Entramado
ISSN: 1900-3803, e-ISSN: 2539-0279
Volumen 22, número 1, enero-junio 2026

Editorial Universidad Libre

Intérprete contextual de expresiones en lengua Nasa Yuwe a español usando Técnicas de Procesamiento de Lenguaje Natural

https://doi.org/10.18041/1900-3803/entramado.1.13152

Darcyn Frynet Embus-Quina
Corporación Universitaria Comfacauca - Unicomfacauca, Popayán, Colombia

Andrés Alexis Elago-Quitumbo
Corporación Universitaria Comfacauca - Unicomfacauca, Popayán, Colombia

Angela María Rodríguez-Vivas
Corporación Universitaria Comfacauca - Unicomfacauca, Popayán, Colombia

Resumen

El pueblo indígena Nasa Páez, ubicado principalmente en el departamento del Cauca, Colombia, enfrenta un riesgo cultural debido a la pérdida progresiva de su lengua ancestral, el Nasa Yuwe, especialmente entre la población joven. Dado que el idioma constituye el núcleo de su cosmovisión, tradición oral y vida comunitaria, este estudio se apoya en técnicas de Procesamiento de Lenguaje Natural con el fin de proponer herramientas tecnológicas que ayuden a preservar el Nasa Yuwe. Si bien, idiomas occidentales como el inglés, español o francés cuentan con numerosas soluciones de traducción automática, implementarlas con el Nasa Yuwe es un gran reto toda vez que esta es una lengua de ultra (o extremadamente) bajos recursos digitales en comparación con las mencionadas. En aras de mantener coherencia en la recopilación y sistematización de frases en Nasa Yuwe, la investigación se delimita al resguardo indígena San Lorenzo de Caldono, con el fin de trabajar sobre una variante lingüística específica. Para evaluar el intérprete desarrollado, se empleó la matriz de confusión de intenciones, la cual arrojó una eficiencia del 100%, medida mediante el F1-score, y una precisión del 100% en el reconocimiento de expresiones culturales, sobre un total de 144 muestras evaluadas. De manera complementaria, la evaluación de la extracción de entidades culturales, realizada mediante la matriz de confusión generada por el modelo DIETClassifier, evidenció una precisión promedio (macro) del 91,1%, un recall del 95,3% y una eficiencia global del 92,8%, considerando 955 entidades anotadas, lo que confirma el adecuado desempeño del sistema en la identificación de componentes semánticos propios de la lengua Nasa Yuwe.

Palabras clave

Lenguas indígenas; preservación; Procesamiento de Lenguaje Natural; traducción automática; corpus bilingüe; Nasa Yuwe; diacríticos.

Registro

Artículo de investigación
Recibido: 19/09/2025
Aceptado: 18/12/2025
Publicado: 21/01/2026

Contextual Interpretation of Nasa Yuwe Language Expressions into Spanish Using Natural Language Processing Techniques

Abstract

The Nasa Páez Indigenous community, located mainly in the department of Cauca, Colombia, faces a cultural risk due to the progressive loss of its ancestral language, Nasa Yuwe, particularly among younger generations. As language constitutes the core of their worldview, oral tradition, and community life, this study draws on Natural Language Processing techniques to propose technological tools aimed at preserving Nasa Yuwe. While Western languages such as English, Spanish, and French benefit from numerous automatic translation solutions, implementing such technologies for Nasa Yuwe represents a significant challenge, as it is an ultra–low-resource language in digital terms when compared to those languages. To ensure coherence in the collection and systematization of Nasa Yuwe expressions, this research is delimited to the San Lorenzo de Caldono Indigenous Reserve, allowing the study to focus on a specific linguistic variant. To evaluate the developed interpreter, an intention confusion matrix was employed, yielding an efficiency of 100%, measured through the F1-score, and a precision of 100% in the recognition of cultural expressions, based on a total of 144 evaluated samples. Additionally, the evaluation of cultural entity extraction, conducted using the confusion matrix generated by the DIETClassifier model, showed an average (macro) precision of 91.1%, a recall of 95.3%, and an overall efficiency of 92.8% over 955 annotated entities, confirming the system's effective performance in identifying semantic components inherent to the Nasa Yuwe language.

Keywords

Indigenous languages; preservation; Natural Language Processing; machine translation; bilingual corpus; Nasa Yuwe; diacritics.

Cómo citar este artículo

EMBUS-QUINA, Darcyn Frynet; ELAGO-QUITUMBO, Andrés Alexis; RODRÍGUEZ-VIVAS, Angela María. Intérprete contextual de expresiones en lengua Nasa Yuwe a español usando Técnicas de Procesamiento de Lenguaje Natural. En: Entramado. Enero - junio, 2026. vol. 22, no. 1. p. 1-21. e-13027 https://doi.org/10.18041/1900-3803/entramado.1.13152

1. Introducción

Colombia es el tercer país con mayor diversidad lingüística en Latinoamérica, con 65 lenguas indígenas reconocidas, muchas de ellas en riesgo de desaparición debido al debilitamiento de la transmisión intergeneracional (Llerena y Díaz, 2023). Por su parte, el Consejo Regional Indígena del Cauca (CRIC) representa a ocho pueblos indígenas de Colombia cuyas lenguas ancestrales igualmente están en riesgo de extinción: Nasa – Páez, Guambiano Yanaconas, Coconucos, Epiraras – siapiraras (Emberas), Totoroes, Inganos y Guanacos (CRIC, 2024).

Teniendo en cuenta que las lenguas indígenas poseen una gran riqueza en diversidad y cultura lingüística, estudios en Colombia y Latinoamérica proponen la implementación de políticas públicas efectivas que incluyan reconocimiento legal, promoción de la educación bilingüe (Gutiérrez Arríaga; Alvarado Rodríguez, 2024), mecanismos que fomenten el uso activo de las lenguas indígenas en diferentes ámbitos (Uribe-Muñoz, 2025) y un plan decenal que las revitalice y documente (Ministerio de Cultura de Colombia, 2020). Así mismo, integrar recursos tecnológicos que acompañen los procesos comunitarios de fortalecimiento lingüístico es un llamado inminente (Bautista, Martínez, Rocha y Montes, 2024).

Al respecto, una alternativa prometedora es el aprovechamiento de las técnicas de Procesamiento de Lenguaje Natural (NLP, por sus siglas en inglés), una rama de la Inteligencia Artificial que ha demostrado impacto positivo en la preservación de medios lingüísticos (AmericasNLP, 2026). En el campo del NLP, lenguas de bajos recursos son aquellas que carecen de material digital escrito u oral. En estos casos, la aplicación de modelos de traducción basados en arquitecturas neuronales, es un gran reto ya que estos modelos asumen la existencia de volúmenes de datos generosos (Salazar-Cárdenas, 2022).

Particularmente, la lengua Nasa Yuwe, que constituye el eje de la identidad, cosmovisión y vida comunitaria del pueblo Nasa, atraviesa un momento crítico (Urbina, 2018) toda vez que la influencia de la cultura occidental y la atracción que ejercen las oportunidades de las grandes ciudades sobre los jóvenes han llevado a que muchos de ellos priorizan el aprendizaje del castellano. Según las proyecciones del Departamento Administrativo Nacional de Estadística -DANE-, para el año 2018, se reportaron 243.176 ciudadanos reconocidos como pertenecientes al pueblo Nasa (DANE, 2019). Esta cifra hace que la lengua Nasa Yuwe se constituya en una lengua de ultra (o extremadamente) bajos recursos en comparación con lenguas occidentales como el español o el inglés.

En el año 2021 nació una iniciativa de reunir a expertos e interesados en la aplicación de NLP a lenguas indígenas de las Américas (AmericasNLP, 2026), mostrando hasta el momento importantes avances principalmente en la aplicación de traducción de máquina (MT, por sus siglas en inglés). Sin embargo, el Nasa Yuwe no ha sido foco de atención en ninguna de las ediciones de este evento. En Colombia se han adelantado investigaciones en torno a creación de un corpus en Nasa Yuwe, sin embargo, algunas de estas investigaciones aún no aplicaron NLP en su solución (Sierra et al., 2015; Sierra, Cobos y Corrales 2016; Sierra et al., 2018; Sierra et al., 2019; Solano, Tobar, Sierra y Cobos (2020); Muñoz et al., 2023; Salazar-Cárdenas, 2022) o crearon un corpus a partir de fuentes estáticas (Salazar-Cárdenas, 2022) que no necesariamente reflejan la cosmovisión del hablante indígena al expresar su punto de vista respecto a un contexto específico.

La presente investigación propone el diseño de un intérprete de expresiones del Nasa Yuwe a español haciendo uso de técnicas de NLP. Este intérprete busca realizar traducciones conscientes del contexto, es decir que tienen en cuenta la cosmovisión del hablante indígena en cada frase expresada más allá de traducir de forma meramente literal. Para lograr este objetivo, las frases del corpus fueron recolectadas directamente en territorio Nasa y etiquetadas de forma manual para preservar el sentido de expresiones relacionadas con el territorio. Es importante señalar que el Nasa Yuwe presenta variaciones dialectales libres según la región con diferencias en la pronunciación. Por esta razón, el proyecto delimita su alcance al resguardo indígena San Lorenzo de Caldono, en el departamento del Cauca, con el fin de trabajar sobre

una variante lingüística específica. Esta delimitación permite alcanzar coherencia en la recopilación de datos y la validación de expresiones evitando confusiones derivadas de la heterogeneidad dialectal.

Las contribuciones del presente estudio son:

1. Un corpus de frases en Nasa Yuwe referentes a debates políticos recolectadas directamente con miembros de la comunidad Nasa-Páez.
2. Etiquetado de las frases de acuerdo a entidades relevantes para contextos de debate.
3. Una arquitectura que involucra conceptos del NLP para el manejo de particularidades de las expresiones, e.g., acentos y comillas.
4. Una prueba de concepto con la herramienta Rasa NLU.

Para evaluar el intérprete se empleó la matriz de confusión, la cual arrojó una eficiencia del 100% y una precisión del 100% en la clasificación de intenciones, y una eficiencia del 92,8% junto con una precisión del 91,1% en la extracción de entidades, en el reconocimiento de expresiones culturales. Estos resultados permiten validar la arquitectura propuesta como una aproximación para aplicación de técnicas de NLP a traducciones de lenguas de ultra bajos recursos conscientes del contexto al tiempo que se reconoce su riqueza fonética y cultural.

Las secciones restantes del artículo se estructuran de la siguiente forma. Un marco de referencia en el cual se definen conceptos fundamentales de la investigación y se describen los trabajos relacionados relevantes al tema. Una sección de materiales y métodos que detalla el proceso llevado a cabo para la implementación del prototipo funcional. Resultados y discusión mostrando cómo se efectuó la evaluación del mismo. Finalmente, se presentan las conclusiones del trabajo.

2. Marco de referencia

Se presenta un marco teórico que se centra en explicar particularidades fundamentales de la lengua Nasa Yuwe, palabras o conceptos claves de NLP y los trabajos relacionados que se toman como base para el presente proyecto.

2.1. Marco teórico

2.1.1 Resguardo indígena San Lorenzo - Caldono

El municipio de Caldono se ubica al nor-occidente del departamento del Cauca en Colombia, tal como se muestra en la Figura 1A. La división política administrativa del municipio, está conformado por 86 veredas, 4 corregimientos (Cerro Alto, Siberia, Pescador y Pital) 6 resguardos indígenas (San Lorenzo de Caldono, Pioyá, Pueblo Nuevo, La Aguada, La Laguna y Las Mercedes), y un cabildo Misak.

El municipio es un territorio multiétnico y pluricultural, está habitado por comunidades en su mayoría por indígenas de la etnia Nasa, Misak, población mestiza y una minoría de población afro. De la población indígena, según los listados censales a febrero de 2020, se destaca al resguardo San Lorenzo de Caldono (Figura 1B), como el de mayor población con el 44% (13.641) del total de población indígena (Municipio de Caldono, 2023).

Un antecedente relevante en los procesos de fortalecimiento lingüístico en San Lorenzo - Caldono lo constituyen las iniciativas pedagógicas impulsadas por orientadores y estudiantes de la escuela de Nasa Yuwe del resguardo de Caldono, Cauca. Con propósito de revitalizar el idioma materno, se han elaborado materiales didácticos que proponen temas y ejercicios básicos orientados a la práctica de la oralidad y la lectoescritura. La Figura 2 muestra un ejemplar del libro “Aprendamos a Hablar y Escribir el Nasa Yuwe” elaborado comunitariamente con propósitos académicos en el resguardo. Se desconoce la fecha exacta de su elaboración.

Figura 1A. División política del Cauca
Fuente: Plan de desarrollo departamental 2026-2019

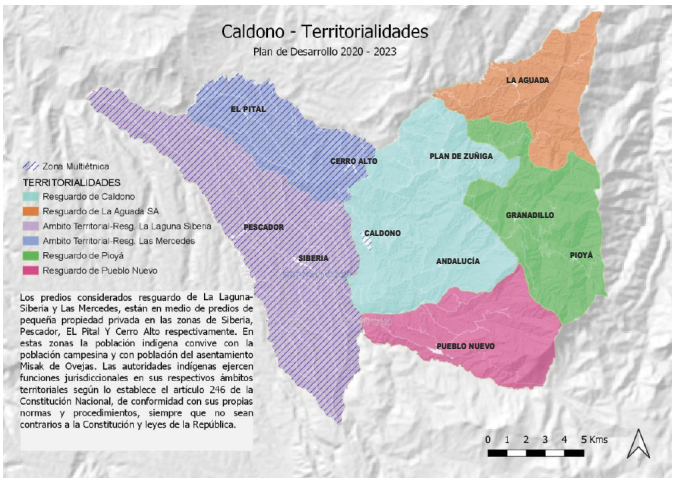

Figura 1B. Territorialidades de Caldono
Fuente: Plan de desarrollo municipal 2020-2023

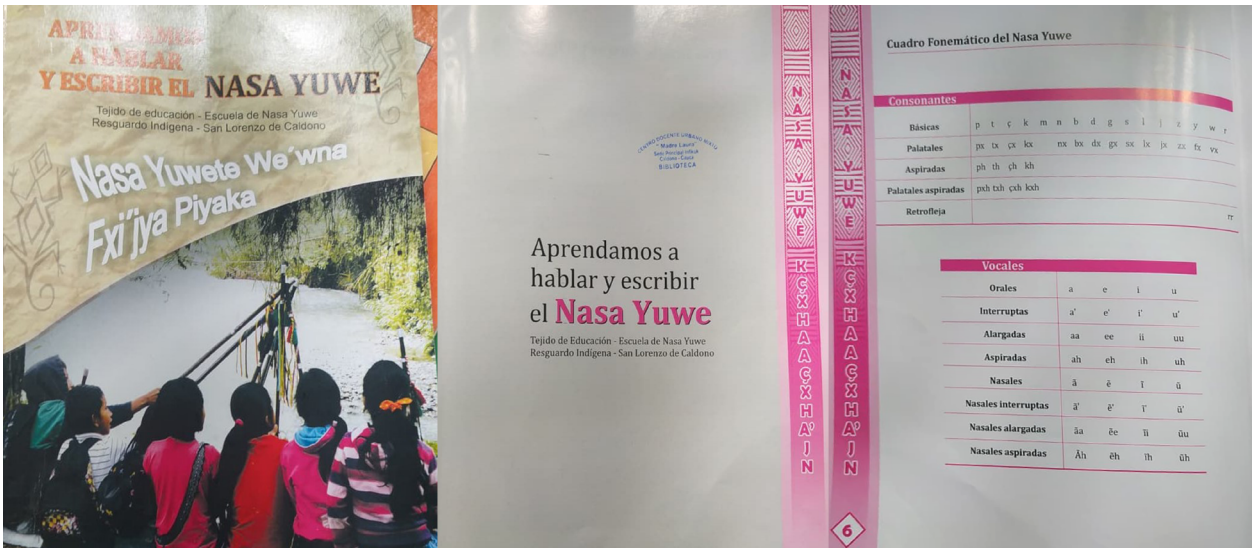

Figura 2. Cartilla "Aprendamos a Hablar y Escribir el Nasa Yuwe"
Fuente: Centro Docente Urbano "Madre Laura", Caldono – Cauca.

Otro ejemplo de recurso lingüístico producto de un esfuerzo académico-colectivo en la Institución Educativa Renacer Páez Pitayó, municipio de Silvia - Cauca se muestra en la Figura 3. Dichos recursos apoyan el reconocimiento del alfabeto Nasa Yuwe trabajado y unificado entre el Consejo Regional Indígena del Cauca (CRIC), la prefectura de Tierradentro y la fundación Juan Tama de Pitayó, el cual hoy en día es utilizado y enseñado en los territorios Nasa.

Figura 3. Cartilla “Trenzando ideas para la construcción de un Pensamiento Nasa”.
Fuente: Institución Educativa Renacer Paez Pitayó, Silvia – Cauca.

2.1.2. Alfabeto Nasa Yuwe

Las letras que conforman el alfabeto del Nasa Yuwe están organizadas por consonantes y vocales que se fundamentan en su fonología (funcionamiento de las letras). La Tabla 1 explica el diagrama del plan etnoeducativo intercultural, diseñado teniendo en cuenta el alfabeto latino (tomado del libro mostrado en la Figura 3).

Tabla 1.
Plan de Vida Etnoeducativo Intercultural teniendo en cuenta el alfabeto latino

El mundo Nasa	Objeto Compartido
Pensamiento idioma usos y costumbres	a, b, ch, d, e, f, g, i, j, k, l, ll, m, n, ñ, o, p, q, r, rr, s, t, u, v, w, x, y, z

Fuente: Elaboración propia, tomando como fuente de información el libro físico titulado: Trenzando ideas para la construcción de un pensamiento Nasa pág. 9

2.1.3. Fonemático del Nasa Yuwe - Tabla 2

Tabla 2.
Consonantes

Básicas	p t ç k m n b d g s l j z y w r
Palatales	px tx çx kx nx bx dx gx sx lx jx zx fx vx
Aspiradas	ph th çh kh
Palatales Aspiradas	pxh txh çxh kxh
Retrofleja	rr

Fuente: Elaboración propia, tomando como fuente de información el libro físico titulado: Aprendamos a hablar y escribir el Nasa Yuwe pág. 6

1. Consonantes básicos: Son aquellos fonemas de la lengua que no tienen ninguna modificación fonológica.
2. Consonantes palatales: Son aquellos sonidos que al pronunciar el dorso de la lengua se acerca al paladar, se pronuncia como silbido, su símbolo es la (x).
3. Consonantes aspiradas: son los que se pronuncian soplados, su símbolo es la (h).
4. Consonante palatal aspiradas: Fonemas que al pronunciar reúnen los dos símbolos o características silbido y soplado (x,h).

En la Tabla 3, se observan los diacríticos, los cuales son signos o letras adicionales utilizadas para escribir las lenguas en el mundo (libro Figura 2). A continuación, se observan los dos tipos de diacríticos saltillos y nasal que se utilizan para escribir la lengua Nasa Yuwe.

Tabla 3.
Vocales

Vocal libre	interrupta	Nasal	Alargadas	Aspiradas
a	a’	ã	aa	ah
e	e’	ẽ	ee	Eh
i	i’	ĩ	ii	Ih
u	u’	ũ	uu	uh

Fuente: Elaboración propia, tomado del libro físico mostrado en la Figura 2, pág. 6.

1. Diacrítico saltillo: El saltillo es una pequeña línea vertical o tilde, que se escribe después de cada vocal, indicando el corte del sonido. (a’te, le’chkue, ki’sen, u’se).
2. Diacrítico nasal: Es la tilde sobre la n para transformarla en ñ, lo cual indica la nasalización de la vocal. (tãsh, ẽs, ĩish, ũsku).

2.1.4. Procesamiento de Lenguaje Natural

El NLP es una rama de la inteligencia artificial que estudia la interacción entre los sistemas computacionales y el lenguaje humano, denominado lenguaje natural. Esta disciplina se orienta al desarrollo de métodos y algoritmos que permiten a las máquinas analizar, comprender e interpretar el lenguaje natural. De acuerdo con Rongali (2025), el NLP integra técnicas de aprendizaje automático, modelos estadísticos y conocimientos lingüísticos para facilitar tareas como la traducción automática, el reconocimiento de voz y el análisis de textos. Así, el NLP se ha considerado como una herramienta fundamental para el desarrollo de aplicaciones inteligentes que optimizan la comunicación entre los seres humanos y las tecnologías digitales.

2.1.5. Tokenización en NLP

La *tokenización* es un proceso fundamental en el campo del NLP, ya que consiste en desfragmentar o dividir un texto en unidades denominadas *tokens* (Bayram et al., 2025). Estas unidades pueden ser palabras, sílabas, caracteres o signos de puntuación (Jurafsky y Martin, 2025). En lenguas de amplia difusión como el español o inglés, la *tokenización* suele hacerse separando las palabras por espacios en blanco y signos de puntuación.

En el caso de lenguas indígenas como el Nasa Yuwe, la tokenización presenta desafíos particulares debido a la presencia de diacríticos (ã, ẽ, ĩ, ũ,) y fonemas propios (pxh, txh, çxh, etc.), los cuales no deben considerarse como unidades divisibles para no distorsionar el significado. Según Bayram et al. (2025), una tokenización inadecuada en lenguas con estructuras fonéticas y morfológicas complejas compromete la integridad lingüística del texto y reduce el desempeño de los modelos de NLP, afectando directamente los resultados obtenidos.

Por ejemplo, la frase en Nasa Yuwe “kwe’sx uma kiwe îtxi” se puede *tokenizar* como: [“kwe’sx”, “uma”, “kiwe”, “îtxi”]. De este modo, el modelo de NLP puede identificar cada palabra de manera independiente y analizar su función en el contexto para su posterior análisis.

2.1.6. Etiquetado

El etiquetado es el proceso mediante el cual se vinculan expresiones, i.e., palabras relevantes o fragmentos exactos que aparecen en una oración, con entidades. Las entidades representan categorías abstractas de información relevantes en una oración (Jehangir et al., 2023). Por ejemplo, en la Figura 4, la frase “Quiero una pizza grande con extra de queso”, la expresión “pizza” se etiqueta como la entidad Producto, “grande” como la entidad Tamaño, y “extra queso” como la entidad Extra. De esta manera, el etiquetado permite transformar un texto en datos estructurados, facilitando que un sistema de NLP comprenda no solo lo que se dice, sino también el significado de cada parte en su contexto.

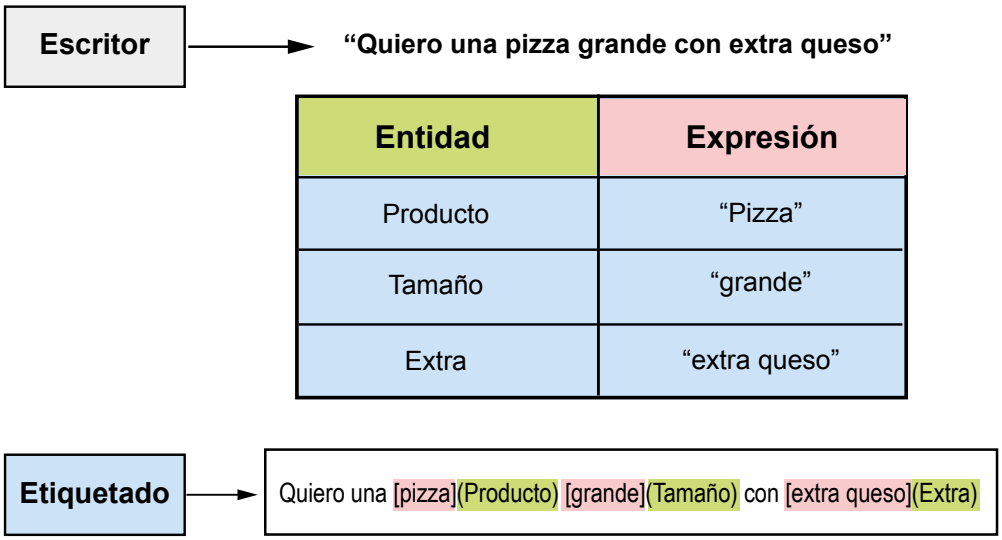

Figura 4. Ejemplo de Etiquetado

Fuente: El ejemplo fue realizado por los autores en la herramienta drawio.

2.1.7. Intenciones

Una intención representa el propósito subyacente de un usuario al enviar un mensaje a su interlocutor. En el contexto del NLP, dicho interlocutor es un asistente conversacional soportado por Inteligencia Artificial (Jurafsky y Martin, 2025). Cada mensaje que ingresa un usuario al asistente se clasifica en una intención específica para que el sistema pueda entender qué quiere lograr o comunicar el usuario. Por ejemplo, en la Figura 4, el usuario escribe “Quiero una pizza grande con extra de queso”, su intención podría denominarse pedido_pizza.

2.2. Trabajos relacionados

2.2.1 Metodologías de trabajo con comunidades indígenas en proyectos tecnológicos.

Las metodologías de trabajo con comunidades indígenas en proyectos tecnológicos se fundamentan en enfoques participativos, interculturales y éticos, reconocen a las comunidades indígenas como sujetos principales y activos en la construcción colectiva del conocimiento. Desde una perspectiva ética y de derechos humanos, la participación activa de los pueblos indígenas en todas las etapas del desarrollo

de sistemas de inteligencia artificial en los modelos de traducción automática, se articula mediante metodologías participativas e interculturales reconociendolos como co-creadores del conocimiento, lo que contribuye a reducir los sesgos algorítmicos y a evitar procesos de exclusión cultural (UNESCO, 2023).

El uso de metodologías de diseño centradas en las personas, como *Design Thinking, Design Sprint* y la técnica *How Might We*, ha permitido promover la participación activa de jóvenes indígenas en la co-creación de soluciones tecnológicas orientadas a reducir barreras lingüísticas y mejorar el acceso a la información en contextos de baja conectividad (Lozano, Carvajal y Álvarez, 2025).

Por su parte, los estudios sobre lenguas indígenas de bajos recursos evidencian que la escasez de datos lingüísticos y el reducido número de hablantes nativos limitan el desarrollo de modelos de traducción automática y otras herramientas basadas en inteligencia artificial. Para enfrentar estas limitaciones, se han empleado metodologías experimentales y comparativas orientadas al diseño y evaluación de modelos de lenguaje especializados (Parankusham, Rizk y KC, 2025).

De manera complementaria, investigaciones basadas en metodologías cualitativas han analizado la percepción y participación comunitaria en el uso de tecnologías de inteligencia artificial aplicadas a la preservación lingüística, evidenciando la importancia de integrar enfoques culturalmente sensibles y éticos para garantizar la aceptación y sostenibilidad de estas herramientas (Soylu y Şahin, 2024). Asimismo, estudios de revisión en contextos latinoamericanos coinciden en que la efectividad de las tecnologías del lenguaje para lenguas indígenas depende de su articulación con políticas públicas, estrategias de difusión y acciones orientadas a reducir la brecha tecnológica y prevenir procesos de exclusión cultural (Aguilar y García, 2023).

2.2.2 NLP aplicado a lenguas indígenas

La principal iniciativa de aplicaciones de NLP a lenguas indígenas en peligro de desaparición en América es sin duda el Workshop Americas NLP (AmericasNLP, 2026). Una de las contribuciones más importantes del evento ha sido la organización de shared tasks competitivos, que ha sido un punto de confluencia de investigadores y estudiantes interesados en el tema. Entre las lenguas recurrentes trabajadas en el Workshop están el Guaraní, Quechua, Bribri y Aymara.

Uno de los ejes centrales del workshop ha sido el MT, en particular en escenarios español ↔ lenguas indígenas. En este contexto, se han explorado técnicas de traducción automática neuronal (NMT), incluyendo modelos sequence-to-sequence con atención, arquitecturas Transformer y enfoques multilingües. Dado el bajo volumen de datos paralelos, una contribución recurrente ha sido el uso de transfer learning, donde modelos preentrenados en lenguas de mayor recurso se adaptan a lenguas indígenas mediante fine-tuning. Asimismo, se han evaluado estrategias como el aprendizaje multilingüe y el entrenamiento conjunto de lenguas tipológicamente relacionadas para mejorar la generalización.

La métrica de evaluación más utilizada ha sido BLEU, complementada en algunas ediciones por ChrF, COMET (Nagoudi, Chen, Abdul-Mageed y Cavusoglu, 2021) y evaluaciones humanas. De manera general, los sistemas participantes alcanzan valores de BLEU bajos a moderados en comparación con lenguas de alto recurso. Los resultados evidencian que lenguas con mayor disponibilidad de datos (como Guaraní o Quechua) alcanzan puntajes superiores y menor varianza entre sistemas, mientras que lenguas con datasets extremadamente pequeños presentan mayor dispersión de resultados y una dependencia más fuerte del enfoque utilizado.

Por otro lado, diversos estudios se han centrado en modelos basados en redes neuronales, como la arquitectura transformer propuesta por Bautista Morales et al. (2024) para traducir lenguas indígenas con estructura Escasos Recursos Lingüísticos (ERL). Por su parte, Nagoudi et al. (2021) sugieren el uso de transfer learning y back-translation, aunque la baja participación de hablantes nativos y la disponibilidad limitada de datos siguen siendo desafíos significativos.

El artículo documentado en Kann et al.,(2022) presenta AmericasNLI, un conjunto de datos paralelo de Natural Language Inference (NLI) para 10 lenguas indígenas de las Américas (Asháninka, Aymara, Bribri, Guaraní, Nahuatl, Otomí, Quechua, Rarámuri, Shipibo-Konibo y Wixarika). El estudio evalúa la capacidad de modelos multilingües preentrenados para realizar clasificación de inferencia en estas lenguas sin entrenamiento directo. Los resultados demuestran que los modelos preentrenados sin adaptación tienen un desempeño pobre, pero la adaptación continua y los métodos basados en traducción mejoran la capacidad de comprender textos con estructuras semánticas complejas en lenguas con muy pocos datos.

Downey, Etxeberria, Müller, y Cotterell (2021) muestran cómo un modelo de segmentación secuencial no supervisado multilingüe puede transferirse con éxito a lenguas extremadamente pobres en datos, como K'iche' (una lengua maya). Comparando con baselines monolingües y modelos entrenados solo en una lengua relacionada (por ejemplo Quechua), los autores encuentran que el preentrenamiento multilingüe mejora consistentemente la calidad de segmentación (≈20.6 F1 en algunos ajustes), lo cual tiene implicaciones profundas para la preparación de pipelines de NLP cuando los datos de segmento son muy escasos.

2.2.3 Proyectos con Nasa Yuwe

Los trabajos de Sierra et al., (2015), Sierra, Cobos y Corrales (2016), Sierra et al., (2018) y Sierra et al., (2019) inician con la construcción de una colección de pruebas lingüísticas en Nasa Yuwe (Sierra et al., 2015) y progresa hacia el diseño de un tokenizador específicamente adaptado a la morfología y variación ortográfica de la lengua (Sierra, Cobos y Corrales, 2016). Posteriormente, el trabajo se amplía con la creación de un corpus anotado basado en metaheurísticas para optimizar el proceso de etiquetado lingüístico bajo restricciones severas de datos y anotación manual (Sierra et al., 2018). Finalmente, se integra en el desarrollo de un sistema de recuperación de información (Sierra et al., 2019).

Complementariamente, Solano et al., 2020 trabajaron en el etiquetado de partes del discurso usando algoritmos metaheurísticos, mejorando la precisión lingüística. En conjunto, estos trabajos evidencian una aproximación progresiva que combina diseño lingüísticamente informado, evaluación experimental y aplicación social.

El trabajo de Muñoz, Jojoa y Catro (2023) aborda uno de los retos más complejos en contextos de ultra-bajos recursos: el procesamiento del habla. Los autores implementaron un sistema de reconocimiento de voz para el Nasa Yuwe, basado en redes neuronales convolucionales (CNN). Así mismo, describen el proceso de recolección y preparación del corpus de audio, así como el diseño de un modelo CNN entrenado para la identificación de patrones fonéticos del Nasa Yuwe. Los resultados experimentales evidencian que, aún con volúmenes reducidos de datos, es posible alcanzar niveles funcionales de reconocimiento.

Salazar-Cárdenas (2022) analiza y compara estrategias de MT con múltiples enfoques que incluyen modelos estadísticos y neuronales, así como técnicas de aprendizaje por transferencia, entrenamiento multilingüe y adaptación de modelos preentrenados, evaluando su impacto en dos lenguas indígenas colombianas: Nasa Yuwe y Wayuunaiki. Se demuestra que los enfoques multilingües y de transferencia superan consistentemente a los modelos entrenados desde cero, especialmente cuando se combinan con preprocesamiento lingüísticamente informado.

Las soluciones previas de aplicación de NLP al Nasa Yuwe se orientan a tareas técnicas específicas y no responden plenamente a las necesidades de comunicación intercultural de las comunidades Nasa, en las que el significado de una expresión depende del contexto cultural, social y situacional en el que se produce. De esta manera, se evidencia la ausencia de herramientas que permitan interpretar las expresiones en su contexto. En este sentido, el presente proyecto responde a una brecha concreta y socialmente relevante.

3. Materiales y métodos

El proceso metodológico se basó en la construcción manual de un corpus lingüístico con participación directa de la comunidad indígena de San Lorenzo – Caldono. Para la implementación del sistema se empleó el framework Rasa (Rasa Technologies, 2025), junto con el lenguaje de programación Python y archivos de configuración en formato YAML, lo que permitió diseñar una arquitectura de software modular (Zhong, Feitosa, Avgeriou, Huang, Li y Zhang, 2024) y adaptable a las particularidades de esta lengua.

El desarrollo del software se realizó mediante un enfoque de prototipado evolutivo en el cual el sistema fue construido y mejorado de forma progresiva a través de pruebas y ajustes continuos.

3.1 Herramienta Rasa

Rasa es un framework de código abierto desarrollado principalmente en el lenguaje de programación Python, orientado a la creación de asistentes conversacionales y chatbots inteligentes. Esta herramienta permite implementar técnicas NLP, como la clasificación de intenciones, la extracción de entidades y la gestión del diálogo. Rasa es apropiado para desarrollar proyectos académicos y comunitarios que trabajan con lenguas de bajos recursos digitales, ya que posibilita el entrenamiento de modelos a partir de corpus propios definidos según criterios lingüísticos y culturales.

Para la configuración del sistema, Rasa emplea archivos en formato YAML, un lenguaje de serialización de datos diseñado para ser legible y fácilmente editable por humanos. Este formato se utiliza para definir intenciones, entidades, ejemplos de entrenamiento, historias conversacionales y reglas del dominio del chatbot, lo que facilita la organización del conocimiento lingüístico y la adaptación del sistema a las particularidades semánticas y culturales de la lengua Nasa Yuwe.

3.2. Corpus lingüístico

Se recolectaron 50 frases en lengua Nasa Yuwe, aportadas directamente por miembros de la comunidad indígena mediante consultas participativas de los integrantes del Tejido del Nasa Yuwe. Este grupo se ha encargado, desde el año 2005 hasta la actualidad, de la elaboración e implementación de planes de estudio orientados al fortalecimiento y revitalización de la lengua materna en las sedes educativas del resguardo San Lorenzo - Caldono, con el propósito de promover la escritura y el uso oral de la lengua sin someterla al español.

Asimismo, el Tejido del Nasa Yuwe ha impulsado estrategias de revitalización lingüística a través de talleres en instituciones educativas del resguardo como como Andalucía, Madres Laura, Plan de Zúñiga, Gualo, Chindaco, la producción de materiales didácticos (libro Figura 2 y Figura 3) y organiza espacios de socialización de las temáticas trabajadas con docentes, tanto hablantes como no hablantes de la lengua, entregando a cada sede educativa el respectivo plan de estudio.

La decisión de trabajar con un corpus reducido respondió a la disponibilidad de datos y al hecho de que se trata de una lengua en riesgo de desaparición, con recursos lingüísticos limitados. Para garantizar la representatividad del material, se seleccionaron frases de uso frecuente que los indígenas Nasa emplean en espacios comunitarios, eventos culturales y encuentros donde existe interacción con entidades gubernamentales o representantes del Estado colombiano.

Las frases recolectadas fueron sometidas a un proceso de limpieza de datos que se realizó manualmente, se incluyó la sustitución de comillas curvadas por comillas simples, la eliminación de espacios en blanco redundantes y de símbolos fonéticos no estandarizados. Asimismo, las expresiones fueron estructuradas de modo que las expresiones correspondientes al idioma Nasa Yuwe se identificarán mediante corchetes [ ], mientras que su entidad equivalente en español se definió explícitamente para facilitar el entrenamiento del modelo.

Posteriormente, se llevó a cabo un proceso de formalización de la escritura, el cual consistió en la unificación de variantes ortográficas y dialectales del Nasa Yuwe en una forma común. Durante este proceso se identificó un inconveniente técnico relacionado con el uso de comillas simples en la apertura o el cierre de los corchetes que delimitan las expresiones en Nasa Yuwe, ya que el framework Rasa no reconocía adecuadamente las entidades cuando dichos caracteres se encontraban dentro del corchete, ejemplo: [**Ptawênxi yuwete'**](Desarmonia_Territorios), en este caso la expresión termina con una comilla simple. Para solucionar este problema, las comillas simples fueron reubicadas fuera del corchete de apertura y cierre, manteniendo su función como delimitadores textuales sin interferir en el reconocimiento de las entidades, lo que permitió un procesamiento correcto.

El compendio de las frases está disponible en el repositorio de github del proyecto: https://github.com/darcygith/nasaYuwe/blob/master/corpus.txt

3.3. Clasificación de intenciones y extracción de entidades

Una vez sistematizado el corpus, se aplicaron técnicas de NLP para la identificación automática de intenciones y entidades, empleando el modelo DIET (Dual Intent and Entity Transformer) integrado en el framework Rasa. Este modelo permite realizar de forma conjunta la clasificación de intenciones y la extracción de entidades dentro de un mismo proceso de aprendizaje. Según Bunk et al. 2020, este modelo se basa en una arquitectura de transformadores que comparte representaciones semánticas entre ambas tareas, lo que permite identificar tanto el propósito comunicativo global del mensaje como los conceptos clave presentes en la expresión lingüística, tales como referencias territoriales, históricas o culturales expresadas en lengua Nasa Yuwe.

Posteriormente a la clasificación de intenciones y la extracción de entidades realizada por el modelo de procesamiento del lenguaje natural, la generación de respuestas se desarrolló mediante acciones personalizadas implementadas en el archivo actions.py, las cuales constituyen la capa lógica del sistema. Las acciones o custom actions utilizan los valores de los slots obtenidos a partir del reconocimiento de entidades para aplicar reglas de interpretación apoyadas en diccionarios contextuales, permitiendo asociar expresiones en lengua Nasa Yuwe con significados en español que preservan su dimensión cultural, simbólica y territorial, conforme al enfoque propuesto en la documentación oficial de Rasa (Rasa, 2023). De este modo, el sistema no se limita a una traducción literal, sino que obtiene interpretaciones explicativas coherentes.

En el corpus, se definieron intenciones y entidades específicas. En el archivo de entrenamiento, se agruparon frases bajo una intención común, por ejemplo, frase_1_2_3, que incluía expresiones relacionadas con la cosmovisión Nasa. Cada frase estaba etiquetada con entidades correspondientes a conceptos culturales, como Territorio_Madre_Tierra, Autonomia_Territorial y Memoria_Historica. La Figura 5 contiene una frase en lengua Nasa Yuwe que ha sido etiquetada con su entidad y expresión respectivamente de color azul y verde, ilustrando cómo se realiza el proceso de anotación para el corpus lingüístico.

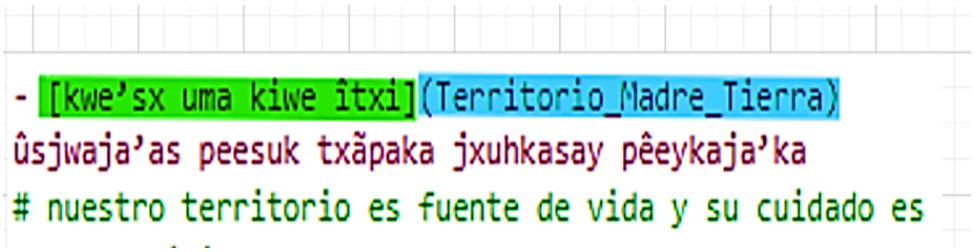

Figura 5. Ejemplo etiquetado, frase en Nasa Yuwe.
Fuente: Elaborado por los autores.

Asimismo, las etiquetas definidas en el corpus se utilizaron como base para el entrenamiento supervisado del modelo DIET (Dual Intent and Entity Transformer), permitiendo asociar cada frase del corpus con una intención específica y sus entidades correspondientes.

En el contexto del NLP, el entrenamiento de modelos consiste en enseñar a un sistema a reconocer patrones lingüísticos a partir de un conjunto de datos etiquetados, mediante el cual el modelo aprende a identificar intenciones, entidades y relaciones dentro del lenguaje, permitiéndole interpretar correctamente frases y generar respuestas coherentes (Jurafsky y Martin, 2025). Este entrenamiento enfocado al NLP es fundamental para comprender el lenguaje de manera contextual, especialmente en lenguas con pocos recursos, donde las traducciones literales podrían distorsionar el significado original.

3.4. Arquitectura

La arquitectura mostrada en la Figura 6 representa el proceso de traducción e interpretación de frases en Nasa Yuwe a español utilizando técnicas de NLP. El proceso inicia con la recolección de frases aportadas (1) por un escritor indígena, las cuales conforman el conjunto de datos iniciales (i.e., el corpus). Estas frases pasan por una etapa de etiquetado y limpieza de datos (2)(3), donde se manejan aspectos técnicos como la eliminación de espacios innecesarios, el uso correcto de comillas y la conservación de símbolos fonéticos propios de la lengua

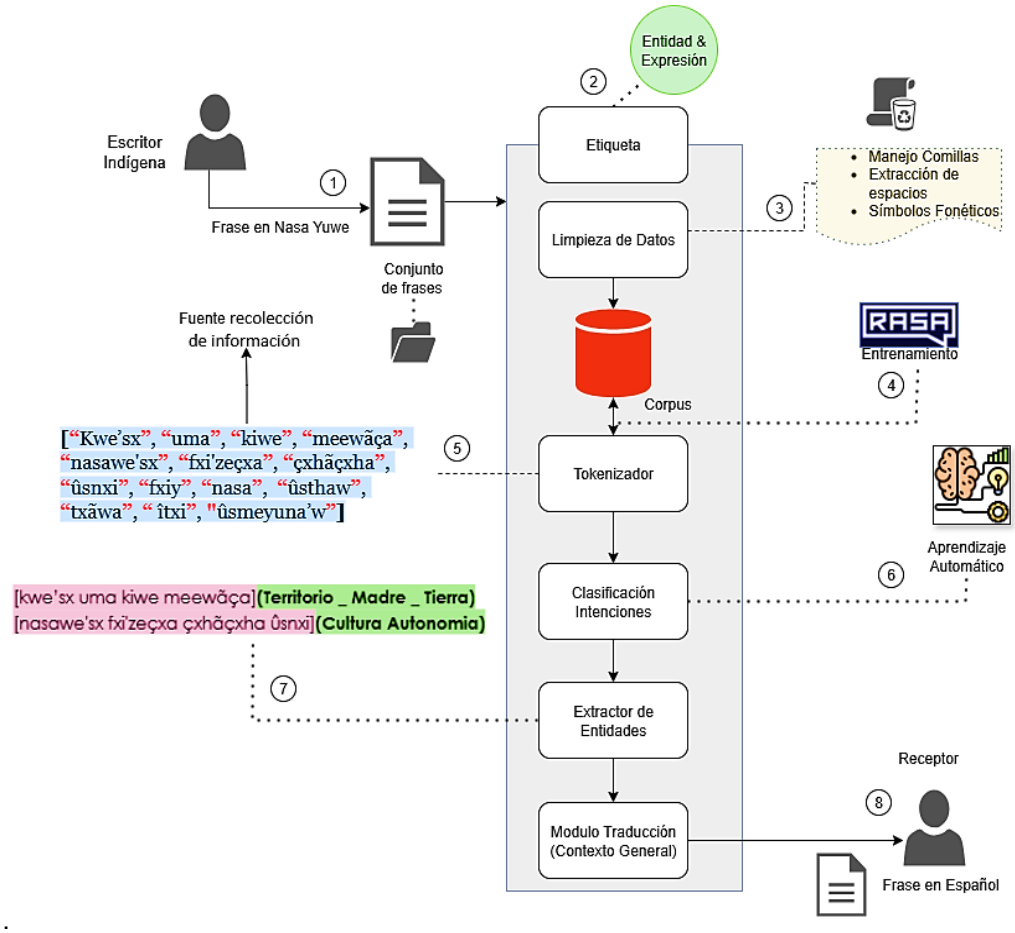

Figura 6. Arquitectura para la interpretación contextual del Nasa Yuwe a español.
Fuente: La arquitectura fue realizada por los autores.

Luego, el corpus procesado se somete a un *tokenizador* (5), que divide las frases en unidades mínimas de análisis. Posteriormente, se aplica un modelo de clasificación de intenciones (6), encargado de identificar el propósito comunicativo de cada enunciado, y un extractor de entidades (7), que asocia palabras o expresiones a categorías culturales relevantes (por ejemplo: Territorio – Madre Tierra, Cultura – Autonomía).

Finalmente, un módulo de traducción contextual (8) interpreta la frase en su conjunto y genera la salida en español, manteniendo el sentido cultural más allá de una traducción literal. Todo este proceso se entrena con RASA y se optimiza mediante aprendizaje automático, permitiendo que el sistema mejore progresivamente su capacidad de comprender y traducir de forma contextual expresiones en Nasa Yuwe.

3.5. Entrenamiento del modelo

El sistema fue entrenado utilizando algoritmos de aprendizaje supervisado a través de la plataforma RASA, la cual permite procesar corpus propios y evaluar el desempeño del modelo (RASA Technologies GmbH, 2025, sección Evaluating your assistant (E2E Testing)). Para ello, el corpus fue dividido en un conjunto de entrenamiento (80%) y un conjunto de prueba (20%), siguiendo prácticas estándar en aprendizaje automático (Raschka y Mirjalili, 2019). El conjunto de entrenamiento permitió ajustar los parámetros del clasificador de intenciones y del extractor de entidades, mientras que el conjunto de prueba se utilizó exclusivamente para evaluar el desempeño del modelo en condiciones de generalización.

Durante la fase de evaluación, RASA comparó las predicciones del modelo (intenciones y entidades detectadas) con las etiquetas reales definidas en el corpus generando matrices de confusión. En el contexto del aprendizaje automático, una matriz de confusión es una herramienta fundamental para la evaluación del desempeño de modelos de clasificación y predicción. Esta matriz permite identificar aciertos y errores mediante cuatro categorías principales: verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos. En un traductor basado en NLP, la matriz permite evidenciar confusiones recurrentes entre unidades lingüísticas, ambigüedades semánticas o fallas en la desambiguación contextual. A partir de ella se derivan las métricas: precisión, exhaustividad (recall), exactitud (accuracy) y medida F1 (Powers, 2020), (Rasa Technologies GmbH, 2025, sección Test Results Analysis).

Las métricas mencionadas proporcionan una medida del grado en que el modelo fue capaz de reconocer correctamente las intenciones comunicativas y los elementos culturales asociados a las expresiones en Nasa Yuwe, evitando depender de traducciones literales y privilegiando la interpretación contextual. Por tanto, este proceso de entrenamiento y evaluación permitió demostrar la viabilidad de emplear técnicas de NLP en lenguas indígenas con recursos ultra limitados, garantizando que los resultados obtenidos tuvieran un respaldo cuantitativo y replicable (Tonja et al., 2024)

4. Resultados

4.1. Corpus generado

El corpus inicial estuvo conformado por 50 frases originales en lengua Nasa Yuwe, recolectadas directamente con apoyo de miembros de la comunidad indígena. Con el fin de mejorar la capacidad de generalización del modelo y reforzar la detección de intenciones y entidades, se aplicó un proceso manual de data augmentation. Este procedimiento consistió en subdividir cada frase en fragmentos más cortos y generar variantes parciales a partir de la oración original (Feng; Gangal; Wei; Chandar; Vosoughi; Mitamura; Hovy, 2021). Como resultado, el corpus final empleado para el entrenamiento estuvo compuesto por 167 ejemplos, correspondientes a las 50 frases originales y sus subdivisiones. En total, el conjunto de datos alcanzó 1933 tokens, distribuidos en 19 intenciones y 29 entidades culturales.

Enlace al repositorio del entrenamiento: https://github.com/darcygith/nasaYuwe/blob/master/data/nlu.yml

4.2. Métricas de desempeño del modelo

El modelo fue evaluado en dos aspectos: clasificación de intenciones y extracción de entidades culturales. La Figura 7 presenta la matriz de confusión correspondiente a la clasificación de un total de 19 intenciones en la cual el modelo no presenta errores. Igualmente la Tabla 4 muestra los resultados de precisión y recall asociados a la Figura 7. Esta "perfección" en los resultados de precisión, recall y puntaje F1 está asociada a la cantidad reducida de frases de entrenamiento y prueba usadas en el prototipo, las cuales están siendo extendidas para realizar pruebas subsiguientes al intérprete.

Tabla 4.
Métricas de rendimiento para la clasificación de intenciones

Intención	Precisión	Recall	Puntaje F1	Soporte
frase_6	1.0	1.0	1.0	6
frase_47	1.0	1.0	1.0	3
frase_37_40_45_19	1.0	1.0	1.0	15
frase_20_21	1.0	1.0	1.0	7
frase_32	1.0	1.0	1.0	5
…	…	…	…	…

Fuente: Elaboración propia a partir de resultados generados por Rasa.

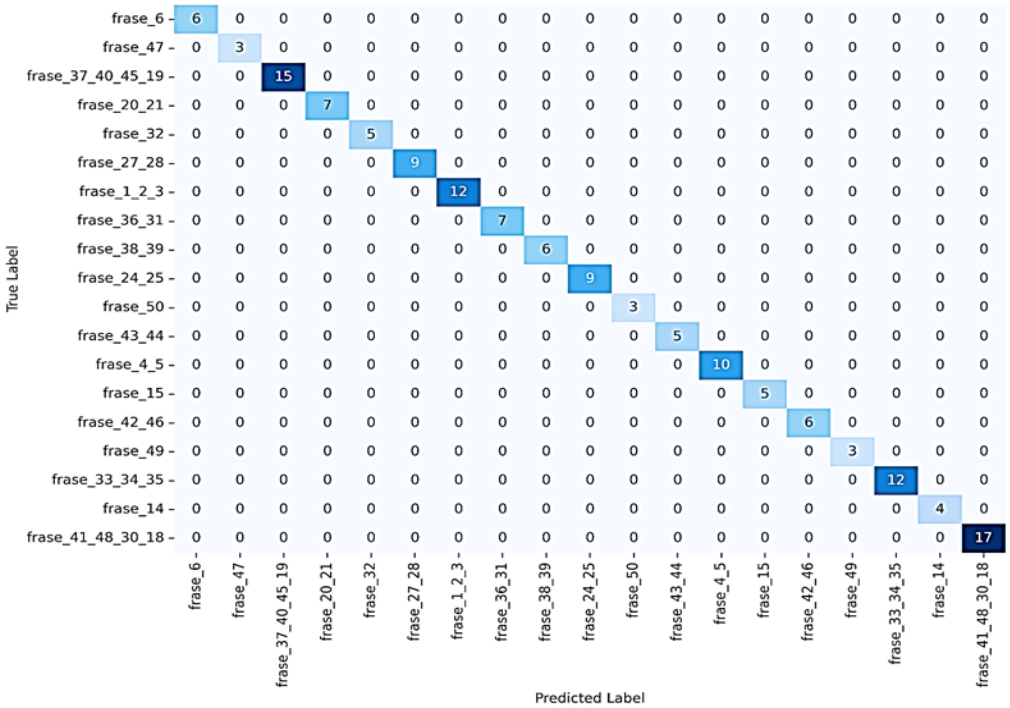

Figura 7. Matriz de confusión de la clasificación de intenciones
Fuente: Herramienta Rasa

Los resultados presentados en la Tabla 5 muestran que las cinco entidades con mayor puntuación F1 (Estados, Cultura_Autonomía, Salud_Educacion_Justicia, Autonomía_Territorial y Territorio_Madre_Tierra) alcanzaron Precisión, Recall y Puntaje F1 de 1.0, es decir, el modelo identificó correctamente todas las instancias de estas entidades sin cometer errores. Este resultado refleja que dichas entidades son fácilmente reconocibles.

Por otro lado, la Tabla 6 muestra que las cinco entidades con menor puntuación F1 (Justicia_Propia, Participación_Comunitaria, Gobierno_Propio, Territorio_Ancestral_Nasa y Derecho_Territorial) presentan métricas ligeramente inferiores y más variables entre Precisión y Recall. Por ejemplo, Justicia_Propia muestra Precisión perfecta (1.0) pero Recall de 0.6667, indicando que algunas instancias reales no fueron detectadas; en contraste, Participación_Comunitaria alcanza Recall perfecto (1.0) pero Precisión de 0.7, señalando que algunas predicciones fueron incorrectas. Estos resultados reflejan que, aunque el modelo mantiene un desempeño alto, las entidades con menor F1 pueden verse afectadas por menor frecuencia, ambigüedad semántica o similitud con otras entidades.

En contraste, la evaluación de la extracción de entidades culturales mostró resultados heterogéneos. Varias de las categorías alcanzaron métricas cercanas al 100%, otras entidades resultaron más complejas de identificar debido a, cómo el tokenizador segmenta el texto en palabras o subunidades antes de la clasificación, lo que puede dificultar la detección precisa de la entidad completa, lo cual se observaron principalmente dificultades asociadas a la baja precisión o a la pérdida de recall.

Tabla 5.
Las cinco entidades con mayor puntuación F1

Entidades	Precisión	*Recall*	Puntaje F1	Soporte
Estados	1.0	1.0	1.0	9
Cultura_Autonomía	1.0	1.0	1.0	18
Salud_Educacion_Justicia	1.0	1.0	1.0	28
Autonomía_Territorial	1.0	1.0	1.0	30
Territorio_Madre_Tierra	1.0	1.0	1.0	50
…	…	…	…	…

Fuente: Elaboración propia a partir de resultados generados por Rasa.

Tabla 6.
Las cinco entidades con menor puntuación F1

Entidades	Precisión	*Recall*	Puntaje F1	Soporte
Justicia_Propia	1.0	0.6667	0.8000	15
Participación_Comunitaria	0.7000	1.0	0.8235	21
Gobierno_Propio	0.7317	1.0	0.8450	30
Territorio_Ancestral_Nasa	0.8064	1.0	0.8928	25
Derecho_Territorial	1.0	0.8125	0.8966	16
…	…	…	…	…

Fuente: Elaboración propia a partir de resultados generados por Rasa.

En la Figura 8 se presenta la matriz de confusión para las entidades culturales. No obstante, la mayoría de categorías fueron clasificadas de manera correcta con el mayor número -lo que evidencia los resultados de la Tabla 5-, en la Figura 9 se observan desaciertos en ciertas entidades que no fueron reconocidas adecuadamente, reflejando las limitaciones del modelo en las categorías.

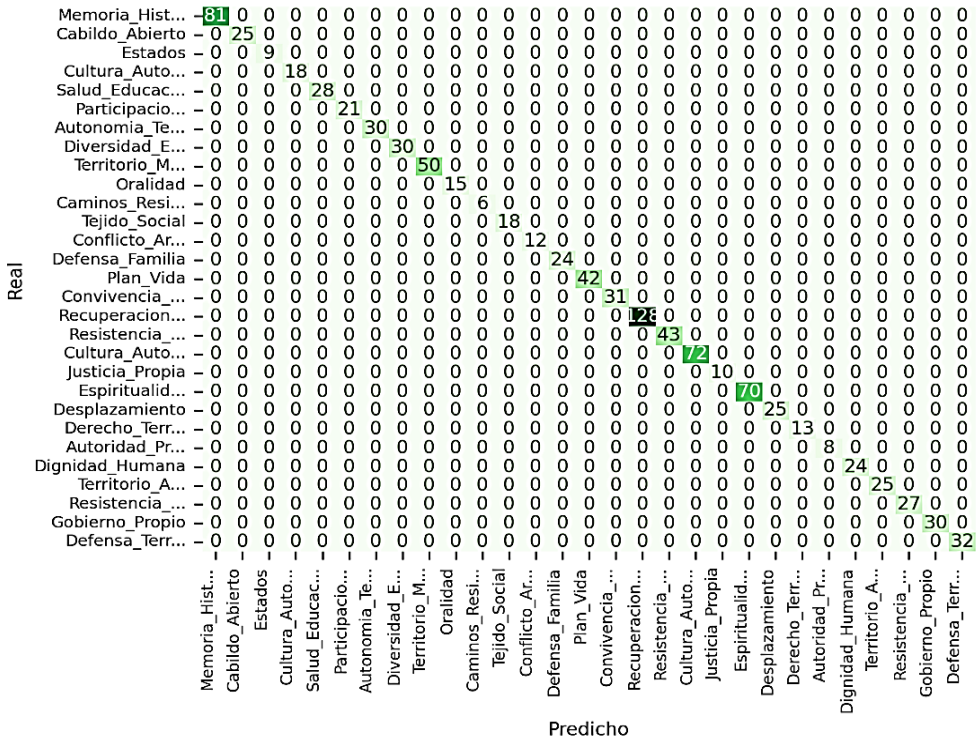

Figura 8. Matriz de confusión de Entidades Clasificadas
Fuente: Herramienta utilizando Python con librerías Pandas y Scikit-learn.

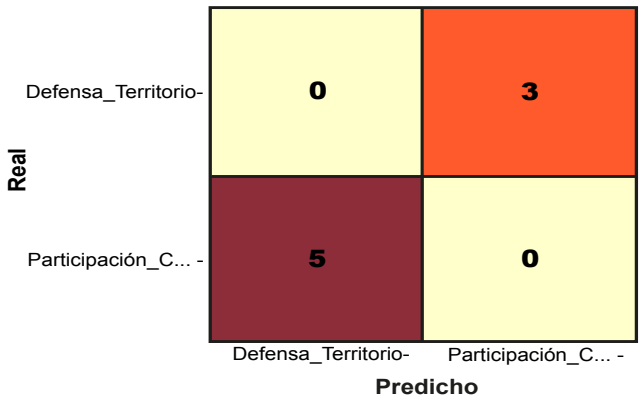

Figura 9. Matriz de confusión de Entidades No clasificadas
Fuente: Herramienta utilizando Python con librerías Pandas y Scikit-learn.

4.3. Resultados cualitativos de interpretación

En la Figura 10, la interpretación evidencia expresiones en Nasa Yuwe que no pueden traducirse de manera literal; el enunciado **"kwe›sx uma kiwe itxi wât fxies fxi›zen ûsthâw nasa we›sxa Thê›sa thwê›jixinitx yahtxna, kwe›s x ksva› wtxi peçkaname txanteywe›sxa"** fue interpretado como una reflexión profunda sobre la relación entre la Madre Tierra, la historia y la sabiduría de los pueblos indígenas, resaltando que la tierra es vida, memoria y espiritualidad. El análisis muestra que su cuidado es colectivo y refleja

la autonomía para decidir sobre su uso y protección, lo cual evidencia que la lengua transmite no solo un mensaje lingüístico, sino también una cosmovisión en la que se integran valores de comunidad, espiritualidad y pertenencia al territorio.

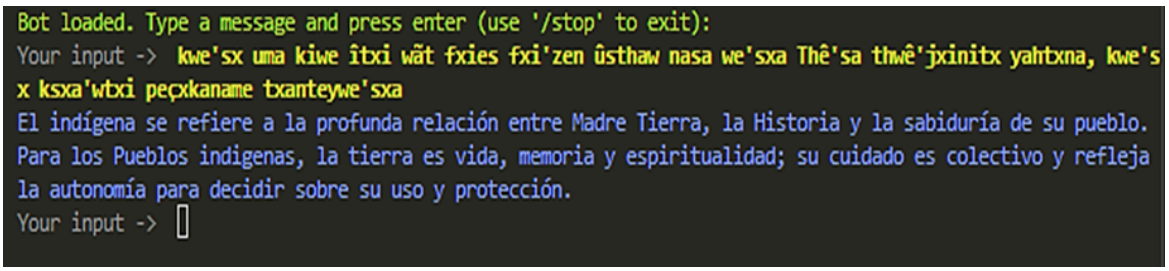

Figura 10. Ejemplo de salida de la consola Rasa después del entrenamiento.
Fuente: Herramienta utilizando Rasa y Python

5. Discusión

Los resultados alcanzados confirman que el uso de NLP en lenguas indígenas es técnicamente viable y culturalmente pertinente. El hecho de que la clasificación de intenciones obtuvo un desempeño bueno demuestra que el modelo capturó con éxito las funciones comunicativas de las expresiones en Nasa Yuwe. La extracción de entidades presentó un rendimiento altamente satisfactorio (98% de exactitud), aunque con limitaciones en categorías culturalmente complejas. La confusión entre Gobierno Propio y Justicia Propia, por ejemplo, refleja la dificultad de distinguir conceptos cercanos dentro de la cosmovisión Nasa, más allá de una simple diferencia léxica.

Comparado con enfoques de traducción literal, el sistema ofrece una ventaja sustancial al mantener el trasfondo cultural de los enunciados, evitando reducciones semánticas que pudieran distorsionar el mensaje original. Sin embargo, el tamaño reducido del corpus limita la robustez del modelo, lo que resalta la necesidad de incrementar la base de datos con frases provenientes de distintos resguardos y contextos comunitarios. Es importante mencionar que la validez final del intérprete la otorga la comunidad indígena, lo cual está contemplado dentro de los alcances de la siguiente etapa del proyecto en la cual ya se está trabajando. Igualmente, se está trabajando en la ampliación del corpus con una nueva categoría de frases referentes a costumbres Nasa. Enlace de la ampliación del corpus: https://github.com/darcysem/nasaYuwe/blob/master/corpus_cultural

Por tanto, esta primera prueba de concepto demuestra que es posible integrar técnicas de NLP y aprendizaje automático en escenarios de revitalización lingüística indígena. El modelo no sólo traduce, sino que interpreta en contexto, aportando a la preservación cultural. No obstante, su consolidación dependerá de la colaboración continua con comunidades hablantes para enriquecer el corpus y reducir ambigüedades.

6. Conclusiones y trabajo futuro

Se diseñó un modelo intérprete del contexto de Nasa Yuwe a español, lo que representa una prueba de concepto del uso de herramientas de NLP que facilita la comprensión contextual de expresiones de líderes indígenas hacia interlocutores hablantes de español. Los resultados obtenidos a partir de las matrices de confusión, evidencian que el modelo desarrollado para la interpretación de la lengua Nasa Yuwe al español, alcanza un desempeño satisfactorio en la identificación de entidades e intenciones. Las categorías relacionadas con Participación_Comunitaria, Memoria_Histórica y Territorio_Madre_Tierra destacan por presentar altos niveles de precisión, lo que evidencia la capacidad del sistema para reconocer entidades e intenciones.

No obstante, se observaron errores puntuales en la clasificación de entidades con significados cercanos, lo que se explica por la similitud semántica de las expresiones. En términos generales, los hallazgos confirman la viabilidad de esta propuesta como una herramienta de apoyo para la traducción contextualizada, aportando a la preservación y difusión de la lengua Nasa Yuwe en escenarios sociales y académicos.

Claramente el proyecto abre camino para varios trabajos futuros. En primer lugar, es indispensable validar la precisión y usabilidad de la herramienta directamente con la comunidad indígena. Esto permitirá una realimentación práctica y alineada con el objetivo principal del proyecto. Actualmente, se está trabajando en dicha validación con la comunidad académica de la Institución Educativa Madre Laura del municipio de Caldono. Adicionalmente, se está proyectando el desarrollo de un chatbot en Nasa Yuwe que servirá como asistente académico a estudiantes de Instituciones Educativas Nasa inicialmente en el área de Ciencias Naturales. Este chatbot se está elaborando igualmente de la mano de la comunidad con el fin de mantener la escencia de la cosmovisión Nasa dentro del mismo.

Finalmente, con el ánimo de pulir el intérprete en su rendimiento, se incorporarán en la evaluación de una siguiente etapa del prototipo, las métricas BLEU y ChRF, una vez que igualmente el corpus se haya robustecido en una medida considerable. La incorporación de estas métricas permitirá hacer análisis comparativos con trabajos del estado del arte que han aplicado NLP a lengua indígenas en peligro. Así mismo, esto permitirá determinar áreas de mejora en la arquitectura propuesta, herramientas, etc.

Sobre los autores

Darcyn Frynet Embus-Quina
Corporación Universitaria Comfacauca – Unicomfacauca, Popayán, Colombia. Estudiante de pregrado.
darcyembus@unicomfacauca.edu.co https://orcid.org/0009-0002-6429-7184

A**ndrés Alexis Elago-Quitumbo**
Corporación Universitaria Comfacauca - Unicomfacauca, Popayán, Colombia. Estudiante de pregrado
andreselago@unicomfacauca.edu.co https://orcid.org/0009-0009-1517-9633

Angela María Rodríguez-Vivas
Docente Investigadora. Corporación Universitaria Comfacauca - Unicomfacauca, Popayán, Colombia.
arodriguez@unicomfacauca.edu.co https://orcid.org/0000-0002-6976-4960

Disponibilidad de datos

Los autores declaran que el artículo contiene todos los datos necesarios y suficientes para la comprensión de la investigación.

Declaración de divulgación

Los autores declaran que no existe ningún potencial conflicto de interés relacionado con el artículo.

Descargo de responsabilidad

Las expresiones, opiniones o interpretaciones expuestas en este artículo son una postura personal de los autores.

Fuentes de financiación

Este artículo es resultado del proyecto “Intérprete Consciente del Contexto de Expresiones en Lengua Nasa Yuwe a Español Basado en Técnicas de Procesamiento de Lenguaje Natural” financiado por la Dirección de Ciencia, Tecnología e Innovación de la Corporación Universitaria Comfacauca – Unicomfacauca, aprobado en la convocatoria interna de proyectos 2025-02 bajo el código VRIE 2025-01S.

Coautoría

Darcyn Frynet Embus-Quina: Investigación y metodología

A**ndrés Alexis Elago-Quitumbo:** Investigación y metodología

Angela María Rodríguez-Vivas: Conceptualización, análisis formal, redacción, revisión y edición

Referencias bibliográficas

1. AGUILAR SANTIAGO, César Antonio; GARCÍA ZÚÑIGA, Hamlet Antonio. Tecnologías del lenguaje aplicadas al procesamiento de lenguas indígenas en México: una visión general. Lingüística y Literatura. 2023. vol. 44, no. 84, p.79-102. https://doi.org/10.17533/udea.lyl.n84a04
2. AMERICASNLP. Recursos y proyectos de procesamiento de lenguas naturales en América. Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas – Universidad Nacional Autónoma de México. 2026 https://turing.iimas.unam.mx/americasnlp/
3. BAUTISTA MORALES, Rolando; MARTÍNEZ RAMÍREZ, Yobani; ROCHA PEÑA, Luis Enrique; MONTES SANTIAGO, Reyna Elisa. Arquitectura de un traductor automático para el idioma Mixteco: un enfoque específico para lenguas indígenas con escasos recursos lingüísticos. En: Revista de Investigación en Tecnologías de la Información. 2024. vol. 12, no. 28. p. 71-81 https://doi.org/10.36825/RITI.12.28.007
4. BAYRAM Mehmet, Ali; FINCAN, Ali Arda; GÜMÜŞ, Ahmet Semih; KARAKAŞ, Sercan; DIRI, Banu; YILDIRIM, Savaş. Tokenization standards for linguistic integrity: Turkish as a benchmark. In: arXiv. 2025. https://doi.org/10.48550/arXiv.2502.07057
5. BUNK, Tobias; VARSHNEYA, Devang; VLASOV, Vladislav; NICHOL, Alan. DIET: Lightweight language understanding for dialogue systems. In: arXiv preprint, 2020. https://doi.org/10.48550/arXiv.2004.09936
6. COLOMBIA MINISTERIO DE CULTURA. Plan decenal de lenguas nativas de Colombia. Bogotá: Ministerio de Cultura de Colombia, 2020 https://www.onic.org.co/images/Cartilla_plan_decenal_de_lenguas_nativas.pdf
7. COLOMBIA MUNICIPIO DE CALDONO CAUCA. Plan de Desarrollo Municipal 2020–2023. Caldono: Alcaldía Municipal de Caldono, 2023 https://www.tangara.gov.co/wp-content/uploads/2023/12/CALDONO-PLAN-DE-DESARROLLO-MUNICIPAL-2020-2023.pdf
8. CONSEJO REGIONAL INDÍGENA DEL CAUCA-CRIC. Estructura organizativa. Popayán, 2024 https://www.cric-colombia.org/portal/estructura-organizativa/
9. DEPARTAMENTO ADMINISTRATIVO NACIONAL DE ESTADÍSTICA - DANE. DANE - Infografía: Perfil Cauca. Bogotá, 2018 https://sitios.dane.gov.co/cnpv/app/views/informacion/perfiles/19_infografia.pdf
10. DEPARTAMENTO ADMINISTRATIVO NACIONAL DE ESTADÍSTICA - DANE. Población indígena de Colombia. Bogotá, 2019 https://www.dane.gov.co/files/investigaciones/boletines/grupos-etnicos/presentacion-grupos-etnicos-2019.pdf
11. DOWNEY, Anna; ETXEBERRIA, Urtzi; MÜLLER, Mathias; COTTERELL, Ryan. Unsupervised Multilingual Sequential Segmentation for Extremely Low-Resource Languages. En: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021). Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021.
12. FENG, Steven Y.; GANGAL Varun; WEI, Jason; CHANDAR, Sarath; VOSOUGHI, Soroush; MITAMURA, Teruko; HOVY, Eduard. A survey of data augmentation approaches for NLP. In: arXiv preprint, 2021 https://doi.org/10.48550/arXiv.2105.03075
13. GUTIÉRREZ ARRIAGA, Óscar Felipe; ALVARADO RODRÍGUEZ, María Elena. La importancia social y jurídica de la conservación de las lenguas indígenas en la Ciudad de México. In: LATAM Revista Latinoamericana de Ciencias Sociales y Humanidades. 2024. vol. 5, no. 5. p. 1133-1148 https://doi.org/10.56712/latam.v5i5.2650
14. JEHANGIR, Basra; RADHAKRISHNAN, Shyam; AGARWAL, Ramesh. A survey on named entity recognition: datasets, tools, and methodologies. In: Natural Language Processing Journal. 2023, vol. 3. e100017 https://doi.org/10.1016/j.nlp.2023.100017
15. JURAFSKY, Daniel; MARTIN, James H. Speech and language processing. 3ª ed. Stanford: Stanford University, 2025 https://web.stanford.edu/~jurafsky/slp3/ed3book_aug25.pdf
16. KANN, Katharina; ETTINGER, Allyson; ANTONIO, Esteban; MURADOĞLU, S. M.; MAGER, Manuel; ONO, Hiroki; ORTIZ, María; VARGAS, Ricardo; WU, Shijie. AmericasNLI: Evaluating Zero-shot Natural Language Inference in Indigenous Languages of the Americas. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022). Dublin, Ireland: Association for Computational Linguistics, 2022
17. LLERENA GARCÍA, Edgardo; DÍAZ PERNETT, Manuel. Herramientas tecnológicas y didácticas para la conservación de la lengua Emberá. En: Lingüística y Literatura. 2023. vol. 44, no. 84, p. 124-153 https://doi.org/10.17533/udea.lyl.n84a06
18. LOZANO LOZANO, Carlos Aníbal; CARVAJAL CALDERÓN, Paola Hasbleidy; ÁLVAREZ GÓMEZ, Cristina. Tecnologías emergentes para la formación sobre consulta previa y liderazgo: una propuesta para jóvenes de la comunidad indígena Wayuú de la Guajira. Universidad Ean. 2025. https://hdl.handle.net/10882/15374
19. MUÑOZ BURBANO, Pablo Enrique; JOJOA GÓMEZ, Pedro Enrique; CASTRO CAICEDO, Fernando Manuel. Implementation

of a Voice Recognition System in the Nasa Yuwe Language Based on Convolutional Neural Networks. In: Revista Ingeniería Solidaria. 2023, vol. 19, no. 1 https://dialnet.unirioja.es/servlet/articulo?codigo=10116949

20. NAGOUDI, El Moatez Billah; CHEN, Wei-Rui; ABDUL MAGEED, Muhammad; CAVUSOGLU, Hasan. IndT5: a text-to-text transformer for 10 indigenous languages. In: arXiv Preprint. 2021. https://doi.org/10.48550/arXiv.2104.07483

21. PARANKUSHAM, Kanishka; RIZK, Rodrigue; KC, Santosh. LakotaBERT: A Transformer-based Model for Low Resource Lakota Language. In: arXiv Preprint. 2025. https://arxiv.org/abs/2503.18212

22. POWERS, David M. W. Evaluation: from precision, recall and F-Measure to ROC, Informedness, Markedness and Correlation. In: arXiv preprint, 2020. https://doi.org/10.48550/arXiv.2010.16061

23. RASCHKA, Sebastian; MIRJALILI, Vahid. Python machine learning. 3ª ed. Birmingham–Mumbai: Packt Publishing, 2019 https://jcer.in/jcer-docs/E-Learning/Digital%20Library%20/E-Books/python-machine-learning-and-deep-learning-with-python-scikit-learn-and-tensorflow-2.pdf

24. RASA-TECHNOLOGIES-GMBH. Custom actions. Rasa Open Source Documentation, 2023 https://rasa.com/docs/rasa/custom-actions

25. RASA-TECHNOLOGIES-GMBH.Rasa documentation, 2025 https://rasa.com/docs/

26. RASA-TECHNOLOGIES-GMBH. Evaluating your assistant (E2E testing). Berlín, 2025 https://rasa.com/docs/pro/testing/evaluating-assistant/

27. RONGALI, Sateesh Kumar. Natural language processing (NLP) in artificial intelligence. In: World Journal of Advanced Research and Reviews. 2025. vol. 25, no. 1, p. 2515-2519 https://doi.org/10.30574/wjarr.2025.25.1.0277

28. SALAZAR CÁRDENAS, Isabel. Machine translation strategies for low-resource Colombian indigenous languages. Universidad de los Andes, 2022. https://hdl.handle.net/1992/62941

29. SIERRA, Luz Marina; COBOS, Carlos Alberto; CORRALES, Juan Carlos; ROJAS CURIEUX, Tulio. Building a Nasa Yuwe language test collection. In: GELBUKH Alexander (ed.). Computational Linguistics and Intelligent Text Processing. Cham: Springer International Publishing, 2015. Lecture Notes in Computer Science, vol. 9041. p.112-123. https://doi.org/10.1007/978-3-319-18111-0_9

30. SIERRA MARTÍNEZ, Luz Marina; COBOS, Carlos Alberto; CORRALES, Juan Carlos. Tokenizer adapted for the Nasa Yuwe language. In: Computación y Sistemas. 2016. vol. 20, no. 3. p. 355-364 https://doi.org/10.13053/CyS-20-3-2455

31. SIERRA MARTÍNEZ, Luz Marina; COBOS, Carlos Alberto; CORRALES MUÑOZ, Juan Carlos; ROJAS CURIEUX, Tulio; HERRERA-VIEDMA, Enrique; PELUFFO-ORDÓÑEZ, Diego Hernán. Building a Nasa Yuwe language corpus and tagging with a metaheuristic approach. In: Computación y Sistemas. 2018. vol. 22. p. 881-894 http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1405-55462018000300881

32. SIERRA MARTÍNEZ, Luz Marina; COBOS, Carlos Alberto; CORRALES, Juan Carlos; ROJAS CURIEUX, Tulio; GÓMEZ, Luis Carlos. Sistema de recuperación de información para apoyar la revitalización del Nasa Yuwe. In: RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação. 2019. no. 17, p. 407-422 https://www.researchgate.net/publication/330844598_Information_Retrieval_System_for_Nasa_Yuwe

33. SOLANO JIMÉNEZ, Miguel Alexis; TOBAR CIFUENTES, José Julio; SIERRA MARTÍNEZ, Luz Marina; COBOS, Carlos Alberto. Adaptation, comparison, and improvement of metaheuristic algorithms to the part-of-speech tagging problem. In: Revista Facultad de Ingeniería Universidad de Antioquia. 2020, vol. 29, no. 54, e11762 https://doi.org/10.19053/01211129.v29.n54.2020.11762

34. SOYLU, Dilek; ŞAHİN, Ayşe. The Role of AI in Supporting Indigenous Languages. In: AI and Tech in Behavioral and Social Sciences, 2024. vol. 2, no. 4, p. 11-18 https://doi.org/10.61838/kman.aitech.2.4.2

35. TONJA, Atnafu Lambebo; BALOUCHZAHI, Fazlourrahman; BUTT, Sabur; KOLESNIKOVA, Olga; CEBALLOS, Héctor; GELBUKH, Alexander; SOLORIO, Thamar. NLP progress in indigenous Latin American languages. Proceedings of NAACL, 2024 https://doi.org/10.18653/v1/2024.naacl-long.385

36. UNESCO Office Montevideo and Regional Bureau for Science in Latin America and the Caribbean; GONZÁLEZ ZEPEDA, Luz Elena; MARTÍNEZ PINTO, Cristina Elena. Inteligencia artificial centrada en los pueblos indígenas: perspectivas desde América Latina y el Caribe. Montevideo: UNESCO Office Montevideo and Regional Bureau for Science in Latin America and the Caribbean, 2023. 53 p. https://unesdoc.unesco.org/ark:/48223/pf0000387814

37. URBINA PULIDO, Fabián-Andrés. Proyecto de traducción del navegador Firefox a la lengua Nasa: un paso en la inclusión digital. En: Inclusión y Desarrollo, 2018. vol. 5, no. 2. p. 125-142 https://doi.org/10.26620/uniminuto.inclusion.5.2.2018.125-142

38. URIBE MUÑOZ, Cristian Mauricio. Linguistic Weakening of Nasa-Yuwe in the Path Yu' Community (Cauca, Colombia). En:

Íkala, Revista de Lenguaje y Cultura. 2025. vol. 30, no. 3. https://doi.org/10.17533/udea.ikala.360057

39. ZHONG, Chenxing; FEITOSA, Daniel; AVGERIOU, Paris; HUANG, Huang; LI-Yue; ZHANG-He. PairSmell: A novel perspective inspecting software modular structure. In: arXiv preprint, 2024 https://arxiv.org/abs/2411.01012