Creating a Parallel Corpus for a Low-Resource, Indigenous Language: Muisca-to-Spanish

Aryan Gulati[†] **Leslie A. Moreno**[†] **Aditya Kumar** **Abhinav Gupta**

Center for AI in Society, University of Southern California, Los Angeles, CA, USA
*{aryangul, morenola, akumar35, abhinavg}@usc.edu

Abstract

We developed a corpus for the Muisca language, an extinct Chibchan language, using 6 unique documents from the Colección Mutis of the Colombian Institute of Anthropology and History. We parse and process these documents to develop the first parallel Muisca-Spanish corpus with 3,000. We also present preliminary results obtained after fine-tuning pre-trained machine translation models on our corpus. Overall, our creation of a Muisca corpus contributes to the preservation and revitalization of endangered languages, allowing future researchers to incorporate the newly created corpus in their linguistic research.

1 Introduction

When faced with limited resources and a restricted pool of unlabeled data, it becomes exceedingly difficult to achieve substantial progress in developing novel methods for language applications. An absence of parallel datasets for languages exacerbates this issue (Tonja et al., 2023). Numerous strategies have been proposed to enhance Neural Machine Translation (NMT) for languages with limited resources. One approach involves the utilization of monolingual data with parallel datasets (Tonja et al., 2023). Indigenous languages spoken in North and South America have seen limited development in Natural Language Processing (NLP). Furthermore, a significant number of these indigenous languages in the Americas are at risk of becoming extinct (Mager et al., 2018). In the Northeastern Andes of Colombia, the Muisca language developed, as early as the 4th Century AD (Izquierdo Pena, 2008). Despite enduring more than five centuries of marginalization, displacement, and the near loss of their cultural identity, the Muisca community has not only survived but has also managed to flourish up to the present day. Currently, the Muisca community of Suba, Bogota, Colombia are actively engaged in the process of re-indigenization by revitalizing their traditional knowledge and fostering the development of a new indigenous identity through ethnogenesis (Sanchez-Castaneda, 2020). In this paper, we introduce the first parallel corpus for machine translation tasks for the Muisca language. We also present preliminary results after fine-tuning a pre-trained sequence-to-sequence model on our corpus.

The Muisca language, also known as Chibcha and as muysccubun within Muisca (meaning "language of the people"), is part of the Chibchan language family. As Muisca was not a written language, research surrounding the linguistic nature of the language are specific to the dialect used in and around Santa Fe at the end of the 16th century and the beginning of the 17th century, whose basis was formed by Fray Pedro Simón (Izquierdo Pena, 2008). The language is categorized as extinct under the Agglomerated Endangerment Scale (AES) (Hammarström et al., 2018).

2 Background

2.1 Data Source

The data we used came from transcriptions of colonial manuscripts describing the vocabulary and grammar of the language, along with parallel Muisca-Spanish sentences. The dataset has been digitized from its original handwritten form and released by (Diana A. Giraldo, 2012). However, as shown in figure 1, the data still contains extraneous symbols and information that was removed when creating the corpus.

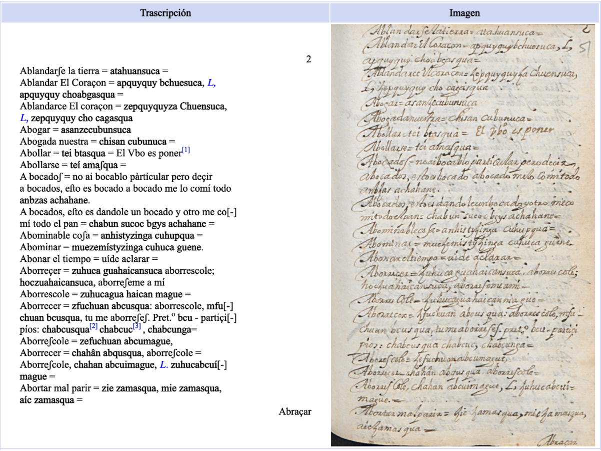
2.2 Subjects

Subjects in Muisca are not categorized by number or gender. An example can be seen below with the word for "dog" (Saravia, 2016).

[†]Equal contribution

Muisca	**English**
To	Dog, dogs
To cha ata	A male dog
To fuhuchá ata	A female dog
To cha mika	Three male dogs
To fuhuchá myhyká	Four female dogs

Table 1: Subjects in Muisca and English



2

Ablandarſe la tierra = **atahuansuca** =
Ablandar El Coraçon = **apquyquy bchuesuca,** *L,*
apquyquy choabgasqua =
Ablandarce El coraçon = **zepquyquyza Chuensuca,**
L, **zepquyquy cho cagasqua**
Abogar = **asanzecubunsuca**
Abogada nuestra = **chisan cubunuca** =
Abollar = **tei btasqua** = El Vbo es poner[1]
Abollarse = **teí amaſqua** =
A bocadoſ = no ai bocablo pàrtícular pero deçir
a bocados, eſto es bocado a bocado me lo comí todo
anbzas achahane.
A bocados, eſto es dandole un bocado y otro me co[-]
mí todo el pan = **chabun sucoc bgys achahane** =
Abominable coſa = **anhistyzinga cuhupqua** =
Abominar = **muezemístyzinga cuhuca guene.**
Abonar el tiempo = **uíde aclarar** =
Aborreçer = **zuhuca guahaicansuca** aborrescole;
hoczuahaicansuca, aborreſeme a mí
Aborrescole = **zuhucagua haican mague** =
Aborrecer = **zfuchuan abcusqua:** aborrescole, **mfu[-]**
chuan bcusqua, tu me aborreſeſ. Pret.º **bcu** - partiçi[-]
píos: **chabcusqua**[2] **chabcuc**[3] , **chabcunga**=
Aborreſcole = **zefuchuan abcumague,**
Aborrecer = **chahân abqusqua,** aborreſcole =
Aborreſcole, **chahan abcuimague,** *L.* **zuhucabcuí[-]**
mague =
Abortar mal parir = **zie zamasqua, mie zamasqua,**
aíc zamasqua =

Abraçar

Figure 1: A section of the transcript which was scraped to create the corpus (Diana A. Giraldo, 2012)

3 Methods and Results

3.1 Data Collection

In this work we present, to the best of our knowledge, the first parallel Muisca-Spanish corpus. We leveraged a web-scraping library, Trafilatura, to extract raw text from the aforementioned data source. We then extracted both the Muisca phrases and their corresponding Spanish translation from *Coleccion Mutis* documents. We extract the Muisca words, which are surrounded by "**" when scraped because all Muisca text is in bold. The corresponding translations follow each phrase and are separated by a "=" character. The consistent nature of this document allows for this simple algorithm to accurately parse all translations from the document.

3.2 Preliminary Modeling

After parsing the corpus, we created a 90/10 train-test split, and fine-tuned a T5-small model (Raffel et al., 2020) for one epoch on the resulting train split. We evaluated this model by reporting BLEU score on the test data. We also compare this performance to other models fine-tuned on parallel corpora for other endangered indigenous language in the section below.

Language	BLEU	Corpus Size
Quechua	1.58	11 MB
Wixarika	3.18	390 KB
Ashaninka	0.03	256 KB
Bribri	0.54	247 KB
Muisca	0.73	230 KB

Table 2: A comparison of the BLEU score of our Muisca corpus compared to models fine-tuned on similar corpora (Mager et al., 2021)

3.3 Results

The performance of our model was assessed using BLEU, a widely used metric for machine-translated text evaluation. Higher BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002) scores correspond to better performance.

Despite being trained on a relatively small dataset of approximately 3000 parallel sentence pairs, equating to 230 KBs, our model achieved a BLEU score of 0.73. While this value may not match up with well-documented languages, it surpasses the baseline scores of some other indigenous languages such as Bribri and Ashaninka, which have larger datasets.

4 Discussion

We have presented, to the best of our knowledge, the first parallel Muisca-Spanish corpus gathered from manuscripts transcribed by the Muysccubun group. Due to the constraints of our dataset, the model had limited exposure to Muyscan vocabulary during the training process. Consequently, the model is unable to provide high-quality, robust translations in most cases. However, with our focus on building an education aid, we hope to meaningfully leverage this corpus to support revitalization efforts, despite its small size.

Our future focus will lie in the creation of an educational tool designed to facilitate Muisca language learning for students. This tool will prioritize interactivity and natural conversation, providing an immersive learning experience. To enhance the learning process, we will integrate machine translation as an integral component, affording students the opportunity to engage with and internalize the intricacies of the Muisca language.

Acknowledgements

We would like to express our sincere gratitude to the Center for AI in Society for their invaluable support and resources, which have significantly contributed to the success of this research. Their commitment to advancing the field of artificial intelligence and its impact on society has been instrumental in shaping the direction of our work. Additionally, we extend our appreciation to the Muysccubun research group for generously providing access to their linguistic data, which served as a crucial component of our study. Their collaboration and dedication to preserving and sharing the Muisca language and culture have enriched the quality and authenticity of our research. We are also thankful to our advisors Dr. Bistra Dilkina and Dr. Robin Jia. This research would not have been possible without the collective efforts and contributions of all those mentioned above.

References

Diego F. Gómez Diana A. Giraldo. 2012. Colección mutis: Muisca (muysccubun). `http://coleccionmutis.cubun.org/index.php?title=Manuscrito_2924_BPRM_todo&printable=yes`. Accessed on March 2023.

Harald Hammarström, Thom Castermans, Robert Forkel, Kevin Verbeek, Michel A. Westenberg, and Bettina Speckmann. 2018. Simultaneous visualization of language endangerment and language description. *Language Documentation Conservation*, 12:359–392.

Manuel Arturo Izquierdo Pena. 2008. The muisca calendar: An approximation to the timekeeping system of the ancient native people of the northeastern andes of colombia. `http://archivos.cubun.org/FILES/images_CUBUN/9/9c/The_Muisca_Calendar.pdf`. Accessed on October 2023.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza. 2018. Challenges of language technologies for the indigenous languages of the americas. *arXiv.org*. Accessed on October 2023.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Rolando Coto-Solano, Alexis Palmer, Elisabeth Mager-Hois, Vishrav Chaudhary, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2021. Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Paola Andrea Sanchez-Castaneda. 2020. Memory in sacred places: The revitalization process of the muisca community. *Urban Planning*, 5(3S1):263+. Accessed on 23 Oct. 2023.

Facundo Manual Saravia. 2016. Curso de aproximación a la lengua chibcha o muisca (nivel 1).

Atnafu Lambebo Tonja, Christian Maldonado-sifuentes, David Alejandro Mendoza Castillo, Olga Kolesnikova, Noé Castro-Sánchez, Grigori Sidorov, and Alexander Gelbukh. 2023. Parallel corpus for indigenous language translation: Spanish-mazatec and Spanish-Mixtec. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 94–102, Toronto, Canada. Association for Computational Linguistics.