Translation systems for low-resource Colombian Indigenous languages, a first step towards cultural preservation

Juan C. Prieto , Cristian A. Martinez, Melissa Robles, Alberto Moreno,
Sara Palacios, Rubén Manrique

The Department of Systems and Computing Engineering
Universidad de Los Andes
{jc.prietoa,ca.martinez2,mv.robles,a.morenoc23,s.palaciosc,rf.manrique}@uniandes.edu.co

Abstract

The use of machine learning and Natural Language Processing (NLP) technologies can assist in the preservation and revitalization of indigenous languages, particularly those classified as "low-resource". Given the increasing digitization of information, the development of translation tools for these languages is of significant importance. These tools not only facilitate better access to digital resources for indigenous communities but also stimulate language preservation efforts and potentially foster more inclusive, equitable societies, as demonstrated by the AmericasNLP workshop since 2021. The focus of this paper is Colombia, a country home to 65 distinct indigenous languages, presenting a vast spectrum of linguistic characteristics. This cultural and linguistic diversity is an inherent pillar of the nation's identity, and safeguarding it has been increasingly challenging given the dwindling number of native speakers and the communities' inclination towards oral traditions. Considering this context, scattered initiatives exist to develop translation systems for these languages. However, these endeavors suffer from a lack of consolidated, comparable data. This paper consolidates a dataset of parallel data in four Colombian indigenous languages - Wayuunaiki, Arhuaco, Inga, and Nasa - gathered from existing digital resources. It also presents the creation of baseline models for future translation and comparison, ultimately serving as a catalyst for incorporating more digital resources progressively.

1 Introduction

In the field of natural language processing (NLP), low-resource languages are characterized by limited written or spoken digital material. Consequently, applying translation models rooted in advanced neural architectures to these languages is challenging due to the models' high dependence on substantial data volumes (Wang et al., 2021). However, recent years have seen a growing trend towards working with low-resource languages for machine translation based on transformers. Notably, efforts to preserve indigenous languages have significantly contributed to advancements in this area (Mager et al., 2018; Ortega et al., 2020a,b; Chen and Fazio, 2021). The Americas, in particular, host numerous endangered indigenous languages spoken by small populations. In response, several researchers have devoted their work towards developing translation models for some of these languages (Ngoc Le and Sadat, 2020). Highlighting this effort, AmericasNLP was convened in 2021 as the first global workshop dedicated to the application of NLP to American indigenous languages (Mager et al., 2021).

The development of translation tools catering specifically to indigenous languages has the potential to confer numerous advantages, including enriched access to digital resources and the promotion of language preservation efforts. As the availability of digital resources proliferates, it becomes increasingly imperative for indigenous communities to have access to information in their respective languages, in order to safeguard their unique cultures and traditional ways of living. Regrettably, the limited resources available in indigenous languages restrict their accessibility to crucial digital materials. The creation of translation tools could serve as a resolution to this issue, enabling communities to translate digital resources into their languages. Beyond simply providing improved access to information, these tools could stimulate language preservation by offering a medium for language revitalization. Furthermore, translation tools could provide a platform for indigenous communities to engage with the global community, contributing to their economic and social development. By fostering cross-cultural understanding and mutual respect, translation tools could play a critical role in constructing more inclusive and equitable societies.

As acknowledged by the National Indigenous

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 7–14
June 21, 2024 ©2024 Association for Computational Linguistics

Organization of Colombia (ONIC) (de Gobierno Indígena – ONIC, 2015), Colombia's linguistic diversity is marked by the existence of 65 indigenous languages, alongside Spanish and two Creole languages. Among the 32 different departments in the territory, the regions of Amazonas and Vaupés, located in the southern sector of Colombia, stand out for their significant diversity of indigenous languages. This assortment of indigenous languages in Colombia is highly distinctive due to their different characteristics. For instance, Colombia accommodates tonal languages similar to those found in Southeast Asia and Central Africa, along with languages representing all four classic morphological types. These include inflectional (as exemplified by Kogui and Arhuaco), agglutinative (like Achagua, Andoque, and Páez), isolating akin to the Malayo-Polynesian languages (such as Embera and Creole), and polysynthetic (such as Kamsá) (de Estudios de Lenguas Aborígenes , C.C.E.L.A).

According to ONIC, the task of preserving these indigenous voices is increasingly formidable, primarily due to their endangerment amid the dwindling number of native speakers. Over half of these languages, alarmingly, have fewer than a thousand active speakers, thus exacerbating their preservation and conservation challenges (de Gobierno Indígena – ONIC, 2015). Further complexities arise from the proclivity of certain communities to uphold their oral traditions over written modes. Take, for example, the Inga community, wherein cultural identity preservation is embedded in the prioritization of oral tradition as the primary vehicle for knowledge transmission. Such communities contend that the detachment from oral traditions can incite a loss of numerous practices that necessitate face-to-face interaction and stimulate dialogue (Rodríguez and Narváez, 2022).

Despite the existence of scattered efforts aimed at creating translation systems for some indigenous Colombian languages, we found no consolidated data that allows for reuse and/or comparison (Sierra et al., 2015; Sierra Martínez et al., 2016, 2018; Fernandez et al., 2013). Only recently have initiatives emerged that have promoted data replicability and openness in Colombian Indigenous languages (Graichen et al., 2023). The primary objective of this work is to consolidate a dataset of parallel data in four Colombian indigenous languages: Wayuunaiki, Arhuaco, Inga, and Nasa. This dataset is a product of the compilation, processing, and alignment of already existing digital resources in these languages. It aims to serve as a starting point to encourage the incorporation of new digital resources progressively. A second objective is to create a set of baseline models for translation that can be used for comparison in future research.

2 Related Work

Low-resource languages, which lack the digital or written material needed to build a corpus or a linguistic collection, include indigenous or endangered languages, region-specific dialects, or languages without substantial digital resources despite the existence of millions of speakers. The shortage of available data, which often results from limited access to technology, opens up opportunities to apply various techniques like Data Augmentation, Back-Translation, and Transfer Learning to mitigate this scarcity.

Studies on diverse languages have explored Neural Machine Translation, Transfer Learning, and advanced model architectures. For instance, an investigation of the Tigrinya Ethiopian language yielded positive results using Transfer Learning (Öktem et al., 2020). In another study comparing three different models, the Transformer Network performed the best with a parallel corpus of Yoruba and English (Adebara et al., 2021). Also, a modification of the Transformer architecture led to improved results for South African languages (van Biljon et al., 2020). Numerous studies demonstrate the effectiveness of Transfer Learning and advanced modeling techniques across global languages. For instance, Finnish and Czech were used as parent languages to assist low-resource Estonian and Slovak languages via Transfer Learning, leading to improvements over the baseline for almost all pre-trained models (Kocmi and Bojar, 2018). The combination of attention layers and byte-pair encoding in Transfer Learning also notably enhanced translation capabilities for Turkic languages (Nguyen and Chiang, 2017). Notably, the pairing of the Transformer architecture with the Back-Translation technique resulted in improved translation quality for several language pairs (Przystupa and Abdul-Mageed, 2019).

American Indigenous Languages, historically deficient in written records, pose unique challenges for language preservation. The AmericasNLP workshop provided a platform to unite global research groups to address these challenges, with a focus on machine translation across various indigenous languages. Numerous techniques and strategies were employed across the participating teams in the different versions, yielding promising results. While some teams achieved success with multilingual neural networks (Vázquez et al., 2021; Knowles et al., 2021; Moreno, 2021), others found Statistical Machine Translation more effective (Parida et al., 2021). Additionally, using unique sources such as Wikipedia and biblical texts to build the corpus yielded significant results, enhancing progress beyond baseline starts (Billah-Nagoudi et al., 2021). In the recent 2023 edition of the Workshop, the winning team's shared-task strategy comprised extending and finetuning several variants of the NLLB-200 (NLLBTeam, 2022). This cutting-edge machine translation model is specifically tailored for environments with scarce resources. Their submission surpassed the baseline by an average chrF of 11% across all languages, yielding especially considerable enhancements for Aymara, Guarani, and Quechua (Gow-Smith and Sánchez Villegas, 2023).

As for the indigenous Colombian languages' translations, Graichen et al. (2023) and Robles et al. (2024) studies are the only works we discovered that present a translation system from Wayuunaiki to Spanish and Ika (Arhuaco) to Spanish. Graichen et al. (2023) applied various unsupervised and semisupervised subword segmentation methods to enrich the data used to train a transformer-based model with linguistic information. According to the results, the incorporation of linguistic knowledge helps the system to generate improved translation. Nonetheless, these methodologies introduced substantial noise into the process.

3 Data

3.1 Langauges

The linguistic diversity of Colombia is characterized by a variety of indigenous languages, including Wayuunaiki, Nasa Yuwe, Arhuaco (Ika), and Inga. Wayuunaiki, predominantly spoken by the Wayuú community in the La Guajira region, is the most widely spoken indigenous language in Colombia. The 2005 DANE census report (de Gobierno Indígena – ONIC, 2015) indicates a population of 270.413 Wayuú individuals, making it the largest indigenous demographic. Wayuunaiki is an agglutinative language, characterized by the combination of independent morphemes to form words.

On the other hand, Nasa Yuwe, primarily spoken by the Nasa people in the Cauca department and smaller regions such as Valle del Cauca, Tolima, and Huila, is the second most spoken indigenous language. Although traditionally classified as part of the Chibchan language family, it is now largely considered an isolated language.

Likewise, the Arhuacos, who inhabit the western and southeastern regions of the Sierra Nevada de Santa Marta, speak the Ika language. Ika, a member of the Chibchan language family, is distinguished by its sentence structure, which involves the addition of various morphemes to a root or lexeme.

Lastly, the Inga community, descendants of the Inca civilization, primarily inhabit the Sibundoy Valley within the Putumayo region, with additional settlements in Nariño and Cauca. Their language, Inga, belongs to the Quechuan family. The linguistic diversity of these communities contributes to the rich cultural tapestry of Colombia.

3.2 Data Collection

Locating documents written in both indigenous languages and Spanish is challenging due to a lack of translated resources, as shown by limited translations of the constitution. The Colombian Center for Studies in Aboriginal Languages (de Estudios de Lenguas Aborígenes, 1994) has only translated the constitution into seven indigenous languages: Inga, Guambiano, Arhuaco, Kamentsa, Kubeo, Nasa Yuwe, and Wayuunaiki. This lack of translated resources extends to religious texts as well. For instance, complete translations of the Bible are only available in a handful of indigenous languages. Specifically, complete or partial translations of the Bible exist in Wayuunaiki, Arhuaco, and Nasa Yuwe.

For the Wayuunaiki language, in addition to the Bible and the constitution, there are various documents that delve into the characteristics of the language and provide sections with translated excerpts. An example is the document "La conjugación del verbo en la lengua Wayuu" (Álvarez, 2016) which offers a comparative perspective on verbal conjugation in Wayuunaiki and Spanish, addressing semantic, morphological, and syntactic aspects. Similar to this, the book "Compendio de la Gramática de la Lengua Wayuu" (Álvarez, 2017) and the article "Panorámica de la fonología y morfología de la lengua Wayuu" (Álvarez González, 2021) detail the important features of the morphology, phonology, and syntax of the language, presenting comparative examples between the two languages. The book “Vamos a hablar nuestra lengua” (Flórez et al., 2020) provides accurate information about writing, grammar, and cultural aspects implicit in everyday expressions and words. The consolidation of this language dataset was ultimately achieved by utilizing a short story (Cue, 2012) and a dictionary (Amaya, 2021). This last resource, compiled in 2021 by Rafael Jose Negrette Amaya, encompasses a total of 74.583 translated phrases and words in both languages.

To consolidate the Wayunnaiki dataset, sentences were extracted from PDFs or web pages. The dataset for the New Testament of the Bible (YouVersion, 2023) was constructed using a web scraping process. This entailed systematically pairing sentences by verse and chapter of the book. In some of the other sources, we relied on additional pre-processing steps via Large Language Models (LLMs). Since the texts did not follow a defined format, GPT-4 (OpenAI, 2023) was employed to extract candidate texts from the selected documents. A prompt template was utilized to identify and tabulate sentences in both Wayuu and Spanish languages, with this process being conducted at three-page intervals. An illustrative example of the prompts used is, “`Identify sections in the text where Wayuu and Spanish sentences co-occur and create a tabulated representation...`”. Then a manual review process was carried out, filtering incomplete translations, as the Spanish sentence contained blank spaces or non-alphabetic characters. The combined use of web scraping and GPT-4 in this manner allowed for the creation of a comprehensive and well-structured dataset, thereby enhancing the overall readability and coherence of the information.

The primary data source for Nasa Yuwe was the constitution (de Estudios de Lenguas Aborígenes, 1994). This document is partitioned into sections: introductory letters, articles, and a dictionary that is a compendium of frequently translated words and phrases from Nasa Yuwe to Spanish. An Optical Character Recognition (OCR) (Smith et al., 2009) process was employed to extract the 23 translated articles, along with introductory letters and acknowledgments. Discrepancies were observed in the introductory letters as the Nasa Yuwe translation occasionally contained more content than the Spanish version. A manual review of the letter text was necessitated to pinpoint precise word translations and sentence terminations and to eliminate any additional Nasa Yuwe content not found in the Spanish text. Additionally, a dictionary (originarios. Lenguas de América) containing 3.729 words and brief phrases in both Spanish and Nasa Yuwe was included. This dictionary is presented in HTML format and was processed using the Beautiful Soup tool (Richardson, 2007), a web scraping library. This was followed by a manual error correction procedure to guarantee the precision of the extraction process further.

For the Arhuaco our biggest data source was the Bible. We again used web scraping and the BeautifulSoup library on selected chapters of the Old and New Testaments (para el Desarrollo de Pueblos Marginados). Our second source for Arhuaco was the constitution (de Estudios de Lenguas Aborígenes, 1994). Given the low quality of the online document, it was processed through a text identification procedure with the assistance of Google’s DocumentAI OCR. This particular API employs a neural network designed to enhance the recognition of text within PDF documents, which facilitates the conversion of visual document data into text, organizing the content into distinct paragraphs. This segmentation significantly simplifies the subsequent concatenation of sentences, thereby streamlining the text analysis and processing tasks. The initial phase of the processing involved the analysis of introductory letters, which were messages in either Spanish or Arhuaco, expressing gratitude, detailing efforts, and explaining the reasons behind the creation of the book. The document was processed in blocks of text, and sentences were consolidated by matching paragraphs and sentences separated by period. After the letters, we moved to process the constitution articles, which involved creating a correspondence between titles in Spanish and Arhuaco and identifying the pairs. Finally, a manual pairing process was carried out for the dictionary section of the constitution. The last source used for Arhuaco was a book titled “Cantando desde la Sierra” (de la comunidad arhuaca de Jewrwa, 2014), which contains various short stories and poems in Arhuaco, accompanied by their Spanish translations. To utilize this, each poem was detected and subsequently matched with its corresponding translation in the other language.

The Inga language was undoubtedly the most challenging. We employ as a primary source a comprehensive dictionary of words and phrases (de Educación Inga de la Organización “Musu Runakuna”, 1997). This dictionary was processed using a methodology similar to the one used for the Wayuunaiki language. This involved the use of the GPT-4 (OpenAI, 2023) for identifying candidate pairs translations within the dictionary. Subsequently, a further step of data cleaning was undertaken to minimize the occurrence of false positives. The second source was the constitution (de Estudios de Lenguas Aborígenes, 1994), which was processed using the same procedure used for Arhuaco, using an Google OCR and a manual cleaning step. Table 1 shows the size of the training data for each language.

4 Baseline construction

Following the recent results obtained by the work of (Robles et al., 2024) and (Gow-Smith and Sánchez Villegas, 2023) we use NLLB-200, a state-of-the-art machine translation model specifically designed for low-resource settings. We experiment with different distilled versions of NLLB-200 with 600M and 1.3B parameters. Each dataset was randomly divided into training, validation, and testing sets, comprising 80%, 10%, and 10% of the total number of sentences, respectively. The resulting partition, along with the models and code, can be accessed at `https://github.com/juanks235/MT-Colombian-Indigenous-Languages`.

4.1 Experimental Setup

In our approach, we execute model training distinctively for each language pair available in our dataset. We refine the embedding matrix to encompass tags for newly added languages, scrutinizing for not recognized tokens and employing text normalization to reduce potential problems related to unrecognized punctuation. The application of normalization ensures the accurate processing of the text, obviating any unknown tokens and providing the promise that the vocabulary of the tokenizer doesn’t require an update for the target language. Nevertheless, if it becomes necessary for the tokenizer’s vocabulary to be updated, we implement an update to include any new or unrecognized tokens previously overlooked. Our experimental framework operates on four A40-48GB, with a batch size of 16, 1000 warmup steps, 57000 training steps, featuring a learning rate of $1e-4$ and a weight decay of $1e-3$. For automated evaluations, we leverage SacreBLEU (Post, 2018) for computing BLEU scores and chrF2++ (Popović, 2017) to measure chrF2.

4.2 Results

Table 2 presents the results obtained with different trained models. As expected, the 1.3B model performed better in translations from Spanish to the target language.

The influence of the inclusion of dictionary data was evaluated for the Arahauco language due to the difficulty of its extraction. The results of the models using all data and excluding the dictionary generally show a low contribution from the dictionary. For Arhuaco, the best model achieved a BLEU of 7.72 and a chrF2 of 24.17 (spanish to target).

A similar evaluation was carried out for NASA. However, in this case, the more challenging extraction process was with letters from the constitution. Therefore, the model trained with all data and the model trained without constitutional letters were evaluated. The results in terms of BLEU suggest a little contribution from the letters, likely due to difficulties encountered in the alignment process due to enriched translations for the indigenous community.

In Wayuu, we evaluated the use of all data and without the dictionary, which corresponds to the largest fragment of the dataset. Contrary to what was expected, the inclusion of the dictionary did not have a positive impact on the results. In fact, the highest scores for the BLEU metric were achieved without the use of this data source, while the ChrF scores did not show a significant difference. Therefore, the most effective model was the one trained without the dictionary using the 1.3B model, which achieved a BLEU score of 15.37 and a ChrF2 score of 32.06.

Finally, for Inga, we found it to be the most challenging language, both for data collection and for the translation model. Our most successful model yielded a BLEU score of 1.71 and a ChrF2 score of 18.40, achieved without the utilization of the dictionary. However, the dictionary represented the largest dataset, and this experiment only considered 212 sentences. The built models can be downloaded and accessed from the repository, hoping to constitute a baseline for future efforts in

Language	Description	Sentences
Wayuunaiki	Dictionary (Amaya, 2021)	74583
Wayuunaiki	Bible (YouVersion, 2023)	6220
Wayuunaiki	Book (Álvarez, 2017)	534
Wayuunaiki	Book (Flórez et al., 2020)	467
Wayuunaiki	Book (Álvarez González, 2021)	229
Wayuunaiki	Book (Álvarez, 2016)	109
Wayuunaiki	Short story (Cue, 2012)	39
Wayuunaiki	Constitution (de Estudios de Lenguas Aborígenes, 1994)	37
Nasa Yuwe	Dictionary (originarios. Lenguas de América)	3729
Nasa Yuwe	Letters (Constitution) (de Estudios de Lenguas Aborígenes, 1994)	57
Nasa Yuwe	Common words (Constitution) (de Estudios de Lenguas Aborígenes, 1994)	53
Nasa Yuwe	Articles (Constitution) (de Estudios de Lenguas Aborígenes, 1994)	23
Arhuaco	Bible (para el Desarrollo de Pueblos Marginados)	5542
Arhuaco	Articles (Constitution) (de Estudios de Lenguas Aborígenes, 1994)	88
Arhuaco	Letters (Constitution) (de Estudios de Lenguas Aborígenes, 1994)	67
Arhuaco	Dictionary (Constitution) (de Estudios de Lenguas Aborígenes, 1994)	46
Arhuaco	Short stories (de la comunidad arhuaca de Jewrwa, 2014)	42
Inga	Dictionary (de Educación Inga de la Organización “Musu Runakuna”, 1997)	3048
Inga	Constitution (de Estudios de Lenguas Aborígenes, 1994)	212

Table 1: Parallel data collected for each language

these languages.

5 Conclusion and Future Work

The preservation of indigenous languages, encompassing their stories, wisdom, and traditions is instrumental in fostering cross-cultural understanding. However, working with low-resource languages such as these often presents unique challenges, particularly in regions like Colombia, which are teeming with linguistic diversity. We constructed a dataset of parallel data in four indigenous Colombian languages, and the resulting dataset is freely accessible and usable for future research projects. Additionally, we developed baseline translation models for each language pair. Our findings demonstrated that the NLLB 1.3B model excelled overall in comparison to the 600M model as expected. Also, a contrast emerged in the range of the BLEU score: from as low as 1.71 (Inga) to as high as 15.37 (Wayuu). Such a significant difference is attributable to the disparities in data volume, with Inga being the most challenging language. We also tested the influence of the inclusion in the training data of some of the sources in particular those that were challenging in the extraction phase. Although this project did not involve direct engagement with community members, future work will prioritize establishing connections with these communities to expand the dataset and evaluate translation systems more thoroughly. Our focus will be on incorporating additional Indigenous Colombian languages and exploring alternative models or architectures to potentially enhance translation outcomes.

6 Acknowledgements

Melissa Robles acknowledges the support of a UniAndes-GoogleDeepMind Scholarship 2023.

References

2012. *Putunkaa Serruma: Duérmete, pajarito blanco. Arrullos y relatos indígenas de cinco etnias colombianas*. Instituto Colombiano de Bienestar Familiar ICBF.

Ife Adebara, Muhammad Abdul-Mageed, and Miikka Silfverberg. 2021. Translating the unseen? yoruba-english mt in low-resource, morphologically-unmarked settings.

Rafael Jose Negrette Amaya. 2021. Osf spanish-wayuunaki.

El Moatez Billah-Nagoudi, Wei-Rui Chen, Muhammad Abdul-Mageed, and Hasan Cavusoglu. 2021. IndT5: A Text-to-Text Transformer for 10 Indigenous Languages. In *Proceedings of the AmericasNLP 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas*. Association for Computational Linguistics.

William Chen and Brett Fazio. 2021. Morphologically-guided segmentation for translation of agglutinative low-resource languages.

Comité de Educación Inga de la Organización “Musu Runakuna”. 1997. *Diccionario Inga*.

Centro Colombiano de Estudios de Lenguas Aborígenes. 1994. *Constitución Política de 1991 traducida a Lenguas Indígenas*.

Target Language	Data	NLLB - Model	Spanish - Target		Target - Spanish	
			BLEU	chrF2	BLEU	chrF2
Arhuaco	All	600M	6.00	23.04	8.29	33.28
Arhuaco	All	1.3B	7.72	**24.17**	8.29	32.98
Arhuaco	Without Dict	600M	7.28	23.06	8.29	32.95
Arhuaco	Without Dict	1.3B	**8.19**	23.08	**8.75**	**33.38**
Nasa	All	600M	2.65	18.18	4.02	**18.70**
Nasa	All	1.3B	2.70	**18.98**	2.22	17.92
Nasa	Without Letters	600M	3.02	15.99	**4.78**	18.21
Nasa	Without Letters	1.3B	**3.87**	15.68	3.19	16.60
Wayuu	All	600M	11.89	30.94	12.50	39.10
Wayuu	All	1.3B	13.81	**32.62**	14.19	40.51
Wayuu	Without Dict	600M	14.38	31.01	17.30	41.26
Wayuu	Without Dict	1.3B	**15.37**	32.06	**18.93**	**42.79**
Inga	All	600M	1.27	20.43	**3.08**	27.21
Inga	All	1.3B	0.78	**21.15**	1.08	27.38
Inga	Without Dict	600M	0.74	19.21	1.86	**30.00**
Inga	Without Dict	1.3B	**1.71**	18.40	1.10	29.30

Table 2: Scores (BLUE, chrF) on test partitions for all languages pairs per NLLB model

El Centro Colombiano de Estudios de Lenguas Aborígenes (C.C.E.L.A). 1994. *Estructuras sintácticas de la predicación: lenguas amerindias de Colombia*.

Autoridad Nacional de Gobierno Indígena – ONIC. 2015. *65 Lenguas Nativas de las 69 en Colombia son Indígenas*.

Alumnos de la comunidad arhuaca de Jewrwa. 2014. *Niwi úmuke pari ayunnuga, Cantando desde la Sierra*.

Dayana Fernandez, Jose Atencia, Ornela Gamboa, and Oscar Bedoya. 2013. Design and implementation of an web api for the automatic translation colombia's language pairs: Spanish-wayuunaiki case. In *Communications and Computing (COLCOM), 2013 IEEE Colombian Conference on*, pages 1–9.

Yasir Bustos Flórez, Norys Jiménez Pitre, and Delio Pontilus. 2020. *Joo'uya waashajaaiwa Wanüiki. Vamos a hablar nuestra lengua*. Institución Educativa Indígena No 4 de Maicao - sede Majayutpana.

Edward Gow-Smith and Danae Sánchez Villegas. 2023. Sheffield's submission to the AmericasNLP shared task on machine translation into indigenous languages. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 192–199, Toronto, Canada. Association for Computational Linguistics.

Nora Graichen, Josef Van Genabith, and Cristina España-bonet. 2023. Enriching Wayúunaiki-Spanish neural machine translation with linguistic information. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 67–83, Toronto, Canada. Association for Computational Linguistics.

Rebecca Knowles, Darlene Stewart, Samuel Larkin, and Patrick Littell. 2021. NRC-CNRC Machine Translation Systems for the 2021 AmericasNLP Shared Task. In *Proceedings of the AmericasNLP 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas*. Association for Computational Linguistics.

Tom Kocmi and Ondř ej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*. Association for Computational Linguistics.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. 2018. Challenges of language technologies for the indigenous languages of the Americas. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Manuel Mager, Arturo Oncevay, Annette Rios, Ivan Vladimir Meza Ruiz, Alexis Palmer, Graham Neubig, and Katharina Kann, editors. 2021. *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*. Association for Computational Linguistics, Online.

Oscar Moreno. 2021. The REPUcs' Spanish–Quechua Submission to the AmericasNLP 2021 Shared Task on Open Machine Translation. In *Proceedings of the AmericasNLP 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas*, pages 241–246. Association for Computational Linguistics.

Tan Ngoc Le and Fatiha Sadat. 2020. Revitalization of indigenous languages through pre-processing and neural machine translation: The case of Inuktitut. In *Proceedings of the 28th International Conference* *on Computational Linguistics*, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301, Taipei, Taiwan. Asian Federation of Natural Language Processing.

NLLBTeam. 2022. No language left behind: Scaling human-centered machine translation.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Pueblos originarios. Lenguas de América. Diccionario páez-español.

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

John E Ortega, Richard Alexander Castro-Mamani, and Jaime Rafael Montoya Samame. 2020b. Overcoming resistance: The normalization of an Amazonian tribal language. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, Suzhou, China. Association for Computational Linguistics.

Fundación para el Desarrollo de Pueblos Marginados. Visor biblia iku.

Shantipriya Parida, Subhadarshi Panda, Amulya Dash, Esau Villatoro-Tello, A. Seza Dogruöz, Rosa M. Ortega-Mendoza, Amadeo Hernández, Yashvardhan Sharma, and Petr Motlicek. 2021. Open Machine Translation for Low Resource South American Languages. In *Proceedings of the AmericasNLP 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas*. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting bleu scores.

Michael Przystupa and Muhammad Abdul-Mageed. 2019. Neural machine translation of low-resource and similar languages with backtranslation. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Leonard Richardson. 2007. Beautiful soup documentation. *April*.

Melissa Robles, Cristian A. Martínez, Juan C. Prieto, Sara Palacios, and Rubén Manrique. 2024. Preserving heritage: Developing a translation tool for indigenous dialects. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24. Association for Computing Machinery.

Geraldyn Otavo Rodríguez and Melissa Lizette Portilla Narváez. 2022. *Relatos ancestrales: una alternativa para la preservación de la identidad cultural oral del territorio Inga*.

Luz Marina Sierra, Carlos Alberto Cobos, Juan Carlos Corrales, and Tulio Rojas Curieux. 2015. Building a nasa yuwe language test collection. In *Computational Linguistics and Intelligent Text Processing*, pages 112–123, Cham. Springer International Publishing.

Luz Marina Sierra Martínez, Carlos Cobos, and Juan Corrales. 2016. Tokenizer adapted for the nasa yuwe language. *Computacion y Sistemas*, 20:355–364.

Luz Marina Sierra Martínez, Carlos Alberto Cobos, Juan Carlos Corrales Muñoz, Tulio Rojas Curieux, Enrique Herrera-Viedma, and Diego Hernán Peluffo-Ordóñez. 2018. Building a Nasa Yuwe Language Corpus and Tagging with a Metaheuristic Approach. *Computación y Sistemas*, 22:881 – 894.

Ray Smith, Daria Antonova, and Dar-Shyang Lee. 2009. Adapting the tesseract open source ocr engine for multilingual ocr. In *MOCR '09: Proceedings of the International Workshop on Multilingual OCR*, ACM International Conference Proceeding Series. ACM.

Elan van Biljon, Arnu Pretorius, and Julia Kreutzer. 2020. On optimal transformer depth for low-resource language translation.

Raúl Vázquez, Yves Scherrer, Sami Virpioja, and Jörg Tiedemann. 2021. The Helsinki submission to the AmericasNLP shared task. In *Proceedings of the AmericasNLP 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas*. Association for Computational Linguistics.

Rui Wang, Xu Tan, Renqian Luo, Tao Qin, and Tie-Yan Liu. 2021. A survey on low-resource neural machine translation.

YouVersion. 2023. Biblia en wayuu, https://www.bible.com/es/bible/1584/mat.1.guc.

José Álvarez. 2016. La conjugación del verbo en la lengua wayuu.

José Álvarez. 2017. Compendio de la gramática de la lengua wayuu.

José Ramón Álvarez González. 2021. Panorámica de la fonología y morfología de la lengua wayuu. 61.

Alp Öktem, Mirko Plitt, and Grace Tang. 2020. Tigrinya neural machine translation with transfer learning for humanitarian response.