The Construction of a Mixe Variant Parallel Corpus

Ivan Meza[1], **Delfino Zacarías**[2], **Martha Elba Ramírez Andrés**[3]
Victoriano Santiago Cayetano[3], **Jonathan Santiago Antonio**[3]
Carlos Mena[4]
[1] Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas
Universidad Nacional Autónoma de México, Mexico
[2] Centro de Investigación en Matemáticas A.C., Mexico,
[3] Unión Nacional de Traductores Indígenas A. C., Mexico,
[4] Barcelona Supercomputing Center, Spain.
ivanvladimir@turing.iimas.unam.mx, delfino.zacarias@cimat.mx, carlos.hernandez@bsc.es

Abstract

We present the progress and challenges of constructing a Mixe-Spanish parallel corpus for Machine Translation. Mixe is a Mexican Indigenous language spoken by more than 100,000 speakers. In particular, we focus on the San Juan Guichicovic Mixe variant (`mir`). The resulting resource is available under an open research license (CC BY-NC-SA). It was created following a previous state-of-the-art methodology for Mexican indigenous languages. In this case, we used paid translators from the variant region. We present a baseline system.

Keywords: Indigenous languages, Machine translation, Mixe, Ayuuk, Spanish

1. Introduction

Mixe[1] is an indigenous language spoken primarily in the southern part of Mexico, particularly in the state of Oaxaca. It belongs to the Mixe–Zoquean language family, which also includes the Zoque languages spoken in neighboring areas. This family is of considerable interest to linguists due to its deep historical roots and its potential relationship with ancient Mesoamerican languages, including those associated with the Olmec civilization.

According to the most recent census data, there are approximately 139,760 speakers of Mixe in Mexico, making it the 16th most widely spoken indigenous language in the country (INEGI, 2020)[2]. This number, however, does not capture the entire speech community, as there is evidence of Mixe-speaking migrant populations in the United States (Asad and Hwang, 2019). Migration—both seasonal and permanent—has played an increasingly important role in the sociolinguistic dynamics of Mixe, leading to new forms of language contact, code-switching, and bilingual practices among younger generations.

Although Mixe is the name most commonly used in English and Spanish, native speakers refer to their language by various autonyms, such as *Ayuuk*, *Ayöök*, *Ayuujk*, *Ayüük*, or *Ayuhk*. These differences are not merely stylistic; they reflect the internal linguistic diversity within what is externally categorized as Mixe. In reality, Mixe is not a single homogeneous language but rather a continuum of closely related varieties that exhibit differences in phonology, lexicon, and syntax.

One linguistic classification identifies 11 distinct Mixe variants, one of which is now considered extinct (Wichmann, 2008). However, more recent scholarship challenges this strictly taxonomic view by introducing the notion of the *comunalect* (Valiñas Coalla, 2010). This concept emphasizes the relationship between language, territory, and community, proposing that each Mixe-speaking community sustains its own local linguistic variety that cannot be fully understood outside its cultural and geographical context. From this perspective, the Mixe linguistic landscape appears as a complex and dynamic ecosystem characterized by mutual intelligibility among neighboring communities and by strong local identities and linguistic autonomy.

A detailed demographic and linguistic analysis of these varieties can be found in Vásquez (2021), who provides an updated overview of Mixe's internal diversity and sociolinguistic situation. This body of research highlights the resilience of the Mixe language in the face of ongoing pressures from Spanish dominance, migration, and modernization, while also demonstrating the richness and adaptability of its multiple living forms.

This work presents our current efforts to create a parallel corpus of the Mixe variant of San Juan Guichicovi (SJGu). This variant forms part of a larger well-identified variant denominated the Isth-

[1] Link to Glottolog Mixe: https://glottolog.org/resource/languoid/id/mixe1286 last visited October 2025).

[2] Statistic taken from https://cuentame.inegi.org.mx/descubre/poblacion/hablantes_de_lengua_indigena/ (last visited October 2025).

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3456–3461
11–16 May 2026.

mus Mixe (mir)[3]. This variant is associated with the municipality of San Juan Guichicovi in the state of Oaxaca, located in the lowland region. Table 1 summarizes some aspects of this variant.

Variant	**SJGu**
Municipality inhabitants	29,802
Mixe speakers in municipality	18,000
Region	lowland
ISO-639-3 code	mir

Table 1: Some characteristics of the San Juan Guichicovy Mixe. Quantities extracted from (Vásquez, 2021).

2. Previous work

To exemplify the language, see the following phrase[4]:

(1) Jantim xyondaak ja koy jadu'un (*SJGu*)
The bunny became happy.

The first publicly available Mixe parallel corpus was released as part of the *JW300* multilingual corpus (Agić and Vulić, 2019). This massive resource originally contained over 300 languages and was built by automatically aligning translations extracted from the *Jehovah's Witnesses* website. Within that collection, the Mixe variant, *Coatlán Mixe* (mco), was included, which belongs to the lowland branch of the Mixe language continuum. Despite its potential usefulness for computational and linguistic research, the dataset was later taken down due to copyright concerns. Consequently, the Mixe portion of the JW300 corpus is no longer publicly accessible and is considered a deprecated research resource (Luccioni et al., 2022). This episode illustrates a significant challenge in developing linguistic resources for underrepresented languages: the reliance on third-party data can limit long-term availability and reproducibility.

Following JW300, a second Mixe parallel corpus was introduced, this time focusing as well on the *San Juan Guichicovi Mixe* (mir) variant (Zacarías Márquez and Meza Ruiz, 2021a,b). This resource was developed through a collaboration that aimed to provide parallel data between Mixe and Spanish for machine translation and linguistic analysis. The corpus includes a mixture of parallel and comparable segments, though its components differ in terms of licensing and accessibility. In some cases, texts in Spanish are freely distributed under open licenses. At the same time, their corresponding Mixe translations remain restricted due to copyright ownership or the absence of explicit authorization from the original authors. This mismatch highlights a recurring problem in the field of language documentation and computational linguistics: the coexistence of open and closed components within the same dataset, which complicates both redistribution and ethical reuse.

The first Mixe parallel corpus was part of the *JW300* corpus (Agić and Vulić, 2019). The variant included in this corpus was the *Coatlán Mixe* (mco), which is also a lowland variant. Unfortunately, this resource is no longer available due to copyright issues with the source of the text. Therefore, this resource should be considered deprecated (Luccioni et al., 2022). The second parallel corpus for Mixe was also focused on the SJGu variant (mir) Zacarías Márquez and Meza Ruiz (2021a,b). This corpus presents a mix of parallel and comparable corpora with varying accessibility and license statuses; in some cases, the phrases are available under an open license in one language but not in the other. This situation arises from the copyright holders.

To mitigate these issues and ensure the long-term openness and transparency of our own resources, one of our main efforts was to design a corpus that could be released under a permissive open license. This approach ensures that both the source and target texts can be freely used, modified, and redistributed within the research community, while respecting the rights of contributors and native speakers. To achieve this, we adopted and extended the methodology proposed in Mager et al. (2018), initially used for the construction of the *Wixarika corpus* Mager et al. (2018)[5].

In the Wixarika project, the corpus was built using public-domain materials—specifically, the Spanish versions of Grimm and Andersen fairy tales—as source texts. These stories were selected not only for their accessibility but also for their narrative diversity and simplicity, which make them suitable for controlled translation and linguistic study. Native speakers then produced high-quality translations into Wixarika, and the resulting bilingual corpus was released under a Creative Commons license (CC BY-NC-SA 4.0), allowing for non-commercial reuse and adaptation.

Our Mixe corpus follows a similar design philosophy, emphasizing ethical data collection, transparency, and reproducibility. However, we intro-

[3]Link to Glottolog Isthmus Mixe: https://glottolog.org/resource/languoid/id/isth1238 (last visited October 2025).

[4]This example is part of a short story collected and written by Albino Pedro Juan, a native speaker and preserver of the language. It is not part of the current corpus.

[5]Corpus available at: https://github.com/pywirrarika/wixarikacorpora (last visited October 2025)

duced two key differences. First, rather than relying solely on volunteer contributions, our project employed professional translators who are fluent in both Mixe and Spanish, ensuring consistency and linguistic accuracy across the dataset. Second, we established an explicit open licensing framework from the outset, covering all components—texts, annotations, and metadata—to avoid any ambiguity in future distribution or use. This strategy aims to provide a sustainable, fully open resource that supports future work in NLP of the Mixe language.

3. Creation of the Corpus

As widely discussed in the literature, multiple challenges are associated with developing computational resources for the Indigenous languages of the Americas (Mager et al., 2018). Among these, one of the most persistent and well-recognized issues is the scarcity of linguistic data. However, as discussed in the previous section, the problem extends beyond the mere absence of data. There are also substantial barriers to creating and sharing linguistic resources, particularly those arising from copyright and intellectual property restrictions.

In many cases, materials that could serve as valuable bilingual data—such as translations, educational materials, or literary works—are produced by specific individuals, often under copyright, which limits their redistribution. Similarly, the Spanish source texts from which these translations derive may themselves be protected, making it legally problematic to publish or share the corresponding Mixe.

The problem is further complicated by the specifics of national copyright legislation. In Mexico, for instance, literary and artistic works enter the public domain only after 100 years have elapsed since the author's death. Consequently, it is common to encounter published translations or original texts whose authors are no longer living, yet whose works remain protected by copyright. This extended duration—significantly longer than in many other jurisdictions—poses a significant barrier to the use of otherwise culturally and linguistically rich materials for research and documentation. While in some countries the principle of *fair use* may allow limited reproduction of copyrighted works for educational or research purposes, Mexican law does not provide a clear equivalent exception. Moreover, the specific application of fair use principles to linguistic data, such as parallel corpora or text mining for NLP, has not been tested in Mexican courts.

Beyond the legal dimension, there are also important ethical considerations. Many contemporary Indigenous authors and translators explicitly assert their moral and cultural rights over their linguistic productions, regardless of the legal framework. Their position often stems from a broader history of appropriation, where Indigenous knowledge, language, and expression have been extracted, digitized, and reused without proper acknowledgment or benefit to the communities of origin. As a result, even when the law might allow specific uses, Indigenous creators may prefer to retain complete control over their works to ensure they are represented and circulated under culturally appropriate conditions.

Therefore, researchers, linguists, and technologists must respect and uphold the wishes of Indigenous translators and authors. Ethical resource development for Indigenous languages must not only comply with legal standards but also align with community expectations and consent frameworks. Rather than treating Indigenous linguistic data as a freely extractable resource, it should be understood as a product of community knowledge and cultural labor. Only by building equitable collaborations and transparent licensing practices can the creation of linguistic resources proceed responsibly, ensuring both scientific advancement and the protection of Indigenous intellectual sovereignty.

Given this complex legal and ethical landscape, our project has adopted a series of strategies to create linguistic samples from Indigenous languages that respect both community rights and academic goals. Specifically, we have implemented the following three approaches:

1. Community engagement and donation-based openness. We first seek to establish direct communication with Indigenous translators and authors. During this process, we explain in accessible terms the relevance and potential impact of making such data available for research. We invite contributors to voluntarily donate their materials and to license them under an open license. This participatory approach recognizes contributors not merely as data providers but as co-creators of knowledge, thereby aligning the corpus development process with the principles of informed consent and collective authorship.

2. Professional translation under open agreements. The second strategy involves hiring professional translators. From the outset, these collaborators are informed that their translations will be distributed under an open license. This formal employment structure ensures that translators are compensated for their labor while maintaining complete transparency about data ownership and licensing terms. It also allows us to produce high-quality, consistent translations that meet linguistic and technical standards.

3. Controlled collection of restricted (closed)

data. While openness remains a guiding principle, the scarcity of available resources often necessitates working with data that cannot be freely shared due to copyright, consent, or institutional restrictions. In such cases, we still collect and curate closed data for internal use in system development, particularly when it is essential for training or evaluating computational models. However, we strictly refrain from redistributing or publishing this data. This strategy inevitably limits the reproducibility of our results, since other researchers cannot access identical training materials.

It is important to emphasize that adopting an open license under the first and second strategies does not necessarily imply complete openness in the unrestricted sense. Open permits can be tailored to include specific limitations—such as prohibiting commercial use or requiring attribution—depending on the contributors' wishes and the project's ethical stance. This nuanced understanding of openness allows for the integration of values such as reciprocity (Lévi-Strauss, 1944) and data sovereignty (Charter, 2016; Carroll et al., 2020).

In the current version of the Mixe corpus, we have followed the second strategy. We hired three translators to translate $2,120$ sentences from the $8,967$ Spanish phrases available in the Wixarika corpus; recall that these phrases belong to the Public Domain. We paid standard rates established by an NGO in the region, and the translators agreed to associate this corpus with an open license with commercial restrictions (CC BY-NC-SA 4.0) [6]. Table 2 shows a characterization of these two resources.

Variant	**SJGu**
Sentences	$2,119$
Token words	$16,680$
Type words	$4,881$
Min. length words	4
Max. length words	135

Table 2: Characteristics of corpora.

4. Baseline systems

With the collected resources, we train a baseline system based on the Transformer architecture (Vaswani et al., 2017). We use an *encoder-decoder* setting implemented using the JoeyNMT architecture (Kreutzer et al., 2019). We also use the same architecture reported there, consisting of the following settings:

- Number of layers: 3
- Number of heads: 4
- Input embedding dimensionality: 64
- Embedding dimensionality: 64
- Position-wise feed-forward: 128

In our experiments, we report BLEU scores and the *character error rate*. Table 3 summarizes our results using the two corpora:

Variant	**SJGu** (mir)
es $\rightarrow$ mir	
Previous BLEU	7.29
BLEU	8.03
Char error	0.16
mir $\rightarrow$ es	
Previous BLEU	5.82
BLEU	10.61
Char error	0.17

Table 3: Performance metrics for baseline systems, previous results from Zacarías Márquez and Meza Ruiz (2021b).

As shown, the current performance of the baseline systems for Mixe is poor, particularly compared to other language pairs with larger available resources. However, given the conditions under which these resources are collected, we believe they represent a milestone towards creating a more extensive resource.

5. Conclusions

We have presented the current state of the Mixe corpus, composed of two resources, each associated with a different variant of the Mixe language. We have constructed an open resource with the help of translators. Based on this resource, we presented two baseline systems—one for each variant—and showed that, for Mixe variant of San Juan Guichicovi (SJGu), including these newly translated phrases improves previously reported performance.

Further work will focus on continuing the collection of linguistic data for Mixe; we aim to incorporate more variants of the Mixe language. Additionally, we are interested in experimenting with a scenario in which we can complement the variants to create a single Machine Translation system for the Mixe language.

[6] This resource is available here: https://gitlab.com/l52mas/ayuuk-spanish (last visited October 2025.)

6. Acknowledgments

The authors thank , an NGO that helped us identify our excellent translators. We also thank the project which provided the funding to support the translators and associated students. Finally, the authors also thank for the computer resources provided through the .

6.1. Ethical considerations and limitations

At this stage, the corpus is too small to support the development of a fully functional machine translation (MT) system. The overall quantity of aligned sentence pairs remains insufficient to train models capable of achieving robust generalization or high-quality translation performance.

In addition to its small size, the corpus is also linguistically narrow, focusing specifically on the *San Juan Guichicovi Mixe* (SJGu) variant, which is itself classified under the broader Mixe variant `mir`. Even within the same lowland subgroup, variation in vocabulary, orthography, and idiomatic expressions may lead to substantial degradation in translation accuracy and fluency.

This limitation becomes even more pronounced when the scope is extended beyond the Mixe language cluster to other members of the Mixe–Zoquean family. As a result, any attempt to generalize the current model to broader linguistic contexts without proper retraining or variant-specific fine-tuning would likely result in reduced performance and misrepresentation of linguistic structures.

7. Bibliographical References

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Asad L Asad and Jackelyn Hwang. 2019. Migration to the united states from indigenous communities in mexico. *The ANNALS of the American Academy of Political and Social Science*, 684(1):120–145.

Stephanie Russo Carroll, Ibrahim Garba, Oscar L Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, Kay Raseroka, Desi Rodriguez-Lonebear, Robyn Rowe, et al. 2020. The care principles for indigenous data governance.

Māori Data Sovereignty Network Charter. 2016. Te mana raraunga.

INEGI. 2020. Censo de población y vivienda.

Julia Kreutzer, Jasmijn Bastings, and Stefan Riezler. 2019. Joey NMT: A minimalist NMT toolkit for novices. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 109–114, Hong Kong, China. Association for Computational Linguistics.

Claude Lévi-Strauss. 1944. Reciprocity and hierarchy. *American Anthropologist*, 46(2):266–268.

Alexandra Sasha Luccioni, Frances Corry, Hamsini Sridharan, Mike Ananny, Jason Schultz, and Kate Crawford. 2022. A framework for deprecating datasets: Standardizing documentation, identification, and communication. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 199–212.

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.

Leopoldo Valiñas Coalla. 2010. Historia lingüística: migraciones y asentamientos. Relaciones entre pueblos y lenguas. *Historia sociolingüística de México*, 1:97–160.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Julio César Gallardo Vásquez. 2021. Población y lengua mixe en el censo 2020.

Søren Wichmann. 2008. Om opdagelsen af et gränseoverskridende nyt sprog. In Jesper Nielsen and Mettelise Fritz Hansen, editors, *De mange veje til Mesoamerika: Hyldestskrift til Una Canger*, pages 63–80. Afdelingen for Indianske Sprog og Kulturer, Institut for Tværkulturelle og Regionale Studier, Københavns Universitet, København.

Delfino Zacarías Márquez and Ivan Vladimir Meza Ruiz. 2021a. Ayuuk-Spanish neural machine translator. In *Proceedings of the First*

Workshop on Natural Language Processing for Indigenous Languages of the Americas, pages 168–172, Online. Association for Computational Linguistics.

Delfino Zacarías Márquez and Ivan Vladimir Meza Ruiz. 2021b. Traductor automático neuronal ayuuk-español. *Research in Computer Science*, 150(5):10.

8. Language Resource References

Manuel Mager, Diónico Carrillo, and Ivan Meza. 2018. The Wixarika-Spanish parallel corpus. In *Latin American and Iberian Languages Open Corpora Forum 2018*.