasta 2,000 formas mediante combinaciones de sufijos de persona, tiempo, modo, evidencialidad y dirección (Adelaar, 2004a) Por ejemplo, la palabra *llank'achkarpusaq* se descompone en:

- Raíz: *llank'a* (trabajar)
- Sufijos: *-chka* (progresivo) + *-rpu* (intensivo) + *-saq* (futuro 1ª persona)

- *Trabajaré intensamente pronto*.

Esta complejidad desafía los tokenizadores estándar: herramientas como SentencePiece segmentan incorrectamente el 38 % de palabras aglutinadas, rompiendo unidades semánticas indivisibles (R. Zevallos et al., 2024). En traducción automática, esto genera errores de omisión en el 63 % de sufijos evidenciales (*-mi*, *-si*), cruciales para transmitir la fuente del conocimiento. Ejemplo:

- *Paraqayamun* → "Está lloviendo [yo lo veo]"
- *Paraqayamunsi* → "Está lloviendo [me contaron]"

Además, el quechua emplea *sufijos posesivos fusionados* que combinan persona y número en un único morfema:

- *-yki* = 2ª persona singular (*wasyiki* = "tu casa")
- *-nchik* = 1ª persona plural inclusiva (*wasinchik* = "nuestra casa" [incluyendo al oyente])

Modelos neuronales como MarianMT confunden estos sufijos en el 27 % de casos, generando ambigüedades relacionales (Rios et al., 2021).

3.1.2. Variación dialectal

El quechua presenta una diversificación dialectal extrema, clasificada en dos ramas principales con inteligibilidad mutua limitada (Cerrón-Palomino, 2003a):

1. **Quechua I (Central)**: Hablado en Perú central (Ancash, Huánuco), con rasgos como la aspiración de oclusivas (*qhatu*).
2. **Quechua II (Meridional)**: Incluye:
   - *Collao* (Cusco, Puno): Conserva oclusivas uvulares /q/.
   - *Chanka* (Ayacucho): Simplifica /q/ a /X/.
   - *Kichwa* (Ecuador): Pérdida de vocales uvulares.

Concepto	Collao	Chanka	Kichwa
Agua	yaku	unu	yaku
Hombre	qhari	nuna	runa
Maíz	sara	awa	sara

Estas variantes exhiben divergencias léxicas críticas:

La falta de estandarización ortográfica agrava el problema: el fonema /q/ se escribe como *q*, *k* o *c* según región (Cusihuamán et al., 1976). Esto fragmenta los recursos computacionales: el corpus *Monolingual-Quechua IIC* contiene 62 % textos Collao vs 19 % Chanka (R. Zevallos et al., 2024), sesgando modelos como *NLLB-200* que muestran 22 % menor precisión en Chanka.

En traducción automática, la variación dialectal requiere estrategias específicas:

- Normalización gráfica: Unificar "llank'ay"(Collao) y "llank'ai"(Chanka) en una representación común.
- Modelos adaptativos: Entrenar variantes por separado usando códigos de idioma (ej: quy para Collao, quz para Chanka). Estudios demuestran que estas estrategias mejoran la precisión semántica en un 31 % frente a modelos monolíticos (Rios, 2015).

3.2. Recursos y herramientas disponibles

El desarrollo de herramientas para el quechua enfrenta una disponibilidad asimétrica: mientras el español cuenta con miles de recursos, el quechua dispone de conjuntos limitados y fragmentados. El principal corpus, Monolingual-Quechua IIC (R. Zevallos et al., 2024), integra 384,184 oraciones (4.4 millones de tokens) de 50 fuentes heterogéneas, con una distribución desigual por dominio: cultura (24 %), salud (18 %), educación (15 %), religión (12 %), política (11 %) y misceláneos (20 %). Pese a su amplitud, adolece de tres limitaciones críticas: 1) solo el 8 % incluye narrativa oral (crucial para términos culturales), 2) predominio del quechua Collao (62 %) sobre el Chanka (19 %), y 3) ausencia de anotación morfológica para sufijos evidenciales (Zevallos et al., 2022). Este desbalance sesga los modelos hacia registros formales, omitiendo variantes dialectales y léxico comunitario.

En cuanto a modelos de traducción, destacan dos iniciativas:

- **MarianMT para quechua-español:** Basado en Transformers, fue fine-tuneado con 42,000

oraciones paralelas de AmericasNLP 2021. Logra un BLEU de 22.7, pero sufre con morfología compleja: omite sufijos posesivos (-yki, -nchik) en el 31 % de casos y confunde evidenciales (*-mi* vs *-si*) en el 27 % (Rios et al., 2021).

- **NLLB-200 de Meta (2022):** Incluye quechua entre 200 lenguas, pero con rendimiento dispar: 18.2 BLEU en Collao vs 14.1 en Chanka. Su principal falla es la traducción literal de términos culturales: "Pachamama" "Madre Tierra"(ignorando Pacha como çosmos-tiempo"), y .ªyni" .ªyuda"(perdiendo reciprocidad) (Rios, 2011)

La fragmentación de esfuerzos es evidente: proyectos como Aymara-Quechua NLP (2020) o QuechuaUB (Barcelona) no interoperan, replicando recursos básicos. Esto explica que solo el 12 % del léxico cultural ("*apu*", "*huaca*") tenga embeddings en repositorios como FastText (Joulin et al., 2016). La consecuencia es una dependencia crítica de back-translation con datos sintéticos no verificados, que introducen errores culturales en el 41 % de traducciones automáticas evaluadas (Neubig y Hu, 2018).

Capítulo IV

La Propuesta

4.1. Tipo y Diseño de la Investigación

Esta investigación se clasifica como cuantitativa y descriptiva. El diseño corresponde a un estudio no experimental, ya que se analizan modelos de traducción automática existentes sin manipular variables, midiendo su impacto en la calidad semántica en un momento específico. El enfoque es comparativo utilizando muestreo no probabilístico de textos en quechua y análisis estadístico descriptivo para responder a los objetivos planteados.

4.2. Proceso metodológico

El proceso metodológico se estructura en cinco fases interrelacionadas:

- Selección y preparación del corpus.
- Generación de traducciones automáticas
- Cálculo de métricas automáticas de calidad semántica*
- Evaluación humana de las traducciones
- Análisis comparativo de resultados

Figura 4.1: Fases de la evaluación de la calidad semántica de las TA quechua-español

4.3. Corpus de Estudio

El dataset empleado para este trabajo es el Monolingual Quechua IIC (J. Zevallos et al., 2022), que consta de 4,408,953 tokens y 384,184 sentencias con variantes de quechua Collao y Chanka de la rama de Quechua II. Este corpus es una compilación de 50 corpus monolingües de diferentes fuentes y que abarco varios dominios como: religión, economa, salud, cultura, politica y misceláneos.

4.4. Modelos de traducción automática

Se evaluarán tres modelos de traducción automática representativos de distintos enfoques tecnológicos:

Tabla 4.1: Comparación de modelos de traducción automática

Modelo	**Tipo**	**Implementación**
Google Translate	Comercial (Transformer multilingüe)	API pública `googletrans`
MarianMT	Especializado en lenguas minoritarias	Modelo `Helsinki-NLP/opus-mt-quc-es` de Hugging Face
Baseline Léxico	Traducción basada en reglas	Diccionario quechua-español (Siminchik) + orden SVO

Nota. *Elaboración propia.*

4.5. Métricas de Calidad Semántica

Para la evaluación de calidad semántica, se emplean dos enfoques complementarios. El primero es un enfoque cuantitativo. Dentro de éste, se han establecido res métricas complementarias para evaluar la equivalencia semántica:

4.5.1. Similitud Coseno con LaBSE

- **Objetivo:** Medir equivalencia conceptual global.
- **Fundamento técnico:** Representación vectorial en espacio semántico multilingüe.
- **Escala:** 0 (sin relación) a 1 (equivalencia total).

$$\text{Similitud}(\mathbf{q}, \mathbf{t}) = \frac{\mathbf{q} \cdot \mathbf{t}}{\|\mathbf{q}\| \, \|\mathbf{t}\|} \tag{4.1}$$

Donde:

- **q y t**: Vectores generados por *LaBSE*.
- $\cdot$: Producto punto.
- $\|\mathbf{q}\|$ | y $\|\mathbf{t}\|$ |: Normas Euclidianas de los vectores.

Esta métrica es particularmente relevante para capturar equivalencias no literales y adaptaciones culturales.

4.5.2. COMET-QE (Quality Estimation)

- **Objetivo:** Evaluar calidad intrínseca sin referencia humana
- **Fundamento técnico:** Modelo transformer preentrenado en evaluación de traducciones
- **Escala:** Puntuación continua (mayor valor indica mejor calidad)

Fórmula (modelo Transformer):

$$\text{COMET-QE} = f_\theta(\text{src}, \text{mt}) \tag{4.2}$$

Donde f_θ es un modelo preentrenado que estima la calidad mediante:

$$f_\theta = \text{TransformerEncoder}(\text{Embed}(\text{src}) \oplus \text{Embed}(\text{mt})) \tag{4.3}$$

4.5.3. chrF++

- **Objetivo:** Evaluar precisión léxica y morfológica
- **Fundamento técnico:** N-grams de caracteres adaptados a lenguas aglutinantes
- **Escala:** 0 a 1 (1 = máxima precisión)

$$\text{chrF}_{++} = (1+\beta^2) \cdot \frac{\text{chrP} \cdot \text{chrR}}{\beta^2 \cdot \text{chrP} + \text{chrR}} \tag{4.4}$$

Donde:

- chrP: Precisión de n-gramas de caracteres (hasta 6-gram).
- chrR: Recall de n-gramas de caracteres.
- β: Peso para el recall (usualmente $\beta = 2$).

4.5.4. Evaluación Humana

Paralelamente, se realiza una evaluación humana con cinco hablantes bilingües que califican fluidez (gramaticalidad y naturalidad) y adecuación (preservación de significado cultural) mediante escalas Likert, analizando 30 frases por modelo para identificar discrepancias entre métricas automáticas y percepción nativa.

4.6. Análisis de Datos

En el análisis de datos se emplearán varios enfoques para evaluar y comparar los resultados obtenidos de los diferentes modelos de traducción automática. Para la **comparación entre modelos**, se aplicará un **ANOVA unidireccional**, con el objetivo de identificar si existen diferencias significativas entre los promedios de los modelos evaluados. Posteriormente, se realizarán **pruebas post-hoc de Tukey** para realizar comparaciones más detalladas entre los pares de modelos, a fin de determinar cuáles son los modelos que presentan diferencias significativas. En todas estas pruebas, el **nivel de significancia** se establecerá en $p < 0{,}05$, lo que indicará si los resultados son estadísticamente significativos.

En cuanto a la **correlación entre las métricas automáticas y las evaluaciones humanas**, se utilizará el **coeficiente de correlación de Pearson** para medir la relación entre los resultados obtenidos a través de las métricas automáticas y las evaluaciones realizadas por los evaluadores humanos. Además, se generarán **diagramas de dispersión** para visualizar de forma gráfica la relación entre las métricas y las evaluaciones por cada métrica, lo que permitirá identificar patrones o inconsistencias en las evaluaciones.

Bibliografía

Adelaar, W. F. (2004a). *The languages of the andes*. Cambridge University Press.

Adelaar, W. F. (2004b). *The languages of the andes*. Cambridge University Press.

Agic, Ž., y Vulic, I. (2019). Jw300: A wide-coverage parallel corpus for low-resource languages..

Arnold, D. (1994). Machine translation: an introductory guide. *(No Title)*.

Bahdanau, D., Cho, K., y Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bender, E. M. (2011). On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, *6*.

Bird, S. (2020a). Decolonising speech and language technology. En *28th international conference on computational linguistics, coling 2020* (pp. 3504–3519).

Bird, S. (2020b). Decolonizing speech and language technology. En *Proceedings of the 58th annual meeting of the acl* (pp. 1–13).

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., y Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, *19*(2), 263–311.

Cardenas, R., Zevallos, R., Baquerizo, R., y Camacho, L. (2018). Siminchik: A speech corpus

for preservation of southern quechua. *ISI-NLP*, *2*, 21.

Cerrón-Palomino, R. (2003a). *Lingúística quechua*. Centro Bartolomé de Las Casas.

Cerrón-Palomino, R. (2003b). Lingüística quechua (edición). *Cusco: Centro de Estudios Regionales Andinos' Bartolomé de las Casas*.

Costa-Jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., ... others (2022). No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Cotterell, R., Kirov, C., Hulden, M., y Eisner, J. (2019). On the complexity and typology of inflectional morphological systems. *Transactions of the Association for Computational Linguistics*, *7*, 327–342.

Cusihuamán, G., y cols. (1976). Gramática quechua, cuzco-collao. *(No Title)*.

De Gibert, O., Pugh, R., Marashian, A., Vázquez, R., Ebrahimi, A., Denisov, P., ... others (2025). Findings of the americasnlp 2025 shared tasks on machine translation, creation of educational material, and translation metrics for indigenous languages of the americas. En *Proceedings of the fifth workshop on nlp for indigenous languages of the americas (americasnlp)* (pp. 134–152).

Faisal, F., Ahia, O., Srivastava, A., Ahuja, K., Chiang, D., Tsvetkov, Y., y Anastasopoulos, A. (2024). Dialectbench: A nlp benchmark for dialects, varieties, and closely-related languages. *arXiv preprint arXiv:2403.11009*.

Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., ... others (2021). Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, *22*(107), 1–48.

Feng, F., y cols. (2022). Language-agnostic BERT sentence embedding. En *Proceedings of acl* (pp. 878–891).

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., y Wang, W. (2020). Language-agnostic bert

sentence embedding. *arXiv preprint arXiv:2007.01852*.

Fernandez-Sabido, S., y Peniche-Sabido, L. (2025). Redefining technology for indigenous languages. *arXiv preprint arXiv:2504.01522*.

Hutchins, W. J. (1986). *Machine translation: past, present, future*. Ellis Horwood Chichester.

Joshi, P., y cols. (2020). The state and fate of linguistic diversity. *Proceedings of the National Academy of Sciences*, *117*(23), 23368–23377.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., y Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., y Mikolov, T. (2016). Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Junczys-Dowmunt, M., y cols. (2018). Marian: Fast neural machine translation in C++. En *Proceedings of acl.*

Kay, M. (1997). The proper place of men and machines in language translation. *machine translation*, *12*, 3–23.

Koehn, P. (2020). *Neural machine translation*. Cambridge University Press.

Lauscher, A., Ravishankar, V., Vulić, I., y Glavaš, G. (2020). From zero to hero: On the limitations of zero-shot cross-lingual transfer with multilingual transformers. *arXiv preprint arXiv:2005.00633*.

Lommel, A. (2018). Translation quality metrics. En *Proceedings of the amta 2018 workshop on the role of authoritative standards in the mt environment* (pp. 69–94).

Lommel, A., y cols. (2014). Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica*, *12*, 455–463.

Lommel, A., Uszkoreit, H., y Burchardt, A. (2014). Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica*(12), 0455–463.

Neubig, G. (2017). Neural machine translation and sequence-to-sequence models: A tutorial. *arXiv preprint arXiv:1703.01619*.

Neubig, G., y Hu, J. (2018). Rapid adaptation of neural machine translation to new languages. *arXiv preprint arXiv:1808.04189*.

Och, F. J., y Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics*, *29*(1), 19–51.

Papineni, K., y cols. (2002). Bleu: a method for automatic evaluation of machine translation. En *Proceedings of acl* (pp. 311–318).

Papineni, K., Roukos, S., Ward, T., y Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. En *Proceedings of the 40th annual meeting of the association for computational linguistics* (pp. 311–318).

Pires, T., Schlinger, E., y Garrette, D. (2019). How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Popović, M. (2017). chrf++: words helping character n-grams. En *Proceedings of the second conference on machine translation* (pp. 612–618).

Rei, R., De Souza, J. G., Alves, D., Zerva, C., Farinha, A. C., Glushkova, T., ... Martins, A. F. (2022). Comet-22: Unbabel-ist 2022 submission for the metrics shared task. En *Proceedings of the seventh conference on machine translation (wmt)* (pp. 578–585).

Rios, A. (2011). Spell checking an agglutinative language: Quechua.

Rios, A. (2015). *A basic language technology toolkit for quechua* (Tesis Doctoral no publicada). University of Zurich.

Rios, A., y cols. (2021). Cultural adequacy in machine translation. En *Proceedings of americasnlp* (pp. 1–15).

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

Shi, J., Amith, J. D., Chang, X., Dalmia, S., Yan, B., y Watanabe, S. (2021). Highland puebla nahuatl speech translation corpus for endangered language documentation. En *Proceedings of the first workshop on natural language processing for indigenous languages of the americas* (pp. 53–63).

Team, N., Costa-Jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., ... others (2022). No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Tiedemann, J., y Thottingal, S. (2020). Opus-mt–building open translation services for the world. En *Annual conference of the european association for machine translation* (pp. 479–480).

Torres, L., y cols. (2023). Digital resource gaps for quechua: A computational analysis. *Language Resources and Evaluation*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Weaver, W. (1952). Translation. En *Proceedings of the conference on mechanical translation.*

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... others (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Zevallos, J., y cols. (2022). *Monolingual-quechua-iic: Technical documentation* (Inf. Téc.). Instituto de Investigación en Ciencia de la Computación.

Zevallos, R., Bel, N., y Farrús, M. (2024). Tema: Token embeddings mapping for enriching low-resource language models. En *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 11423–11435).

Zhang, T., Kishore, V., Felix, W., y Weinberger, K. Q. (2020). Evaluating text generation with

bert. *ICLR, Bertscore*.

Zoph, B., Yuret, D., May, J., y Knight, K. (2016). Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.

Anexo