www.editada.org

International Journal of Combinatorial Optimization Problems and Informatics, 17(5), 2026, 168-182. ISSN: 2007-1558.
https://doi.org/10.61467/2007.1558.2026.v17i5.1451

RAÍZ: A Hybrid Transformer-Based Machine Translation Architecture with Linguistic Preprocessing and Community Validation for Low-Resource Indigenous Languages (Spanish–Nahuatl and Spanish–Mazahua)

Gabriela Salas-Cabrera[1], Aníbal Alejandro Gallegos-Macías[2], Francisco Rafael Trejo-Macotela[1]*

[1] Polytechnic University of Pachuca, Mexico.

[2] Consejo de Ciencia, Tecnología e Innovación de Hidalgo, Mexico.

E-mail: salascabrerag@gmail.com, anibal.gallegos@hidalgo.gob.mx, trejo_macotela@upp.edu.mx

* Corresponding author: trejo_macotela@upp.edu.mx

Abstract. Machine translation rarely reaches the world's Indigenous languages, which remain marginal to modern natural language processing owing to scarce digital corpora, inconsistent orthography, dialectal variation and morphological richness that standard models were not designed to absorb. **Objective:** this paper introduces and evaluates RAÍZ, a hybrid machine translation architecture conceived for such low-resource settings, and assesses whether pairing neural learning with explicit linguistic knowledge and community participation improves translation quality. **Methodology:** RAÍZ combines a multilingual Transformer backbone (ByT5, fine-tuned with supervised learning) with linguistically informed components — phonetic normalisation, morphological segmentation, multilingual semantic representation, community lexical retrieval and a native-speaker review loop. The framework was evaluated on Spanish–Nahuatl and Spanish–Mazahua translation using a parallel corpus of 42,000 aligned sentence pairs (23,000 and 19,000 respectively), split 80/10/10, and assessed through a comparative evaluation against mT5, ByT5 and NLLB baselines using BLEU, chrF and COMET, supplemented by qualitative judgements of clarity, naturalness, semantic fidelity and cultural appropriateness. **Key results:** the Spanish–Nahuatl task achieved 28.6 BLEU, 56.8 chrF and 74% COMET, while the Spanish–Mazahua task reached 27.4 BLEU, 54.9 chrF and 71% COMET, with stable performance across both language pairs despite differences in corpus size. **Contribution:** the study offers a reproducible, scalable methodological framework that integrates corpus development, dataset traceability (registered with the Instituto Nacional de los Pueblos Indígenas), baseline comparison and community-based assessment within a single workflow, treating native speakers as collaborators rather than afterthoughts. **Impact:** the findings provide evidence that low-resource translation improves when neural learning is combined with linguistically informed and culturally aware processing, and offer a transferable foundation for future research on Indigenous language technology, language preservation and community-centred NLP.

Keywords: Low-resource machine translation; Indigenous language processing; Nahuatl–Spanish translation; Mazahua–Spanish translation; Hybrid neural architecture; Transformer-based multilingual model; Linguistic preprocessing; Community-based validation; Multilingual neural machine translation; Low-resource NLP; Language preservation technology; Culturally grounded AI.

Article Info
Received May 29, 2026
Accepted Jul 6, 2026

1 Introduction

Indigenous languages do more than carry messages between speakers. They hold historical knowledge, collective memory, cultural practices and community identities that have accumulated over generations. Yet their presence in today's digital environment remains thin. The gap shows up most clearly in machine translation, information retrieval, conversational agents, educational technology and digital accessibility tools. One explanation stands out: natural language processing (NLP) has long concentrated on languages that enjoy large corpora, mature linguistic resources and serious computational infrastructure.

The past decade has transformed NLP through deep neural networks and Transformer-based architectures, which have sharply improved how models capture contextual relationships in language understanding and translation (Vaswani et al., 2017). The catch is that these models tend to thrive only where training data are plentiful, writing conventions are fairly standardised and textual resources are rich. Indigenous languages rarely meet those conditions. Data are scarce, orthography varies, and digital representation is patchy. There is therefore a real need for translation approaches that do not simply rely on neural learning, but pair it with explicit linguistic knowledge and validation procedures sensitive to local language contexts.

RAÍZ was designed with this need in view. It is a hybrid machine translation architecture built specifically for low-resource Indigenous languages, one that joins advances in deep learning with linguistic expertise and community participation. The goal is not merely to push automatic evaluation scores higher. Rather, the architecture offers a methodological framework that can cope with orthographic variation, morphological complexity, multilingual transfer, semantic preservation and the cultural suitability of translated content.

Nahuatl served as the primary validation language, with Mazahua included as a comparative case. Testing the framework across these two settings lets us examine both its effectiveness and its capacity to adapt to different Indigenous language scenarios. The corpus and associated datasets used in the study have been registered as linguistic resources with the Instituto Nacional de los Pueblos Indígenas (INPI), which reinforces the traceability, preservation and institutional recognition of the materials. Should the editorial process require it, the relevant registration details can be supplied in the final version of the manuscript.

This article presents, contextualises and assesses RAÍZ as a hybrid alternative for improving machine translation quality in low-resource Indigenous language environments. The paper begins by reviewing the relevant literature and the current state of the art, then describes the experimental methodology and the characteristics of the corpus. It goes on to examine the quantitative and qualitative results, and closes by discussing the main limitations of the study and possible directions for future research.

2 Background and State of Art

Natural language processing has changed almost beyond recognition over the past few decades. The earliest systems leaned on manually defined linguistic rules; statistical learning, recurrent neural architectures and, most recently, Transformer-based models followed in turn. This trajectory reflects a wider movement away from symbolic representations of language and towards data-driven methods that model contextual information with growing sophistication. Linguistic structure, probabilistic reasoning and semantic representation have therefore become progressively woven into contemporary NLP systems (Jurafsky & Martin, 2023).

The success of these models, however, is not unconditional. Across a broad range of understanding and generation tasks they perform impressively, but that performance generally presupposes access to extensive, well-curated datasets. In low-resource settings the requirement is hard to meet, and Indigenous languages illustrate the difficulty clearly: their corpora are often small, uneven in quality, or marked by orthographic inconsistency. Before Transformers took hold, recurrent architectures such as Long Short-Term Memory (LSTM) networks marked a genuine step forward, easing the vanishing-gradient problem and improving how contextual dependencies were modelled across sequential data (Hochreiter & Schmidhuber, 1997).

Neural machine translation owes much of its shape to encoder–decoder frameworks and attention-based mechanisms. The decisive moment came with the Transformer, whose self-attention mechanism models global relationships among tokens while reducing reliance on sequential computation and improving training efficiency (Vaswani et al., 2017). Multilingual

models soon built on this foundation: mT5 carried the text-to-text paradigm to more than a hundred languages (Xue et al., 2021), while ByT5 showed that operating directly on byte-level representations makes a model more resilient to textual noise, spelling variation and tokenisation problems (Xue et al., 2022). Both descend from the broader T5 framework, which folded a wide range of NLP tasks into a single text-to-text formulation and demonstrated the promise of transfer learning across heterogeneous linguistic domains (Raffel et al., 2020).

NLLB pushed multilingual coverage further, aiming to extend translation to languages with limited digital resources. Its evaluation combined the Flores-200 benchmark with human assessment across thousands of translation directions (NLLB Team et al., 2022). Yet broad coverage alone does not settle the questions that matter most for Indigenous languages. Dialectal diversity, complex morphology, cultural appropriateness and systematic community validation tend to be under-represented in general-purpose systems, and it is precisely these gaps that have fuelled interest in hybrid approaches pairing neural modelling with linguistically informed processing.

Research on the Indigenous languages of the Americas has gathered momentum through initiatives such as AmericasNLP, which has created shared evaluation spaces devoted to machine translation, morphological adaptation and language technology development for Indigenous communities. The earlier editions of these shared tasks exposed persistent difficulties: data scarcity, orthographic variation, and evaluation practices that rarely fit Indigenous language settings (Mager et al., 2021). More recent work suggests that building successful language technology depends on more than model architecture alone. Corpus availability, cultural representation, evaluation procedures that reflect local linguistic realities, and the active participation of native speakers throughout the research process all matter (de Gibert et al., 2025).

A review of the literature brings several recurring challenges into focus, none of them fully resolved. To begin with, many multilingual architectures lean almost entirely on statistical learning and never make explicit use of phonetic or morphological knowledge. Second, automatic evaluation metrics say little about naturalness, clarity or cultural appropriateness. Third, orthographic diversity and dialectal variation can undermine systems that depend on conventional tokenisation. Finally, growing attention has turned to the ethical and methodological importance of documenting corpus provenance, ensuring transparency in data governance, and involving native speakers in validation. RAÍZ was designed with these interconnected problems in mind.

To place the proposed framework within the wider research landscape, Table 1 summarises the principal methodological trends, multilingual architectures and shared-task initiatives that have shaped recent work on machine translation for low-resource languages. Alongside their main contributions, the table identifies limitations that remain relevant to Indigenous language scenarios and indicates how these observations fed into the design of RAÍZ.

Table 1. Related Work and Methodological Relevance for RAÍZ.

Research line / model	Main contribution	Relevance	Observed limitation	Relationship with RAÍZ
Transformers	Self-attention and global contextual modelling	Enables neural machine translation	Requires substantial data and computational resources	Neural foundation of the translation module
mT5	Multilingual text-to-text model trained on 101 languages	Supports transfer learning across languages and tasks	Limited coverage for many Indigenous languages	General multilingual baseline
ByT5	Byte-level processing	Robust against noise and orthographic variation	Longer sequences and increased computational cost	Useful backbone for written variation
NLLB	Large-scale multilingual translation	Designed for low-resource languages	Does not independently resolve cultural validation challenges	Translation-oriented specialised baseline
AmericasNLP	Shared tasks for Indigenous languages of the Americas	Enhances visibility of Indigenous datasets, models, and evaluation practices	Persistent limitations regarding corpora and evaluation metrics	Supports the need for community-centred evaluation

The comparative analysis in Table 1 supports a hybrid, community-oriented perspective, particularly where purely data-driven solutions struggle to accommodate linguistic diversity, orthographic variability and culturally situated forms of language use.

3 Experimental Procedures

The study followed an applied technological approach with descriptive, explanatory, experimental and comparative dimensions. The analysis drew on the Spanish–Nahuatl and Spanish–Mazahua parallel corpora, on the translations produced by the evaluated models, and on the judgements supplied by native speakers during the qualitative validation stage.

Corpus preparation followed a sequence common in NLP workflows: text cleaning, sentence alignment verification, orthographic consistency checking, and linguistically informed preprocessing (Bird et al., 2009). Once these steps were complete, the dataset contained 42,000 parallel sentence pairs — 23,000 Spanish–Nahuatl and 19,000 Spanish–Mazahua. The corpus was split into training, validation and testing subsets in an 80/10/10 ratio, and the same partitioning was applied across every experiment to keep the methodology consistent and the comparisons fair.

Table 2 sets out the distribution of the parallel corpus and the experimental split. Beyond reporting how sentence pairs were allocated across the subsets, it also summarises the procedures used to ensure dataset traceability, documentation and institutional registration.

Table 2. Corpus Distribution, Experimental Split, and Data Governance.

Language pair	Total pairs	Training 80%	Validation 10%	Test 10%	Data governance
Spanish-Nahuatl	23,000	18,400	2,300	2,300	Reviewed and documented parallel corpus
Spanish-Mazahua	19,000	15,200	1,900	1,900	Reviewed and documented parallel corpus
Total	42,000	33,600	4,200	4,200	Dataset/corpus with traceability and institutional registration before the Instituto Nacional de los Pueblos Indígenas (INPI)

As Table 2 shows, the Spanish–Nahuatl corpus holds more aligned sentence pairs than the Spanish–Mazahua corpus. This difference may affect the amount of linguistic information available during training, but keeping both language pairs within a common experimental setting allows us to examine how the proposed architecture transfers across distinct low-resource conditions. The data-governance practices also contribute to the transparency and reproducibility of the research.

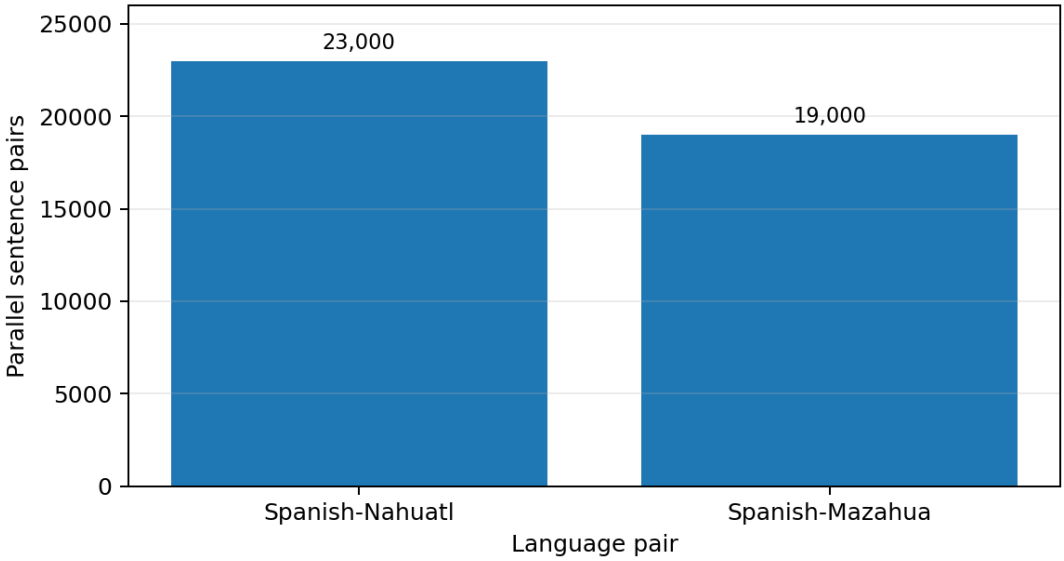

Figure 1. Corpus Distribution by Language Pair.

Figure 1 shows the distribution of sentence pairs for each language combination. The horizontal axis corresponds to the translation pairs and the vertical axis to the number of available aligned sentences. The figure makes the unequal size of the two subsets immediately apparent, and the interpretation of the experimental results takes this imbalance into account.

3.1 RAÍZ Architecture

RAÍZ was conceived as a hybrid architecture that joins neural modelling to linguistically informed processing. Its starting assumption is that Indigenous languages should not be treated as mere sequences of textual tokens, but as complex linguistic systems shaped by phonetic variation, rich morphological structures, community-specific language practices and culturally situated forms of expression. Earlier work supports this view, showing that subword and morphological representations can ease data sparsity and improve translation performance in low-resource settings (Sennrich et al., 2016).

Table 3 summarises the main components of the architecture and their functions. Rather than relying on a single end-to-end translation model, RAÍZ brings together a set of complementary mechanisms that address the linguistic, semantic, retrieval-oriented and community-based aspects of translation.

Table 3. Components of the RAÍZ Hybrid Architecture.

Component	Function	Expected contribution
Phonetic normalisation	Reduces orthographic variation and approximates equivalent forms	Minimises textual noise and improves input regularity
Morphological segmentation	Separates meaningful units and morphological patterns	Enhances generalisation in languages with high morphological productivity
Multilingual semantic representation	Projects sequences into a shared semantic space	Facilitates transfer between Spanish and Indigenous languages
Community lexical retrieval and memory	Retrieves validated equivalences and linguistic registers	Promotes terminological consistency and lexical preservation
Transformer module	Generates translations through contextual learning	Provides neural capacity for sequential and semantic modelling
Community validation	Incorporates review by native speakers	Evaluates clarity, naturalness, fidelity, and cultural appropriateness

Table 3 reflects the modular character of the framework. Translation does not emerge from one computational stage; it arises from the interaction of several interconnected components. This design suits low-resource Indigenous languages well, because it makes it easier to integrate explicit linguistic knowledge, to withstand orthographic and morphological variation, and to refine the system iteratively through community participation. Neural representation is thus complemented by evaluation criteria grounded in real language use.

The retrieval layer also draws on principles from information retrieval and lexical access systems, improving consistency across semantically related expressions and validated terminology (Manning et al., 2008).

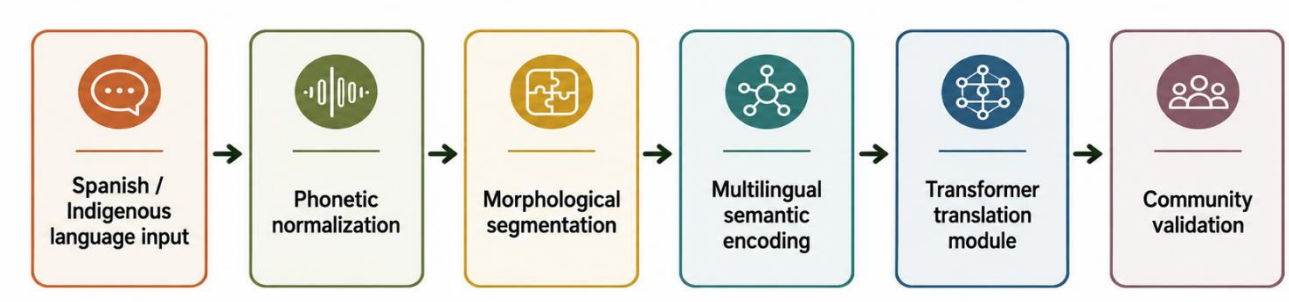

Figure 2. RAÍZ Hybrid Translation Pipeline.

Figure 2 depicts the overall workflow within RAÍZ. Incoming text first passes through phonetic normalisation and morphological segmentation before being mapped into multilingual semantic representations. The Transformer-based translation module then generates the target-language output, which native speakers review. As the figure illustrates, community validation is an integral part of the translation cycle and feeds directly into the iterative improvement of the framework.

3.2 Architectural and Computational Foundations of RAÍZ

The conceptual foundations of RAÍZ reach back into the history of natural language processing, and in particular the move from rule-based and statistical approaches towards neural sequence-modelling architectures (Jurafsky & Martin, 2023; Manning, Raghavan, & Schütze, 2008). The earliest NLP systems depended heavily on handcrafted linguistic rules and probabilistic n-gram models. Those methods worked well enough within restricted domains, but they scaled and adapted poorly when confronted with morphologically rich languages and scarce linguistic resources.

Recurrent architectures — Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) models — represented a genuine advance in sequence modelling by preserving contextual information across linguistic sequences (Cho et al., 2014; Hochreiter & Schmidhuber, 1997). They demonstrated that meaning does not reside in isolated lexical items alone, but also in the relationships distributed throughout a sequence (Goldberg, 2017). Even so, recurrent models frequently struggled with long-range dependencies, gradient degradation and reduced effectiveness when training data were limited.

RAÍZ was developed against this backdrop as a hybrid framework that combines multilingual Transformer-based semantic representations with complementary linguistic resources and community-centred validation mechanisms (Vaswani et al., 2017; Xue et al., 2021). Rather than depending entirely on end-to-end neural inference, the architecture incorporates lexical retrieval, linguistic preprocessing, contextual refinement and native-speaker review to strengthen semantic coherence and translation reliability in low-resource Indigenous language settings. In this way, the framework aims to balance the expressive power of deep learning with the advantages of explicit linguistic knowledge and culturally informed evaluation practices.

The semantic modelling capabilities of the Transformer component rest on scaled dot-product attention, a mechanism that represents contextual relationships across multilingual token sequences (Vaswani et al., 2017).

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

In this formulation, Q, K, and V denote the query, key and value matrices respectively, while the scaling factor helps stabilise the attention distribution during multilingual semantic representation learning.

The operational workflow within RAÍZ integrates multilingual semantic representation, contextual linguistic refinement, retrieval-assisted adaptation and community-informed validation into a unified low-resource translation framework. Algorithm 1 summarises the principal computational stages involved in training, inference and qualitative evaluation.

Algorithm 1. General Operational Workflow of RAÍZ

Require: Parallel Spanish–Nahuatl and Spanish–Mazahua corpora
Ensure: Trained multilingual translation framework and evaluated translation outputs
1. Load and preprocess the parallel corpora
2. Apply orthographic normalisation and linguistic cleaning
3. Perform morphological segmentation and auxiliary linguistic annotation
4. Partition the dataset into training, validation, and testing subsets
5. Initialise the multilingual ByT5 backbone architecture
6. Integrate phonetic, morphological, and semantic refinement modules
7. Configure contextual retrieval memory mechanisms
8. **for** each training iteration **do**
9. Encode multilingual input sequences
10. Integrate contextual linguistic representations
11. Retrieve semantically relevant contextual information when available
12. Generate multilingual translation outputs
13. Assess translation consistency and semantic preservation
14. **end for**
15. Compute BLEU, chrF, and COMET evaluation metrics
16. Perform qualitative validation with native speakers
17. Analyse linguistic consistency and contextual appropriateness
18. Refine the framework according to community-informed feedback

This workflow allows RAÍZ to function as a modular translation framework in which neural semantic modelling is complemented by linguistic knowledge and context-aware adaptation mechanisms tailored to low-resource Indigenous language settings.

Together, these architectural components provided the basis for the experimental implementation of RAÍZ, which called for a training configuration capable of supporting multilingual semantic representations under conditions of limited linguistic resources.

3.3 Explainability and Linguistic Traceability

RAÍZ stands apart from fully opaque end-to-end translation systems. It was designed as a modular framework in which several stages of the translation process remain identifiable and open to linguistic interpretation. Orthographic preprocessing, lexical retrieval support, contextual refinement and native-speaker validation together provide a degree of traceability across the workflow, which makes it easier to spot outputs that are semantically inconsistent or culturally inappropriate.

The interpretability of the framework does not rest on post hoc explanation techniques alone. Transparency emerges instead from the explicit incorporation of linguistically informed processing stages and community-based validation practices (Doshi-Velez & Kim, 2017; Rudin, 2019). This arrangement allows lexical ambiguities, contextual mismatches and semantic adjustments to be examined collaboratively during evaluation, supporting a translation framework that is at once more transparent and more responsive to the linguistic and cultural realities of Indigenous language communities.

3.4 Training Configuration and Implementation Details

The experimental setup was designed with the specific challenges of low-resource Indigenous language translation in mind. RAÍZ was implemented by supervised fine-tuning of the multilingual ByT5 Transformer architecture (Xue et al., 2022), chosen for its demonstrated ability to cope with orthographic variability and morphologically rich linguistic structures.

Before training, the corpus passed through several preprocessing stages: phonetic normalisation, morphological segmentation, data cleaning and multilingual token standardisation. These steps were meant to reduce the inconsistencies that arise from spelling variation and morphological fragmentation, both of which appear frequently in Nahuatl and Mazahua textual resources.

Training ran with an effective batch size of 32 over 20 epochs and an initial learning rate of 3×10^{-5}. Parameter optimisation relied on the AdamW algorithm, while model selection was guided by validation performance to limit the risk of overfitting during fine-tuning. Translation quality was tracked with BLEU (Papineni et al., 2002), chrF (Popović, 2015) and COMET (Rei et al., 2020), which together offer complementary views of lexical accuracy, character-level correspondence and semantic adequacy.

The experiments were run on GPU-accelerated hardware: an NVIDIA A100 accelerator, a 16-core CPU environment and 32 GB of RAM, within a PyTorch implementation framework. This configuration provided enough stability for multilingual optimisation while keeping the experimental results reproducible.

Table 4 summarises the principal training parameters and the computational resources used during the evaluation of RAÍZ.

Table 4. Training Configuration and Computational Infrastructure of RAÍZ.

Category	Parameter	Configuration
Training Strategy	Optimisation approach	Supervised fine-tuning
	Backbone architecture	ByT5
	Effective batch size	32
	Number of epochs	20
	Initial learning rate	3×10^{-5}
	Optimisation method	AdamW
	Model selection criterion	Best validation performance
Evaluation Framework	Automatic evaluation metrics	BLEU, chrF, COMET
	Evaluation scope	Lexical, character-level, and semantic assessment
Computational Infrastructure	GPU	NVIDIA A100
	CPU environment	16-core processor
	RAM	32 GB
	Deep learning framework	PyTorch

3.5 Baseline Models and Evaluation Metrics

To gauge the contribution of the proposed framework, three widely recognised multilingual architectures were chosen as baselines: mT5, ByT5 and NLLB. mT5 served as a multilingual text-to-text reference; ByT5 allowed us to examine byte-level representations under conditions of orthographic variability; and NLLB offered a comparison against a translation architecture built specifically to widen multilingual coverage for low-resource languages.

Evaluation relied on BLEU, chrF and COMET. BLEU measures n-gram correspondence between generated and reference translations (Papineni et al., 2002); chrF concentrates on character-level similarity and proves especially informative in languages marked by orthographic and morphological variation (Popović, 2015); and COMET complements the two by estimating translation quality through neural evaluation models that generally align more closely with human judgements (Rei et al., 2020). Given the well-documented shortcomings of automatic metrics, the quantitative analysis was supplemented by qualitative assessment involving native speakers.

Table 5 summarises both the automatic evaluation metrics and the qualitative validation procedures adopted in the study. The evaluation framework was deliberately structured to combine quantitative evidence with human-centred linguistic assessment, so that translation quality could be examined from perspectives reaching beyond lexical overlap.

Table 5. Evaluation Metrics and Qualitative Validation Criteria.

Metric / procedure	What it measures	Usefulness in this study	Main limitation
BLEU	Lexical overlap based on n-grams	Standard comparison with previous studies	May penalise valid translations expressed differently
chrF	Character-level overlap	Suitable for morphological and orthographic variation	Does not fully capture cultural meaning
COMET	Neural estimation of semantic quality	Greater sensitivity to fidelity and adequacy	May present limitations for languages with limited coverage
Community evaluation	Human judgement by native speakers	Assesses clarity, naturalness, and appropriateness	Depends on speaker availability, consensus, and dialectal diversity

The strategy set out in Table 5 rests on the conviction that no single metric can represent translation quality in Indigenous language contexts. BLEU, chrF and COMET capture complementary aspects of lexical correspondence, character-level consistency and semantic preservation, while community-based evaluation adds insight into naturalness, cultural appropriateness and communicative authenticity. For this reason, the assessment framework adopted here embraces a multidimensional perspective, one better aligned with the linguistic diversity and cultural complexity of low-resource Indigenous language translation.

4 Results and Discussion

RAÍZ produced consistent quantitative results across the two translation tasks examined in this study. For Spanish–Nahuatl, the framework reached 28.6 BLEU, 56.8 chrF and 74% COMET. For Spanish–Mazahua, the corresponding values were 27.4 BLEU, 54.9 chrF and 71% COMET. Taken together, these figures indicate that the proposed architecture maintains a satisfactory balance between lexical correspondence, character-level consistency and semantic preservation in both language pairs.

Lexical similarity between generated outputs and reference translations was estimated with BLEU, one of the most widely used metrics in machine translation evaluation (Papineni et al., 2002).

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{2}$$

In this formulation, BP denotes the brevity penalty applied when candidate translations are shorter than their references, while p_n corresponds to the modified n-gram precision used to quantify lexical overlap between candidate and reference texts.

Table 6 summarises the performance of RAÍZ on the Spanish–Nahuatl and Spanish–Mazahua tasks according to BLEU, chrF and COMET. Alongside the numerical results, the table offers an interpretative perspective intended to ease the analysis of lexical consistency, semantic preservation and the transferability of the framework across different low-resource linguistic contexts.

Table 6. Extended Quantitative Results of RAÍZ by Language Pair.

Language pair	Model	BLEU	chrF	COMET	Interpretation
Spanish–Nahuatl	RAÍZ	28.6	56.8	74%	Competitive lexical and semantic consistency under low-resource conditions
Spanish–Mazahua	RAÍZ	27.4	54.9	71%	Consistent performance; evidence of methodological transferability

A comparison of the values in Table 6 reveals relatively stable behaviour across both language pairs, despite differences in corpus size and linguistic characteristics. The slightly stronger performance observed for Spanish–Nahuatl probably reflects the larger volume of training data available for that pair. Even so, the Spanish–Mazahua results remain comparatively close, which suggests that the effectiveness of the framework does not depend entirely on increased data availability. This observation matters in Indigenous language settings, where substantial training resources are rarely at hand. The combination of linguistic preprocessing, multilingual semantic representations and community-oriented validation appears to yield a level of robustness that extends beyond conventional data-intensive learning.

To place these findings within the broader context of multilingual machine translation research, Table 7 compares the principal characteristics, strengths and limitations of the baseline architectures considered in this study. The comparison goes beyond numerical performance and also examines how each framework handles orthographic variation, linguistic structure and community participation in low-resource Indigenous language scenarios.

Table 7. Comparative Positioning of RAÍZ and Baseline Multilingual Architectures.

Model	Architectural focus	Principal strength	Principal limitation	Relationship with RAÍZ
mT5	Multilingual text-to-text Transformer	Cross-lingual transfer learning	Limited adaptation to Indigenous linguistic variability	General multilingual baseline
ByT5	Byte-level multilingual Transformer	Robustness against orthographic variation and noisy text	Increased computational cost and absence of explicit sociolinguistic modelling	Principal neural backbone
NLLB	Large-scale multilingual translation architecture	Strong multilingual coverage and semantic modelling	Generalist architecture without integrated community validation	Translation-oriented comparative reference
RAÍZ	Hybrid linguistically informed architecture	Integration of linguistic preprocessing, semantic representation, retrieval refinement, and community validation	Requires specialised corpus preparation and native-speaker participation	Proposed contextual and culturally grounded framework

The comparison makes clear that existing multilingual architectures have made substantial progress in semantic representation and cross-lingual transfer learning. Yet most of these systems continue to operate within predominantly general-purpose training paradigms. RAÍZ, by contrast, treats explicit linguistic preprocessing, contextual lexical refinement and native-speaker validation as core components of the translation process. From this perspective, the contribution of RAÍZ should not be judged solely through quantitative performance. Its value also lies in introducing a hybrid methodological framework designed to support linguistically informed and socially grounded machine translation for Indigenous languages.

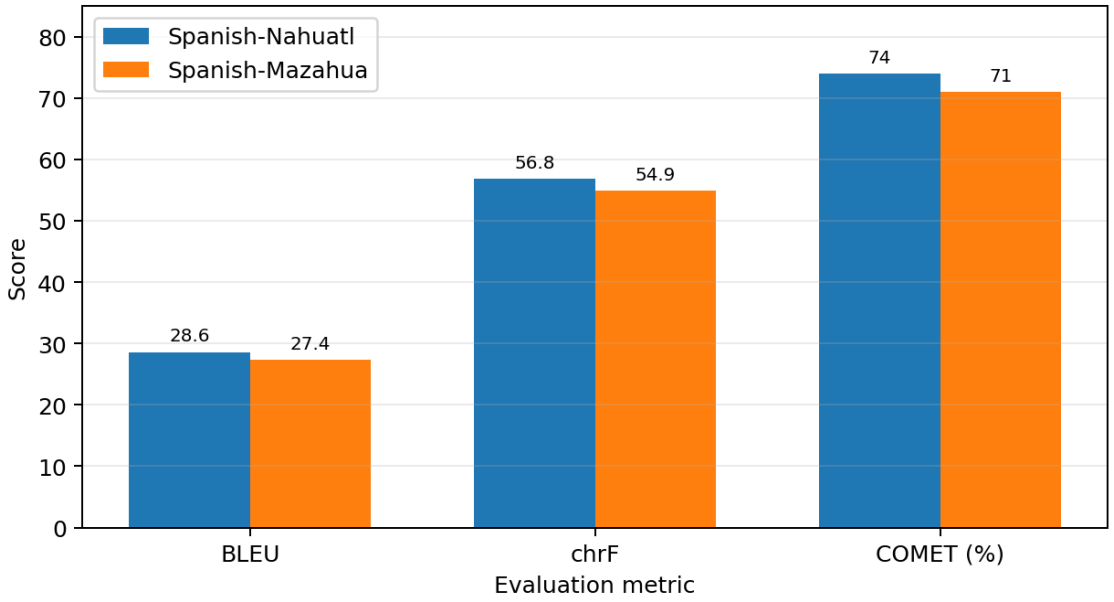

Figure 3. RAÍZ Automatic Evaluation Results.

Figure 3 presents the automatic evaluation results for both language pairs. BLEU, chrF and COMET scores appear along the horizontal axis, while the vertical axis shows the corresponding performance values. The figure reveals a similar performance pattern in both experimental scenarios. Spanish–Nahuatl consistently attains slightly higher scores, probably reflecting the larger corpus, but Spanish–Mazahua follows the same general trend despite having fewer parallel sentence pairs.

The relationship between BLEU and chrF is especially informative. The results suggest that the framework is not simply reproducing lexical sequences found in the training data; it is also preserving relevant information at the level of characters and subword structures. Such behaviour matters greatly in Indigenous language contexts, where orthographic variation and morphological complexity frequently disrupt exact word-level matching. COMET scores add further evidence that the generated translations retain a substantial portion of the source-text meaning. These results should nonetheless be read with caution, because neural evaluation metrics remain relatively underexplored for many Indigenous languages. Dialectal diversity, spelling variation and culturally specific semantic structures are often under-represented in the multilingual datasets used to train large-scale evaluation models (Freitag et al., 2022).

While automatic metrics offer useful quantitative indicators of translation quality, they do not fully capture contextual appropriateness or culturally grounded language use. For this reason, qualitative evaluation with native speakers formed an essential part of the assessment strategy adopted in this study.

The qualitative review complemented the automatic metrics by examining dimensions such as clarity, naturalness, semantic fidelity and cultural appropriateness. Feedback gathered during this stage showed that translations achieving acceptable automatic scores can still carry problems related to register selection, lexical preference, cultural nuance or community-level naturalness. These observations reinforce the importance of building community participation directly into the evaluation process, rather than treating it as a secondary validation stage.

Table 8 summarises the criteria used during the community-based assessment conducted with native speakers. These criteria were chosen to complement automatic metrics by introducing dimensions tied to intelligibility, linguistic authenticity, semantic preservation and culturally situated language practices.

Table 8. Qualitative Validation Criteria Used with Native Speakers.

Criterion	Definition	Validation procedure	Contribution
Clarity	The translation can be understood by speakers	General intelligibility review	Prevents grammatically opaque outputs
Naturalness	The sentence sounds authentic within the target language	Judgement by native speakers	Detects literal or artificial translations
Semantic fidelity	Preserves the meaning of the source text	Source–output comparison	Reduces semantic loss or unintended additions
Cultural appropriateness	Respects community usage, registers, and contextual norms	Contextual review	Prevents technically correct but socially inappropriate equivalences

The validation framework set out in Table 8 illustrates that translation quality in Indigenous language contexts extends well beyond grammatical correctness and lexical similarity. Expressions that look acceptable according to automatic metrics may still strike members of the target-language community as unnatural, culturally inappropriate or semantically imprecise. Native-speaker participation therefore adds a crucial interpretative layer, one that strengthens both the linguistic reliability and the social relevance of the proposed translation framework.

5 Conclusions

This study introduced RAÍZ, a hybrid machine translation framework developed for low-resource Indigenous languages. The experimental results indicate that combining phonetic normalisation, morphological segmentation, multilingual semantic representations, community-oriented lexical memory and native-speaker validation can contribute positively to translation quality in resource-constrained settings. The evidence gathered throughout the evaluation suggests that these components work in a complementary manner, offering advantages beyond those typically provided by general-purpose multilingual architectures alone.

The contribution of RAÍZ is not confined to the quantitative results reported here. Equally important is the methodological perspective underpinning the framework, which links corpus development, dataset traceability, experimental design, baseline comparison, automatic evaluation and community-based assessment within a single research workflow. Such integration is especially valuable in Indigenous language contexts, where the effectiveness of a language technology cannot be judged exclusively through numerical performance indicators. Practical usefulness, linguistic acceptability and cultural appropriateness remain equally important dimensions of evaluation.

The study also brought several limitations to light. Although the corpus comprises 42,000 parallel sentence pairs, differences in data availability between language pairs remain evident. Dialectal variation poses an additional challenge, since lexical, phonetic and morphosyntactic differences can occur across regional varieties of the same language and demand dedicated validation procedures. Limitations associated with automatic evaluation metrics were likewise observed, particularly in languages that remain under-represented in existing evaluation frameworks. Another practical constraint concerns the availability of native speakers for large-scale and sustained validation activities. Moreover, extending the framework to additional Indigenous languages will require new corpora, language-specific adaptations and continued collaboration with local communities.

Table 9 summarises the principal limitations identified during the development and evaluation of RAÍZ, together with the research directions that emerge from these observations. Rather than viewing these issues solely as constraints, they can also be understood as opportunities for the continued refinement of language technologies designed for low-resource Indigenous language environments.

Table 9. Identified Limitations and Future Research Perspectives.

Detected limitation	Effect on the study	Future research line
Corpus size	The corpus is useful but insufficient to cover all usage domains	Expand parallel datasets and thematic domains
Language balance	Nahuatl contains more sentence pairs than Mazahua	Develop comparable subsets and report results according to sample size
Dialectal variation	Not all linguistic variants are represented	Document variants, regions, and normalisation criteria
Automatic metrics	BLEU, chrF, and COMET do not fully capture cultural appropriateness	Combine automatic metrics with human and community evaluation
Community validation	Depends on speaker availability and linguistic consensus	Expand reviewer panels and inter-rater agreement protocols
Transferability	The model requires adaptation before application to other languages	Evaluate additional Indigenous languages and alternative baseline models

The synthesis in Table 9 highlights that the challenges surrounding Indigenous language machine translation extend well beyond computational considerations. Linguistic diversity, corpus development, evaluation practices and community engagement are closely interconnected factors that influence both translation quality and the long-term sustainability of technological initiatives. Future work should therefore address not only architectural improvements, but also collaborative approaches to data governance, culturally informed validation practices and the creation of reusable linguistic resources that support language preservation and revitalisation.

Future developments should be seen as a continuation of the research trajectory initiated in this study, rather than as a separate stage that follows its completion. Expanding the corpus to additional domains, increasing dialectal coverage, incorporating further Indigenous languages, evaluating emerging multilingual architectures, broadening the participation of speakers from different regions and producing reusable educational and community resources all represent promising directions for subsequent work. These priorities offer opportunities to improve scalability, linguistic coverage, explainability and the long-term deployment of community-oriented translation technologies.

The methodological principles underpinning RAÍZ may also be adapted to other low-resource languages and alternative translation architectures, provided that traceability, linguistic validation, community participation and ethical data stewardship remain central to the development process. Under these conditions, the framework has the potential to serve not only Nahuatl and Mazahua, but also to act as a transferable methodological foundation for culturally grounded, interpretable and inclusive language technologies.

6 Data Availability, Corpus Registration, and Ethical Considerations

The corpus used in this research was built through a structured process that included data selection, cleaning, alignment, linguistic revision and quality verification. The resulting datasets and corpora have been documented as linguistic resources registered with the Instituto Nacional de los Pueblos Indígenas (INPI). This registration supports the traceability of the resource, acknowledges its linguistic and cultural significance, and reinforces the need to manage it under appropriate ethical principles. Decisions about public or restricted access should therefore be guided by community agreements, institutional requirements, authorised permissions and the protection of culturally sensitive information.

Native-speaker participation formed a central part of the evaluation strategy adopted in this study. Relying on automatic metrics alone would have provided only a partial assessment of translation quality. Within Indigenous language research,

community involvement should be seen not merely as a consultative activity, but as an essential source of linguistic, semantic and cultural expertise that contributes directly to the quality and relevance of the resulting language technology.

References

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. O'Reilly Media.

De Gibert, O., Pugh, R., Marashian, A., Vazquez, R., Ebrahimi, A., Denisov, P., Rice, E., Gow-Smith, E., Prieto, J., Robles, M., Manrique, R., Moreno, O., Lino, A., Coto-Solano, R., Alvarez, A., Agüero-Torales, M., Ortega, J. E., Chiruzzo, L., Oncevay, A., . . . Mager, M. (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 *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)* (pp. 134–152). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.americasnlp-1.16

Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1702.08608

Freitag, M., Rei, R., Mathur, N., Lo, C.-K., Stewart, C., Avramidis, E., Kocmi, T., Foster, G., Lavie, A., & Martins, A. F. T. (2022). Results of WMT22 metrics shared task: Stop using BLEU—Neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)* (pp. 46–68). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.wmt-1.2

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing* (3rd ed. draft, January 7, 2023) [Draft manuscript]. https://web.stanford.edu/~jurafsky/slp3/old_jan23/ed3book_jan72023.pdf

Mager, M., Oncevay, A., Ebrahimi, A., Ortega, J., Rios, A., Fan, A., Gutierrez-Vasques, X., Chiruzzo, L., Giménez-Lugo, G., Ramos, R., Meza Ruiz, I. V., Coto-Solano, R., Palmer, A., Mager-Hois, E., Chaudhary, V., Neubig, G., Vu, N. T., & Kann, K. (2021). Findings of the AmericasNLP 2021 shared task on open machine translation for Indigenous languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas* (pp. 202–217). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.americasnlp-1.23

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

NLLB Team, Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Mejia Gonzalez, G., Hansanti, P., . . . Wang, J. (2022). *No language left behind: Scaling human-centered machine translation* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2207.04672

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

Popović, M. (2015). chrF: Character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation* (pp. 392–395). Association for Computational Linguistics. https://doi.org/10.18653/v1/W15-3049

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research, 21*(140), 1–67.

Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 2685–2702). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.213

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1715–1725). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1162

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008.

Xue, L., Barua, A., Constant, N., Al-Rfou, R., Narang, S., Kale, M., Roberts, A., & Raffel, C. (2022). ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics, 10*, 291–306. https://doi.org/10.1162/tacl_a_00461

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2021). mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 483–498). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.41