NLP Progress in Indigenous Latin American Languages

Atnafu Lambebo Tonja [♦,♥], **Fazlourrahman Balouchzahi** [♦], **Sabur Butt** [♣], **Olga Kolesnikova** [♦], **Hector Ceballos**[♣], **Alexander Gelbukh** [♦], **Thamar Solorio** [♥,♠]

[♦] Instituto Politécnico Nacional, Mexico
[♣] Tecnológico de Monterrey Mexico
[♥] MBZUAI, Masdar City, UAE
[♠] University of Houston, Houston, USA

Abstract

The paper focuses on the marginalization of indigenous language communities in the face of rapid technological advancements. We highlight the cultural richness of these languages and the risk they face of being overlooked in the realm of Natural Language Processing (NLP). We aim to bridge the gap between these communities and researchers, emphasizing the need for inclusive technological advancements that respect indigenous community perspectives. We show the NLP progress of indigenous Latin American languages and the survey that covers the status of indigenous languages in Latin America, their representation in NLP, and the challenges and innovations required for their preservation and development. The paper contributes to the current literature in understanding the need and progress of NLP for indigenous communities of Latin America, specifically low-resource and indigenous communities in general.

1 Introduction

In a rapidly changing world with the advent of cutting-edge technologies, the emergence of Artificial Intelligence (AI), and the development of highly intelligent systems, it is essential to draw our attention to communities that seem to dwell in a different realm, detached from the whirlwind of global progress. These individuals, those who do not possess the knowledge or ability to speak the world's dominant languages, risk becoming increasingly marginalized and alienated from the ever-changing global landscape. But who are they, and why are they so crucial to the broader tapestry of human existence? These questions beckon us to delve deeper into the significance of preserving and promoting indigenous languages.

The people who speak indigenous languages represent a rich tapestry of cultural diversity (Nakagawa and Kouritzin, 2021). They are the bearers of unique worldviews, traditions, and ancestral knowledge that have been passed down through generations. These languages are not just a means of communication; they are vessels of history, folklore, and the wisdom of their communities. When we neglect or lose these languages, we forfeit a vital part of our collective human heritage. Each indigenous language holds the key to preserving a distinct cultural identity and the intangible yet invaluable assets that come with it.

Indigenous languages face an even greater degree of underrepresentation within the field of Natural Language Processing (NLP). Joshi et al. (2020) have highlighted a striking fact: over 88% of the world's languages, spoken by approximately 1.2 billion people, have been overlooked and continue to be neglected in the realm of language technologies. Blasi et al. (2022) have further emphasized that while linguistic NLP tasks such as morphology analysis exhibit a more inclusive approach to language diversity, user-facing NLP tasks like machine translation (MT) tend to be less accommodating. In today's information age, NLP techniques have become pervasive in their application on the internet, significantly shaping the content we encounter daily. Consequently, the absence of NLP technology support for endangered languages restricts their visibility to users. This unfortunate situation exacerbates the issue of linguistic marginalization, as regular exposure to a language is pivotal for its continued use and development. Conversely, most NLP research exhibits a bias toward languages with abundant resources, neglecting a wide array of linguistic typologies (Joshi et al., 2020) and often relying on the availability of extensive datasets. Incorporating endangered languages into NLP research can serve to assess the generalizability of NLP models (Bender, 2011) and promote the pursuit of universal and resource-efficient approaches. The lack of alignment between the under-representation of indigenous languages in

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 6972–6987

NLP and the pressing need for their inclusion motivates the creation of this study. Our aim is to bridge the gap between these Latin American communities and researchers, facilitating a more profound and reciprocal dialogue. By compiling insights, challenges, and innovations related to indigenous languages, we hope to provide researchers with a clearer roadmap, helping them prioritize what is vital and pressing. This paper seeks to offer a comprehensive overview of the current needs in the realm of NLP, ultimately fostering a more inclusive, collaborative approach to technological advancements that respect the unique perspectives and aspirations of indigenous Latin American people.

We summarize the contributions of this paper as follows:

- Report and discussion of the current state-of-the-art NLP research efforts for indigenous Latin American Languages.
- Analysis of challenges and opportunities in contributing to this space.
- We provide recommendations for researchers interested in indigenous Latin American Languages based on community feedback.

2 Overview of Indigenous Languages of Latin America

Indigenous peoples, who make up around 5% of the world's population, collectively preserve more than 7,000 distinct languages, showcasing their remarkable linguistic diversity (Nations, 2023). In Latin America, notable indigenous languages include Quechua, which traces its origins to what is now Ecuador and Peru with around ten million speakers; Guarani, which serves as the official language of Paraguay with over six million speakers; Nahuatl with approximately two million speakers in Mexico, Aymara is spoken by about two million people in Peru, Chile, and Bolivia, and Mapundung (Mapuche), an influential indigenous language with unclear origins in Chile and Argentina (Nations, 2023). Indigenous languages hold a significance extending well beyond mere markers of identity or community affiliation. They encapsulate the ethical values derived from ancestral wisdom, fostering a profound connection with the land, and stand as a vital cornerstone for preserving indigenous heritage and nurturing the aspirations of younger generations. This critical situation of indigenous languages is substantiated by data from (Nordhoff and Hammarström, 2012), revealing the existence of around 86 language families and 95 language isolates in the region, with an alarming number of these languages classified as endangered. The imminent threat of extinction primarily stems from state policies. Some governments actively engage in eradicating these languages, including extreme measures such as criminalizing their use, reminiscent of the historical context during the early colonial era in the Americas. Furthermore, the plight of indigenous languages is exacerbated by the denial of the existence of indigenous peoples in certain nations. This denial relegates their languages to mere dialects, affording them less recognition in comparison to national languages, thereby accelerating their decline (Degawan, 2019).

The linguistic panorama in Latin America showcases an extraordinary assortment of indigenous languages, encompassing diverse language families, isolates, and unclassified linguistic forms (Campbell et al., 2020). With an estimated tally of approximately 650 languages, including both existing and extinct varieties, this region stands as a testament to the multifaceted cultural and historical legacy it holds (Wikipedia, 2023). Distinctive language families like Mayan, Na-Dené, Algic, Arawakan, Tupian, Quechuan, Cariban, and Uto-Aztecan, among others, contribute significantly to this intricate linguistic fabric, each carrying profound significance in reflecting unique socio-cultural heritages throughout Latin America. This linguistic richness permeates seamlessly across the entire region, presenting a shared legacy of indigenous linguistic diversity. Current efforts aimed at documenting and preserving these languages emphasize their crucial role in upholding the deep cultural heritage and identity of indigenous communities across Latin America (Wikipedia, 2023).

The map presented in Figure 1 offers a visual representation of the indigenous languages that continue to be actively spoken across Latin America (Dockrill, 2023). A recent report from Statista (Statista, 2023) highlighted that approximately 7.5% of the Latin American population use native or indigenous languages as their mother tongues.

Figure 1: Modern Indigenous Languages in Latin America. Source https://adockrill.blogspot.com/2012/05/map-of-contemporary-latin-america.html

3 Overview of efforts in NLP research for indigenous Latin American languages

This section discusses the progress of NLP works for Indigenous Latin American languages. We used ACL Anthology[1] to search NLP works and searched the literature using keywords: Latin America, indigenous language of Latin America, indigenous language of (*) – * denotes Latin American languages.

3.1 Number of indigenous languages present in NLP research per country

Figure 2 shows the number of indigenous languages included in NLP research/work per country. Out of 33 countries in the Americas, we found that NLP works for indigenous languages spoken in 14 languages with five languages, and the rest of the countries have 1 to 4 languages of indigenous languages represented in NLP research, with 22 languages in total. Following Mexico, Brazil and Peru have 15 and 12 languages, respectively, represented in NLP research. Argentina comes next with nine languages, followed by Chile, Paraguay, and Bolivia, which each have five languages included, and the rest of the countries have 1 to 4 languages included.

The number of languages depicted in the figure

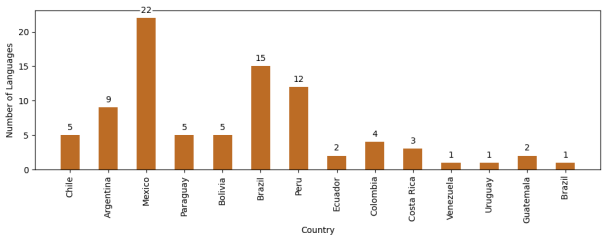


Figure 2: Number of indigenous languages present in NLP research per country

per country shows that most languages are left behind in NLP research in each country. For example, in Mexico, the government recognizes 68 indigenous languages. However, only around half (22) of the indigenous languages are represented in NLP research; in Peru, over 70 indigenous languages are spoken, but only 12 languages are present in NLP research.

3.2 Number of publications per language vs task

Figure 3 illustrates the number of publications available for each language vs tasks. From the figure, we observe that Quechua has more research and publications in different downstream NLP tasks, while languages like Shipibo-Konibo, Nahuatl, Aymara, and Guarani also have a good representation, covering more than 2 NLP tasks.

[1] https://aclanthology.org/

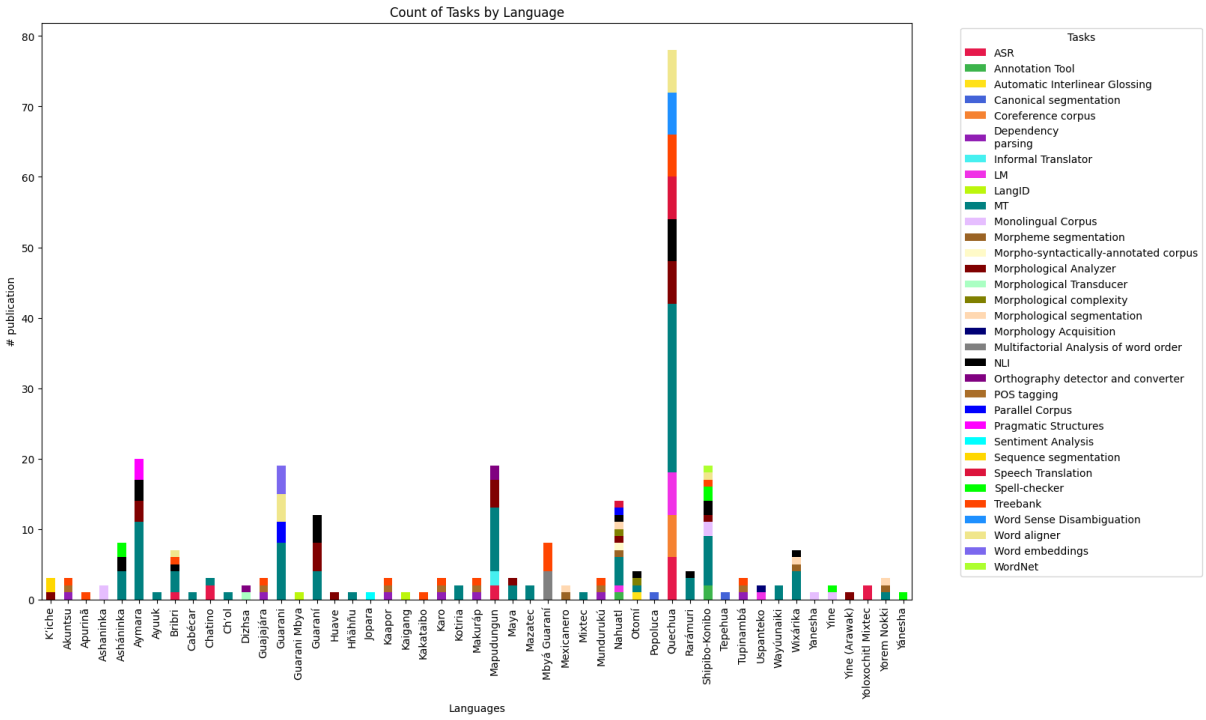


Figure 3: Number of publications per language vs tasks. We did not include publications, tasks, and languages from shared tasks like AmericasNLP in these statistics.

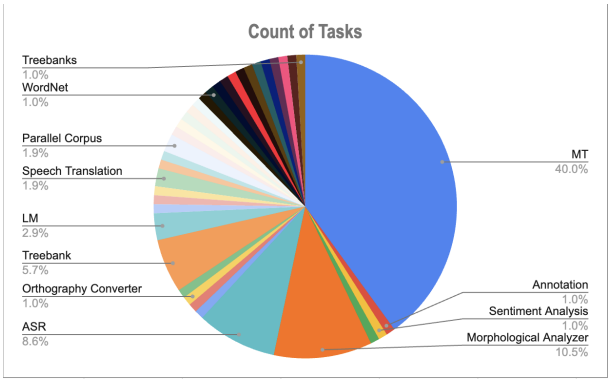


Figure 4: Total number of publications per task

Still, the majority of the languages present in papers only have one publication or NLP task. This clearly shows the under-representation of indigenous Latin American languages in NLP research. This highlights a critical gap in linguistic diversity in NLP research in indigenous languages of Latin America, emphasizing the need for increased focus and resources towards underrepresented languages to foster a more inclusive and equitable technological advancement.

3.3 Total number of publications per task

Figure 4 depicts the overall number of publications per task. As we can see from the figure, the machine translation task (MT) is the most researched NLP tasks, with ≈ 40% of overall publications, while Treebank, Morphological Analyzer, and Automatic Speech Recognition (ASR) have more than ≈ 5% publications.

We also found 12 MT shared task papers submitted in shared tasks organized by AmercasNLP in 2021, 2022, and 2023 that propose different methods to tackle MT problems in diverse indigenous Latin American languages.

3.4 Publication per year, venues and type

Figure 5 shows the overall research efforts of NLP publications for Latin American languages. As we can see from the figure, workshops are the dominant type of publication venue, surpassing conferences, and when we observe the specific venues, **AmericasNLP** and **LREC** are the leading venues over others. When we look at the years and number of publications, we see that the number of publications is increasing over time, and we can see dramatic changes starting in 2021. This clearly shows that the **AmericasNLP** workshop, which started in 2021, increased opportunities for researchers working in Latin American languages to publish their work. We observed more publications in LREC, a conference dedicated to publishing resources for different languages. We can observe the publica-

tions in top NLP conferences like ACL, EMNLP, EACL, etc.

4 Reflection from the community and researchers

To better understand the challenges of working in indigenous languages and recommend a suitable research direction for interested researchers, we conducted a survey with students, indigenous language speakers, and researchers working on indigenous languages.

Survey Design – Our survey focuses on two groups of people: (1) researchers who are familiar with NLP or/and working on indigenous Latin American languages and (2) the indigenous language community.

Participant Selection – We contacted researchers (including professors) and students from two research institutions in Latin American countries and researchers from regional NLP groups. We explained the aim of the survey and its output to them, and we received 27 responses in total from them. To survey indigenous language speakers and the general public, we used ClickWorker[2], a survey outsourcing platform to hire participants from all Latin American countries. We received 350 responses from them; all participants hired by ClickWorker are paid for their work.

Survey Questions – Our survey consists of 19 questions, from which nine questions ask about participants' general information like Gender, Age, Education, Country, indigenous languages they speak/are familiar with, and their occupation/experience. The remaining questions are about indigenous languages and NLP, challenges and needs, collaboration, and support. All the responses are anonymous and will not cause any harm to the participants. We received board approval from our institution before conducting the study. Figure 6 shows statistics of the survey responses.

We categorized our survey questions into four main topics as follows:- **1)** What challenges do researchers and indigenous communities encounter in NLP research for indigenous languages? **2)** What does the community need to save/preserve their languages? **3)** What opportunities do we currently have to develop different NLP tools for these languages? **4)** What future direction should be considered to increase the research works for these languages?

What challenges do researchers and indigenous communities encounter in NLP research for indigenous languages?

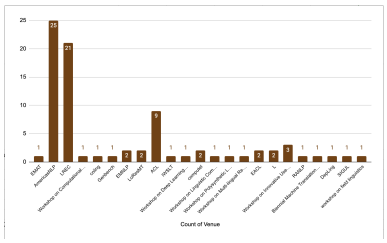
NLP for indigenous languages involves developing computational tools to process and understand these languages, which often have rich oral histories but limited written records, at the intersection of technology, linguistics, and cultural preservation (Ward, 2018). Preserving and revitalizing native languages is crucial for indigenous communities as it helps them maintain their cultural identity (Marmion et al., 2014). Moreover, it allows researchers to broaden the scope of NLP to include more diverse and inclusive languages, moving beyond the currently dominant ones in the field (Joshi et al., 2020).

Regarding the challenges faced by researchers and the indigenous community, we received the following responses: *Researchers* - highlighted the following key issues:- ***lack of access to resources***, ***lack of effort from the scientific community***, ***lack of indigenous language speaker involvement in the creation of the gold standard dataset***. In contrast, respondents from *indigenous community* - emphasized the following challenges: ***lack of inclusion of the indigenous community in research process***, ***little interest from the government***, ***the greatest inclusion of anglicisms and new technology.***

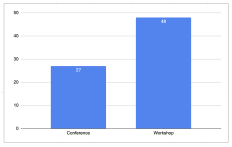
Based on the responses received, it is evident that there are multiple challenges at play when it comes to NLP research in indigenous Latin American languages. We observed different responses regarding challenges from researchers and communities. Researchers are more focused on the availability of resources and scientific community efforts, whereas responses from the indigenous communities are more focused on including the indigenous community in the research process and the government's interest. The foremost problem is the lack of resources, which implies that these communities need additional tools and support to become active participants in scientific research. This problem is further aggravated by the scientific community's need to be more proactive in incorporating indigenous knowledge and perspectives, suggesting an imbalanced approach to the research and development process.

Moreover, the under-representation of indigenous language speakers in creating key datasets

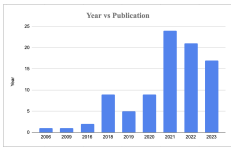
[2] clickworker.com



(a) Venues

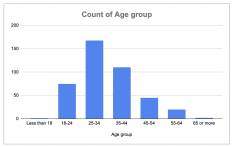


(b) Publication type

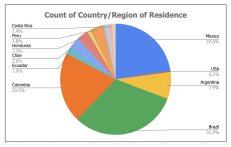


(c) Year vs Publication

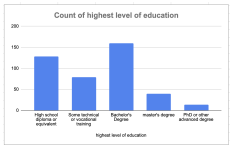
Figure 5: Publication per year, venues, and publication types



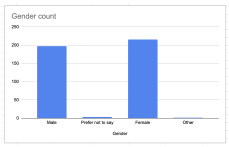
(a) Age distribution



(b) Country distribution



(c) Education statistics



(d) Gender distribution

Figure 6: Demographic statistics of the participants

is a significant barrier to inclusivity. This exclusion restricts the diversity and richness of the data and undermines the relevance and usefulness of research outcomes for indigenous populations.

These issues are further exacerbated by the lack of interest shown by governmental bodies, which indicates more extensive systemic neglect. The prevalence of Western concepts, the use of Anglicisms, and new technologies suggest a cultural overshadowing. It is necessary to give more consideration and importance to indigenous ways of knowing and doing. In science and technology, indigenous voices and needs are marginalized, leading to a homogenized and less inclusive approach to research and technological advancement.

What does the community need to save/preserve their languages?

Linguists advocate for the use of one's native language as an essential initial measure in safeguarding any linguistic heritage. Without such efforts, indigenous languages risk gradual erosion under the influence of dominant group languages. A critical inquiry arises: are our technological strides fostering increased utilization of native languages, thereby empowering speakers of these languages? Or are they reinforcing the dominance of widely spoken languages? Essentially, do these advancements contribute to the preservation of languages?

The insights from survey respondents highlighted the distinct yet interconnected roles of stakeholders, such as technology companies, government entities, and academic institutions. Technology companies emerge as crucial contributors, providing essential financial and technical resources. These resources can be subsidized for the indigenous communities to facilitate better outreach and access to technology. Governments, as outlined by respondents, are urged to play a proactive role by implementing incentive policies. The ***role of academic institutions*** is also pivotal in facilitating collaborative research. By bringing together a diverse range of experts and actively involving indigenous communities, these institutions address the unique challenges NLP poses for indigenous languages. Developing educational resources and training programs should also emerge as a strategy to promote a better understanding of NLP technologies within indigenous communities. Investors are encouraged to ***formulate and invest in AI applications that actively contribute to the preservation of indigenous culture***, stressing the inclusion of the indigenous community in the process and having them guide the priorities on what is needed for their communities and avoiding acculturation. In alliance with educational institutions, governments are advised to focus on providing broad education for indigenous communities, allocating resources to include them in technological advancements, thus preventing the disappearance of native languages. Institutions are positioned as key players in this comprehensive approach, with a responsibility to facilitate new technologies, provide economic assistance, and offer training and support for the holistic development and preservation of indigenous languages and cultures in Latin America.

Central to these suggestions is the imperative to

involve members from indigenous communities at all stages of development, ensuring representation and cultural sensitivity. The call for representative data collection highlighted the importance of ***diverse datasets*** to train NLP models effectively. ***Transparency, documentation, and ethical audits*** were recommended practices by the respondents to address algorithmic bias and promote accountability. The survey emphasized respecting copyright and cultural rights and the development of translation tools that consider linguistic and cultural nuances.

Cultural sensitivity and understanding are paramount in integrating NLP technologies within indigenous contexts. Recommendations included avoiding generalizations, conducting extensive studies on indigenous peoples to inform technology development, and better understanding their culture for a respectful approach. Respondents emphasized profound respect for indigenous peoples' customs, thoughts, and traditions, advocating for active listening and keeping NLP applications separate from indigenous culture to prevent misappropriation. Additionally, ***effective communication and collaboration*** with indigenous communities were highlighted as crucial. This involved direct contact with leaders and representatives, leveraging online resources for independent dissemination, and actively involving communities in projects and initiatives. Open-source initiatives were encouraged for transparency, and integrating indigenous community members in projects was recommended for genuine inclusion.

Respect and privacy were repeated subjects in the responses, emphasizing the importance of establishing agreements that honored cultural boundaries and avoided privacy invasion. Recommendations included respecting people's decisions, adapting to community preferences, and conducting activities at the community's pace and within their cultural context. This collective emphasis on respect and cultural understanding forms a crucial foundation for ethical and meaningful engagements with indigenous communities. The respondents also addressed the importance of ***government and policy involvement*** for the welfare of indigenous communities. Recommendations included government intervention, human rights organizations, public policies, and union work. The comments stressed the necessity of government support for effective and sustainable initiatives benefiting indigenous communities.

A multifaceted approach to ***education and awareness*** is recommended, with an emphasis on comprehensive awareness programs, education, changing paradigms, and targeted initiatives to disseminate accurate information. Teaching the language and creating awareness through meaningful work have been suggested as effective ways to instill mutual respect among individuals, with a specific focus on educational workshops as practical tools for achieving these goals. Practical steps and actions were proposed, including hiring native speakers, direct contact with communities, on-site research, and putting indigenous languages on platforms. The implementation of compensation systems, limits, and sanctions were suggested for fair collaboration and accountability. Adapting NLP to the reality of indigenous communities was recommended for relevant applications, supporting and preserving indigenous languages and cultures.

The survey comments also highlighted strategic recommendations for ***research and development***, emphasizing increased investment, detailed studies, regularizing efforts with protocols, and implementing oversight and monitoring systems. Improving resources, expanding language databases, and investing in the preservation of languages and dialects were key strategies. The survey underscored the importance of contributing resources for comprehensive studies, utilizing advanced technologies, and staying current through updated research and surveys. The overarching theme was a strategic and well-supported effort to advance research and understanding of indigenous languages and cultures.

What future direction should be considered to increase the research works for these languages?

Despite the challenges discussed earlier in this section, promising opportunities exist to promote NLP research and tools for these languages. Based on Figure 4, it is evident that a substantial portion of researchers in these languages are exploring MT. However, there remains a crucial need for researchers to focus on tasks that have received less attention. This requires a thorough examination of the current status and necessary advancements for each language, including the identification of tasks present or absent within their respective linguistic contexts. Drawing from previous efforts to support extremely low-resource languages (Maldonado-Sifuentes et al.; Llitjós et al.,

2005; Gasser, 2011; Mager et al., 2021), we present a discussion on certain tasks and assess the languages in which these tasks can be effectively implemented.

- Machine Translation (MT) - Most of NLP research in indigenous Latin American languages centers on MT, as indicated in Figure 3. While certain languages like Quechua, Shipibo-Konibo, Nahuatl, Aymara, Guarani, Mapudungun, and Wixarika have garnered considerable attention in MT research, numerous others, such as Yánesha, Yoloxochitl Mixtec, Tepehua, Otomi, Popoluca, remain relatively unexplored, lacking both research initiatives and associated tools within this domain.

- Morphology - is crucial for understanding the linguistic intricacies of Indigenous Latin American languages, which are esteemed for their cultural richness and offer valuable insights into human linguistic diversity. Despite their significance, these languages often lack resources for morphological research and technological development. Researchers have initiated various morphology-related tasks, such as morpheme segmentation, annotated corpora, analyzers, and transducers, particularly focusing on languages like Quechua and Nahuatl. However, there remains a significant gap in tools and research for many other languages, as illustrated in Figure 3. Some of these languages are Apurinã, Asháninka, Bribr, Cabécar, Jopara, etc.

- Speech Recognition and Translation - Figure 3 shows that several efforts have been made for ASR and speech translation for Quechua, Yoloxochitl Mixtec, and Wixarika, but the majority of languages such as Raramuri, k'iche, Akuntsu, Apurinã, Aymara, etc. still need tools and researches in this domain.

- Named Entity Recognition (NER) and Part-of-Speech (POS) tagging - represent two essential NLP tasks for which there is a notable scarcity of tools and research dedicated to indigenous languages. As illustrated in Figure 3, the majority of languages, such as Quechua, Shipibo-Konibo, Guarani, Kakataibo, Jopara, among others, suffer from a lack of resources for these specific tasks.

- Treebank - is a parsed text corpus that annotates syntactic or semantic sentence structure (Taylor et al., 2003), are fundamental for the development of various NLP tools in any low-resource languages. Lexical databases are crucial in enhancing NLP systems by offering a foundation for understanding a language's lexicon. This necessity is highlighted by the statistics presented in Figure 3, where Shipibo-Konibo stands as the sole exception, underscoring the need for such resources for these tasks.

- Language Identification (LI) - is another primary requirement for facilitating the development of language technologies for these languages. Figure 3 shows that LI can be an ideal initiative in NLP tasks for languages such as Mazatec, Maya, Mixtec, Kaapor, Popoluca, and many more.

- Natural Language Inference (NLI), as depicted in Figure 3, along with machine translation (MT) and speech recognition tasks, possesses some resources, albeit limited, in languages like Quechua, Asháninka, Aymara, Bribri, Guarani, Nahuatl, and Shipibo-Konibo, while other languages such as Mapudungun, Mexicanero, Yorem Nokki, Yine, Popoluca, among others, lack any resources for NLI.

- Word embeddings - as indicated by the statistics in Figure 3, are notably absent for these languages, with the exception of Guarani, despite their significance and the demonstrated value as essential tools for any language.

Country vs frequency of responses

Figure 7 shows the frequency of responses vs the country of respondents for the first question: *What challenges do researchers and indigenous communities encounter in NLP research for indigenous languages?* As we can observe from the figure, the responses from participants from different countries reflect common challenges. When we see the frequency of the responses that **Limited resource, Lack of access, Language/culture preservation and Lack of interest from government** have a higher frequency than the others.

Limited Resources - Latin American indigenous communities face a multitude of challenges in adopting NLP and AI, primarily due to limited resources. Financial and technological constraints

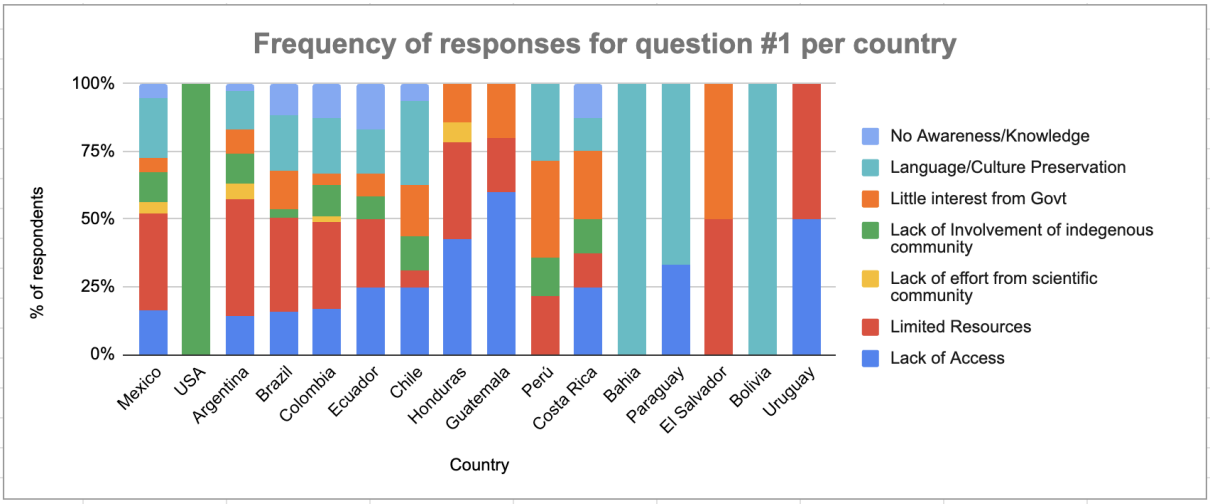


Figure 7: Frequency of response distribution across different Latin American countries

hinder their access to advanced technologies, while the lack of digitized linguistic resources threatens language preservation. Limited data availability and difficulties in accessing modern technology further impede progress. Economic factors, including insufficient investment, pose significant barriers to technology adoption.

Lack of Access - manifests in various forms, including limited resources for technological infrastructure and education. Many communities struggle with inadequate internet connectivity, hindering their ability to utilize modern technologies effectively. Additionally, geographic isolation and economic constraints exacerbate the issue, preventing access to essential tools and resources. Lack of education and training further compound the problem, as does the preservation of cultural identity, which can sometimes conflict with the adoption of new technologies. We added a detailed explanation about this response in Section B.

5 Takeaways

5.1 Opportunities for NLP researchers and indigenous communities

As discussed in Section 3, we observed promising works progress for Latin American indigenous languages, especially a dramatic increase in publications for those languages in recent years. We observed the following takeaways from Section 3.

Indigenous languages – We observe the promising number of publications for languages like Nahuatl, Quechua, Shipibo-Konibo, Bribri, Mapudungun, Aymar, and Wixarika in different NLP domains.

NLP tasks – We observe different works done for indigenous languages of Latin America; from those NLP tasks, MT is one of the leading tasks for those languages. We also observe high-level NLP tasks like the Pre-trained language model for Quechua and low-level NLP tasks like spell checker, morphological analyzer, and Treebank for most languages.

Community impact – We also observe the impact of communities like AmericasNLP by creating environments for researchers to present and publish their works, which Figure 5 notices.

5.2 Community reflection

Feedback from researchers and indigenous communities pointed out interesting points regarding challenges, community needs, and future directions. Researchers working on these languages should include the community in every research process to gain the community's trust and obey the community's customs.

6 Conclusion

In this work, we explore the research progress in indigenous Latin American languages, and we conduct a survey study to identify the challenges when working on these languages, as well as the needs and future directions of the research in NLP research. We hope this study will show some direction for researchers interested in indigenous Latin American languages and low-resource indigenous languages in general.

7 Limitations

This study is limited to Latin American languages and NLP research efforts on those languages. The analysis focused only on showing the research efforts in the area of NLP; we did not include a detailed literature review for downstream NLP tasks. For the overview, we only used ACL Anthology papers. The summary of the responses from the community and researchers may or may not be generalized to a broader society or languages. As discussed in the paper, the motivation of this study is to show the works that have been done for indigenous Latin American languages and to understand the challenges and needs of the community and the scientific community.

8 Ethical consideration

We conducted the survey after board approval from our institution and agreement from the respondents, who agreed to participate in the study. Participants also consented to the use of their data and responses for research work and publication. All the survey participants are anonymous, and we did not include questions that expose their anonymity. The Clickworker platform compensates all community members who participated in this survey.

References

Oliver Adams, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird, and Alexis Michaud. 2019. Evaluating phonemic transcription of low-resource tonal languages for language documentation. In *11th International Conference on Language Resources and Evaluation, LREC 2018*, pages 3356–3365. European Language Resources Association (ELRA).

Marvin M Agüero-Torales, David Vilares, and Antonio G López-Herrera. 2021. On the logistical difficulties and findings of jopara sentiment analysis. *arXiv preprint arXiv:2105.02947*.

Nouman Ahmed, Natalia Flechas Manrique, and Antonije Petrovi. 2023. Enhancing spanish-quechua machine translation with pre-trained models and diverse data sources: Lct-ehu at americasnlp shared task. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 156–162.

Cristian Ahumada, Claudio Gutierrez, and Antonios Anastasopoulos. 2022. Educational tools for mapuzugun. *arXiv preprint arXiv:2205.10411*.

Carlo Alva and Arturo Oncevay. 2017. Spell-checking based on syllabification and character-level graphs for a peruvian agglutinative language. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 109–116.

Jonathan D Amith, Jiatong Shi, and Rey Castillo Garcia. 2021. End-to-end automatic speech recognition: Its impact on the workflow for documenting yoloxóchitl mixtec. In *First Workshop on NLP for Indigenous Languages of the Americas. 11 June 2021. https://www. aclweb. org/anthology/2021. americasnlp-1.8. pdf*.

Emily M Bender. 2011. On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6.

Damián Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505.

Frederic Blum. 2022. Evaluating zero-shot transfers and multilingual models for dependency parsing and pos tagging within the low-resource language family tupían. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 1–9.

Marcel Bollmann, Rahul Aralikatte, Héctor Murrieta Bello, Daniel Hershcovich, Miryam de Lhoneux, and Anders Søgaard. 2021. Moses and the character-based random babbling baseline: Coastal at americasnlp 2021 shared task. In *First Workshop on Natural Language Processing for Indigenous Languages of the Americas, June 11, 2021*, pages 248–254.

Gina Bustamante, Arturo Oncevay, and Roberto Zariquiey. 2020. No data to crawl? monolingual corpus creation from pdf files of truly low-resource languages in peru. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2914–2923.

Lyle Campbell, Thiago Chacon, and John Elliott. 2020. Contact and south american languages. *The Handbook of Language Contact*, pages 625–648.

Paulo Cavalin, Pedro Domingues, Julio Nogima, and Claudio Pinhanez. 2023. Understanding native language identification for brazilian indigenous languages. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 12–18.

Malgorzata Ćavar, Damir Ćavar, and Hilaria Cruz. 2016. Endangered language documentation: Bootstrapping a chatino speech corpus, forced aligner, asr. In *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)*, pages 4004–4011.

Wei-Rui Chen and Muhammad Abdul-Mageed. 2022. Improving neural machine translation of indigenous languages with multilingual transfer learning. *arXiv preprint arXiv:2205.06993*.

William Chen and Brett Fazio. 2021. Morphologically-guided segmentation for translation of agglutinative low-resource languages. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pages 20–31.

Luis Chiruzzo, Pedro Amarilla, Adolfo Ríos, and Gustavo Giménez-Lugo. 2020. Development of a guarani-spanish parallel corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2629–2633.

Luis Chiruzzo, Santiago Góngora, Aldo Alvarez, Gustavo Giménez-Lugo, Marvin Agüero-Torales, and Yliana Rodríguez. 2022. Jojajovai: A parallel guarani-spanish corpus for mt benchmarking. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2098–2107.

Johanna Cordova and Damien Nouvel. 2021. Toward creation of ancash lexical resources from ocr. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 163–167.

Rolando Coto-Solano, Sharid Loáiciga, and Sofía Flores-Solórzano. 2021. Towards universal dependencies for bribri. In *Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021)*, pages 16–29.

Ona De Gibert, Ral Vzquez, Mikko Aulamo, Yves Scherrer, Sami Virpioja, and Jrg Tiedemann. 2023. Four approaches to low-resource multilingual nmt: The helsinki submission to the americasnlp 2023 shared task. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 177–191.

Jorge Antonio Leoni de León. 2010. Computational linguistics in costa rica: an overview. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, pages 40–45.

Minnie Degawan. 2019. Indigenous languages: Knowledge and hope. *The UNESCO Courier, Many Voices, One World*, 1:2220–2293.

Alan Dockrill. 2023. Map of contemporary latin america.

CM Downey, Shannon Drizin, Levon Haroutunian, and Shivin Thukral. 2021. Multilingual unsupervised sequence segmentation transfers to extremely low-resource languages. *arXiv preprint arXiv:2110.08415*.

Mingjun Duan, Carlos Fasola, Sai Krishna Rallabandi, Rodolfo M Vega, Antonios Anastasopoulos, Lori Levin, and Alan W Black. 2019. A resource for computational experiments on mapudungun. *arXiv preprint arXiv:1912.01772*.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Meza-Ruiz, et al. 2021. Americasnli: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. *arXiv preprint arXiv:2104.08726*.

Abteen Ebrahimi, Arya D McCarthy, Arturo Oncevay, Luis Chiruzzo, John E Ortega, Gustavo A Giménez-Lugo, Rolando Coto-Solano, and Katharina Kann. 2023. Meeting the needs of low-resource languages: The value of automatic alignments via pretrained models. *arXiv preprint arXiv:2302.07912*.

Isaac Feldman and Rolando Coto-Solano. 2020. Neural machine translation models with back-translation for the extremely low-resource indigenous language bribri. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3965–3976.

Michael Gasser. 2011. Computational morphology and the teaching of indigenous languages. In *Indigenous Languages of Latin America Actas del Primer Simposio sobre Enseñanza de Lenguas Indígenas de América Latina*, page 52.

Michael Ginn and Alexis Palmer. 2023. Robust generalization strategies for morpheme glossing in an endangered language documentation context. *arXiv preprint arXiv:2311.02777*.

Santiago Góngora, Nicolás Giossa, and Luis Chiruzzo. 2021. Experiments on a guarani corpus of news and social media. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 153–158.

Santiago Góngora, Nicolás Giossa, and Luis Chiruzzo. 2022. Can we use word embeddings for enhancing guarani-spanish machine translation? In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 127–132.

Edward Gow-Smith and Danae Sánchez Villegas. 2023. Sheffield's submission to the americasnlp shared task on machine translation into indigenous languages. *arXiv preprint arXiv:2306.09830*.

Nora Graichen, Josef van Genabith, and Cristina Espaa-bonet. 2023. Enriching wayunaikispanish 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.

Ximena Gutierrez-Vasques and Victor Mijangos. 2018. Comparing morphological complexity of spanish, otomi and nahuatl. *arXiv preprint arXiv:1808.04314*.

Ximena Gutierrez-Vasques, Gerardo Sierra, and Isaac Hernandez Pompa. 2016. Axolotl: a web accessible parallel corpus for spanish-nahuatl. In *Proceedings of the Tenth International Conference on Lan-*

guage Resources and Evaluation (LREC'16), pages 4210–4214.

Marcelo Yuji Himoro and Antonio Pareja Lora. 2022. Preliminary results on the evaluation of computational tools for the analysis of quechua and aymara. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5450–5459.

Petr Homola and Matt Coler. 2013. Pragmatic structures in aymara. In *Proceedings of the second international conference on dependency linguistics (DepLing 2013)*, pages 98–107.

Alex Jones, Rolando Coto-Solano, and Guillermo González Campos. 2023. Talamt: Multilingual machine translation for cabécar-bribri-spanish. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 106–117.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293.

Katharina Kann, Abteen Ebrahimi, Kristine Stenzel, and Alexis Palmer. 2022. Machine translation between high-resource languages in a language documentation setting. In *Proceedings of the first workshop on NLP applications to field linguistics*, pages 26–33.

Katharina Kann, Manuel Mager, Ivan Meza-Ruiz, and Hinrich Schütze. 2018. Fortification of neural morphological segmentation models for polysynthetic minimal-resource languages. *arXiv preprint arXiv:1804.06024*.

Angelika Kiss and Guillaume Thomas. 2019. Word order variation in mbyá guaraní. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 121–129.

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 First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 224–233.

Anastasia Kuznetsova and Francis Tyers. 2021. A finite-state morphological analyser for paraguayan guaraní. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 81–89.

Lori Levin. 2009. Adaptable, community-controlled, language technologies for language maintenance. In *Proceedings of the 13th Annual conference of the European Association for Machine Translation*.

Ariadna Font Llitjós, Roberto Aranovich, and Lori Levin. 2005. Building machine translation systems for indigenous languages. In *Second Conference on the Indigenous Languages of Latin America (CILLA II), Texas, USA*.

Manuel Mager, Rajat Bhatnagar, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2023a. Neural machine translation for the indigenous languages of the americas: An introduction. *arXiv preprint arXiv:2306.06804*.

Manuel Mager, Özlem Çetinoğlu, and Katharina Kann. 2020. Tackling the low-resource challenge for canonical segmentation. *arXiv preprint arXiv:2010.02804*.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza. 2018a. Challenges of language technologies for the indigenous languages of the americas. *arXiv preprint arXiv:1806.04291*.

Manuel Mager, Elisabeth Mager, Katharina Kann, and Ngoc Thang Vu. 2023b. Ethical considerations for machine translation of indigenous languages: Giving a voice to the speakers. *arXiv preprint arXiv:2305.19474*.

Manuel Mager, Elisabeth Mager, Alfonso Medina-Urrea, Ivan Meza, and Katharina Kann. 2018b. Lost in translation: Analysis of information loss during machine translation between polysynthetic and fusional languages. *arXiv preprint arXiv:1807.00286*.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios Gonzales, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, et al. 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.

Manuel Mager, Arturo Oncevay, Elisabeth Mager, Katharina Kann, and Ngoc Thang Vu. 2022. Bpe vs. morphological segmentation: A case study on machine translation of four polysynthetic languages. *arXiv preprint arXiv:2203.08954*.

Diego Maguiño-Valencia, Arturo Oncevay, and Marco Antonio Sobrevilla Cabezudo. 2018. Wordnet-shp: Towards the building of a lexical database for a peruvian minority language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Christian E Maldonado-Sifuentes, Jason Angel, Grigori Sidorov, and Alexander Gelbukh. Towards the inclusion of indigenous languages in mainstream nlp research: Challenges, relevance, and a roadmap proposal. *Procesamiento De Lenguaje Natural Para Las Lenguas Indígenas*, page 161.

Doug Marmion, Kazuko Obata, and Jakelin Troy. 2014. *Community, identity, wellbeing: the report of the*

Second National Indigenous Languages Survey. Australian Institute of Aboriginal and Torres Strait Islander Studies Canberra.

Delfino Zacarías Márquez and Ivan Meza-Ruiz. 2021. Ayuuk-spanish neural machine translator. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 168–172.

Diego Barriga Martínez, Victor Mijangos, and Ximena Gutierrez-Vasques. 2021. Automatic interlinear glossing for otomi language. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 34–43.

Gerardo Sierra Martínez, Cynthia Montaño, Gemma Bel-Enguix, Diego Córdova, and Margarita Mota Montoya. 2020. Cplm, a parallel corpus for mexican languages: Development and interface. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2947–2952.

Rodolfo Mercado-Gonzales, José Pereira-Noriega, Marco Antonio Sobrevilla Cabezudo, and Arturo Oncevay. 2018. Chanot: An intelligent annotation tool for indigenous and highly agglutinative languages in peru. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Christian Monson, Ariadna Font Llitjós, Roberto Aranovich, Lori Levin, Ralf Brown, Eric Peterson, Jaime Carbonell, and Alon Lavie. 2006. Building nlp systems for two resource-scarce indigenous languages: Mapudungun and quechua. *Strategies for developing machine translation for minority languages*, page 15.

Héctor Erasmo Gómez Montoya, Kervy Dante Rivas Rojas, and Arturo Oncevay. 2019. A continuous improvement framework of machine translation for shipibo-konibo. In *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, pages 17–23.

Taesun Moon, Katrin Erk, and Jason Baldridge. 2009. Unsupervised morphological segmentation and clustering with document boundaries. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 668–677.

Oscar Moreno. 2021. The repu cs'spanish–quechua submission to the americasnlp 2021 shared task on open machine translation. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 241–247.

El Moatez Billah Nagoudi, Wei-Rui Chen, Muhammad Abdul-Mageed, and Hasan Cavusogl. 2021. Indt5: a text-to-text transformer for 10 indigenous languages. *arXiv preprint arXiv:2104.07483*.

Satoru Nakagawa and Sandra Kouritzin. 2021. Identities of resignation: Threats to indigenous languages from neoliberal linguistic and educational practices. *Journal of Language, Identity & Education*, 20(5):296–310.

United Nations. 2023. Background - international day of the world's indigenous peoples.

Sebastian Nordhoff and Harald Hammarström. 2012. Glottolog/langdoc: Increasing the visibility of grey literature for low-density languages. In *the 8th international conference on language resources and evaluation [lrec 2012]*, pages 3289–3294. ELRA.

Arturo Oncevay. 2021. Peru is multilingual, its machine translation should be too? In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 194–201.

Arturo Oncevay, Gerardo Cardoso, Carlo Alva, César Lara Ávila, Jovita Vásquez Balarezo, Saúl Escobar Rodríguez, Delio Siticonatzi Camaiteri, Esaú Zumaeta Rojas, Didier López Francis, Juan López Bautista, et al. 2022. Schaman: Spell-checking resources and benchmark for endangered languages from amazonia. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 411–417.

John Ortega and Krishnan Pillaipakkamnatt. 2018. Using morphemes from agglutinative languages like quechua and finnish to aid in low-resource translation. In *Proceedings of the AMTA 2018 workshop on technologies for MT of low resource languages (LoResMT 2018)*, pages 1–11.

John E Ortega, Rodolfo Zevallos, and William Chen. 2023. Quespa submission for the iwslt 2023 dialect and low-resource speech translation tasks. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 261–268.

Elizabeth Pankratz. 2021. qxoref 1.0: A coreference corpus and mention-pair baseline for coreference resolution in conchucos quechua. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 1–9.

Shantipriya Parida, Subhadarshi Panda, Amulya Dash, Esaú Villatoro-Tello, A Seza Doğruöz, Rosa M Ortega-Mendoza, Amadeo Hernández, Yashvardhan Sharma, and Petr Motlicek. 2021. Open machine translation for low resource south american languages (americasnlp 2021 shared task contribution). In *First Workshop on Natural Language Processing for Indigenous Languages of the Americas (NAACL-HLT 2021)*, pages 218–223. Association for Computational Linguistics (ACL).

Begoa Pendas, Andrés Carvallo, and Carlos Aspillaga. 2023. Neural machine translation through active learning on low-resource languages: The case of spanish to mapudungun. In *Proceedings of the Workshop on Natural Language Processing for Indigenous*

Languages of the Americas (AmericasNLP), pages 6–11.

Robert Pugh, Marivel Huerta Mendez, Mitsuya Sasaki, and Francis Tyers. 2022. Universal dependencies for western sierra puebla nahuatl. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5011–5020.

Robert Pugh and Francis Tyers. 2021. Investigating variation in written forms of nahuatl using character-based language models. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 21–27.

Robert Pugh and Francis Tyers. 2023. A finite-state morphological analyser for highland puebla nahuatl. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 103–108.

Robert Pugh, Francis Tyers, and Quetzil Castaeda. 2023. Developing finite-state language technology for maya. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 30–39.

Annette Rios, Anne Göhring, Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis. 2012. A tree is a baum is an árbol is a sach'a: Creating a trilingual treebank.

Lorena Martín Rodríguez, Tatiana Merzhevich, Wellington Silva, Tiago Tresoldi, Carolina Aragon, and Fabrício F Gerardi. 2022. Tupían language ressources: Data, tools, analyses. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 48–58.

Cárdenas Ronald and Zeman Daniel. 2018. A morphological analyzer for shipibo-konibo. In *Association for Computational Linguistics*.

Alex Rudnick. 2011. Towards cross-language word sense disambiguation for quechua. In *Proceedings of the Second Student Research Workshop associated with RANLP 2011*, pages 133–138.

Jack Rueter, Marília Fernanda Pereira de Freitas, Sidney Da Silva Facundes, Mika Hämäläinen, and Niko Partanen. 2021. Apurin\~ a universal dependencies treebank. *arXiv preprint arXiv:2106.03391*.

Lane Schwartz. 2022. Primum non nocere: Before working with indigenous data, the acl must confront ongoing colonialism. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, volume 2.

Jiatong Shi, Jonathan D Amith, Xuankai Chang, Siddharth Dalmia, Brian Yan, and Shinji Watanabe. 2021a. Highland puebla nahuatl speech translation corpus for endangered language documentation. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 53–63.

Jiatong Shi, Jonathan D Amith, Rey Castillo García, Esteban Guadalupe Sierra, Kevin Duh, and Shinji Watanabe. 2021b. Leveraging end-to-end asr for endangered language documentation: An empirical study on yolox\'ochitl mixtec. *arXiv preprint arXiv:2101.10877*.

Rolando Coto Solano. 2021. Explicit tone transcription improves asr performance in extremely low-resource languages: A case study in bribri. In *Proceedings of the first workshop on natural language processing for Indigenous languages of the Americas*, pages 173–184.

David Stap and Ali Araabi. 2023. Chatgpt is not a good indigenous translator. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 163–167.

Research Department Statista. 2023. Share of indigenous language speakers latin america by country 2018 | statista.

Liling Tan. 2023. Few-shot spanish-aymara machine translation using english-aymara lexicon. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 168–172.

Ann Taylor, Mitchell Marcus, and Beatrice Santorini. 2003. The penn treebank: an overview. *Treebanks: Building and using parsed corpora*, pages 5–22.

Guillaume Thomas. 2019. Universal dependencies for mbyá guaraní. In *Proceedings of the third workshop on universal dependencies (udw, syntaxfest 2019)*, pages 70–77.

Atnafu Lambebo Tonja, Christian Maldonado-Sifuentes, David Alejandro Mendoza Castillo, Olga Kolesnikova, Noé Castro-Sánchez, Grigori Sidorov, and Alexander Gelbukh. 2023a. Parallel corpus for indigenous language translation: Spanish-mazatec and spanish-mixtec. *arXiv preprint arXiv:2305.17404*.

Atnafu Lambebo Tonja, Hellina Hailu Nigatu, Olga Kolesnikova, Grigori Sidorov, Alexander Gelbukh, and Jugal Kalita. 2023b. Enhancing translation for indigenous languages: Experiments with multilingual models. *arXiv preprint arXiv:2305.17406*.

Adriano Ingunza Torres, John Miller, Arturo Oncevay, and Roberto Zariquiey Biondi. 2021. Representation of yine [arawak] morphology by finite state transducer formalism. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 102–112.

Francis Tyers and Samuel Herrera Castro. 2023. Towards a finite-state morphological analyser for san mateo huave. In *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 30–37.

Francis Tyers and Robert Henderson. 2021. A corpus of k'iche'annotated for morphosyntactic structure. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 10–20.

Francis Tyers and Nick Howell. 2021. A survey of part-of-speech tagging approaches applied to k'iche'. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 44–52.

Francis Tyers, Robert Pugh, and Valery Berthoud. 2023. Codex to corpus: Exploring annotation and processing for an open and extensible machine-readable edition of the florentine codex. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 19–29.

Alonso Vasquez, Renzo Ego Aguirre, Candy Angulo, John Miller, Claudia Villanueva, Željko Agić, Roberto Zariquiey, and Arturo Oncevay. 2018. Toward universal dependencies for shipibo-konibo. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 151–161.

Raúl Vázquez, Yves Scherrer, Sami Virpioja, and Jörg Tiedemann. 2021. The helsinki submission to the americasnlp shared task. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*. The Association for Computational Linguistics.

Monica Ward. 2018. Qualitative research in less commonly taught and endangered language call.

Jonathan Washington, Felipe Lopez, and Brook Lillehaugen. 2021. Towards a morphological transducer and orthography converter for western tlacolula valley zapotec. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 185–193.

Wikipedia. 2023. Indigenous languages of the americas - language families and unclassified languages.

Roberto Zariquiey, Claudia Alvarado, Ximena Echevarria, Luisa Gomez, Rosa Gonzales, Mariana Illescas, Sabina Oporto, Frederic Blum, Arturo Oncevay, and Javier Vera. 2022a. Building an endangered language resource in the classroom: Universal dependencies for kakataibo. *arXiv preprint arXiv:2206.10343*.

Roberto Zariquiey, Arturo Oncevay, and Javier Vera. 2022b. Cld^2 language documentation meets natural language processing for revitalising endangered languages. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 20–30.

Rodolfo Zevallos, Luis Camacho, and Nelsi Melgarejo. 2022a. Huqariq: A multilingual speech corpus of native languages of peru for speech recognition. *arXiv preprint arXiv:2207.05498*.

Rodolfo Zevallos, John Ortega, William Chen, Richard Castro, Nuria Bel, Cesar Toshio, Renzo Venturas, Hilario Aradiel, and Nelsi Melgarejo. 2022b. Introducing qubert: A large monolingual corpus and bert model for southern quechua. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 1–13.

Francis Zheng, Machel Reid, Edison Marrese-Taylor, and Yutaka Matsuo. 2021. Low-resource machine translation using cross-lingual language model pretraining. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 234–240.

A Papers we found from ACL Anthology

(Zheng et al., 2021; Levin, 2009; Tonja et al., 2023a,b; Vázquez et al., 2021; Mager et al., 2023a; Chiruzzo et al., 2022; Tyers et al., 2023; Agüero-Torales et al., 2021; De Gibert et al., 2023; Cavalin et al., 2023; Monson et al., 2006; Zevallos et al., 2022a; Mager et al., 2018a; Parida et al., 2021; Pugh et al., 2023; Chen and Abdul-Mageed, 2022; Pugh et al., 2022; Duan et al., 2019; Nagoudi et al., 2021; Schwartz, 2022; Feldman and Coto-Solano, 2020; Graichen et al., 2023; Washington et al., 2021; Kiss and Thomas, 2019; Zariquiey et al., 2022b; Ćavar et al., 2016; Thomas, 2019; Ginn and Palmer, 2023; Cordova and Nouvel, 2021; Moon et al., 2009; Adams et al., 2019; Rios et al., 2012; Montoya et al., 2019; Knowles et al., 2021; Kuznetsova and Tyers, 2021; Pendas et al., 2023; Góngora et al., 2021; Ebrahimi et al., 2021; Ortega et al., 2023; Himoro and Lora, 2022; Zevallos et al., 2022b; Ortega and Pillaipakkamnatt, 2018; Martínez et al., 2020; Tyers and Castro, 2023; Márquez and Meza-Ruiz, 2021; Shi et al., 2021a; Gutierrez-Vasques and Mijangos, 2018; Stap and Araabi, 2023; Martínez et al., 2021; Pugh and Tyers, 2023, 2021; Bollmann et al., 2021; Gutierrez-Vasques et al., 2016; Mager et al., 2018b; Kann et al., 2018; Mager et al., 2023b; Kann et al., 2018; Amith et al., 2021; Solano, 2021; Jones et al., 2023; Tan, 2023; Shi et al., 2021b; Bustamante et al., 2020; Góngora et al., 2022; Mager et al., 2020, 2022; Mercado-Gonzales et al., 2018; Oncevay, 2021; Oncevay et al., 2022; Maguiño-Valencia et al., 2018; Torres et al., 2021; Zariquiey et al., 2022a; Ronald and Daniel, 2018; Alva and Oncevay, 2017; Moreno, 2021; Vasquez et al., 2018; Gow-Smith and Villegas, 2023; Ahmed et al., 2023; Ahumada et al., 2022; Downey et al., 2021; Pankratz, 2021; Rudnick, 2011; Ebrahimi et al., 2023; Chen and Fazio, 2021; Blum, 2022; Homola

and Coler, 2013; Rodríguez et al., 2022; Kann et al., 2022; Rueter et al., 2021; Coto-Solano et al., 2021; de León, 2010; Chiruzzo et al., 2020; Tyers and Henderson, 2021; Tyers and Howell, 2021)

B Country vs frequency of responses explanation

No awareness/Knowledge: This lack of awareness manifests in various forms, including limited understanding of the technologies themselves, uncertainty about the availability and accuracy of datasets for indigenous languages, and a general lack of information about AI and NLP applications tailored to their needs. Additionally, there is a sense that the lack of knowledge leads to a lack of interest and devaluation of indigenous culture and people. Some attribute this lack of knowledge to educational barriers, linguistic challenges, and cultural resistance to change.

Language/Culture Preservation: Language preservation emerges as a primary concern, as many indigenous languages lack representation in digital tools and face the risk of being lost or altered over time. Cultural preservation is also highlighted, with a focus on maintaining traditions, customs, and ancestral knowledge. The lack of options and resources in indigenous languages exacerbates the challenge, as does the resistance to modern technology due to its potential impact on linguistic identity. Despite these obstacles, these communities have a strong desire to preserve their languages and cultures, indicating the need for tailored solutions that respect their unique identities and address their specific linguistic and cultural needs.

Little interest from Govt: This limited interest is evidenced by the absence of public policies and initiatives aimed at utilizing these technologies for the benefit of indigenous communities. Additionally, there is a lack of resources and support from governments for the development and preservation of indigenous languages and cultures. Institutional marginalization, discrimination, and lack of recognition exacerbate the challenges faced by these communities, leading to social isolation and hindering their access to essential resources and opportunities.

Lack of involvement of indigenous community : This lack of involvement stems from various factors, including resistance to change and new technologies within indigenous communities, limited education and knowledge about technological aids, and cultural barriers. Additionally, there is a perception of exclusion and discrimination, with indigenous languages and cultures often overlooked or undervalued in the digital world. Furthermore, there is a need for greater inclusion and integration of indigenous perspectives and languages in the development of NLP and AI technologies to ensure that these tools adequately serve the needs of indigenous communities and contribute to preserving their languages and cultures.

Lack of effort from scientific community: This lack of effort is characterized by insufficient research and interest in developing solutions tailored to the needs of indigenous languages and cultures. Many feel that the scientific community should prioritize adapting NLP and AI technologies to native languages to preserve cultural identity and ensure inclusivity.

Limited Resources: Latin American indigenous communities face a multitude of challenges in adopting NLP and AI, primarily due to limited resources. Financial and technological constraints hinder their access to advanced technologies, while the lack of digitized linguistic resources threatens language preservation. Limited data availability and difficulties in accessing modern technology further impede progress. Economic factors, including insufficient investment, pose significant barriers to technology adoption.

Lack of Access: This lack of access manifests in various forms, including limited resources for technological infrastructure and education. Many communities struggle with inadequate internet connectivity, hindering their ability to utilize modern technologies effectively. Additionally, geographic isolation and economic constraints exacerbate the issue, preventing access to essential tools and resources. Lack of education and training further compound the problem, as does the preservation of cultural identity, which can sometimes conflict with the adoption of new technologies.