A survey of natural language processing resources for indigenous languages spoken in Argentina

Anonymous ACL submission

Abstract

Argentina has a rich linguistic diversity, which includes about 14 indigenous languages, many of which are gradually diminishing in usage. Natural language processing tools could help the indigenous communities to revitalize their languages, and to foster improved integration within the broader society. In this work we present an overview of the linguistic diversity of indigenous languages spoken languages in Argentina, a survey of computational resources, including regional work, for these languages and variants and identify challenges and opportunities of working with the under-resourced indigenous languages spoken in Argentina.

1 Introduction

At present, the most spoken languages in South America are two European languages, Spanish and Portuguese. As regards the indigenous languages, from the presumably extreme language diversity existent in the area before the arrival of Europeans (Adelaar 2012), only a bit more than 420 indigenous languages survive (Grimes 1996, Adelaar 2010). Linguistically, the diversity of those languages covers aspects such as the genetic (i.e., variety of linguistic families), typology (i.e., variety of language types as regards grammatical properties) and geographical distribution of its languages. According to (Dixon and Aikhenvald, 2006), the linguistic variety of South America is peculiar due to the noticeable discontinuity in the distribution of languages across the continent, which sets it apart from other areas of the world.

The linguistic diversity of Argentina stands out in the area due to the combination of its native population and the multiple historical waves of immigrants. For instance, Argentina is the country with the highest regional immigration rate in South America, according to the International Organization for Migration (IOM, 2022), being social-economic vulnerability the main reason to migrate.

Despite the linguistic and cultural richness of this country, research in indigenous language have not occupied a meaningful place. For instance, there is a only one survey of indigenous languages in Argentina (Censabella, 1999b) and none for computational resources. Based on (UNESCO, 2022), some challenges of indigenous languages are related to the colonial past of their nations. Despite of the suggestions for the (United Nation, Agenda 2030, 2017), the following topics are pending issues in the Argentinian public policy agenda: detailed indigenous demographic data, indicators that reflects indigenous peoples' situation, nationwide cultural awareness of cultural diversity, among others. As long as these issues remains unaddressed, the barriers of computational and linguistic research on indigenous languages will remain.

The need of availability of natural language processing resources for indigenous languages is of utmost interest for language revitalization, culture preservation and numerous applications for reducing inequalities. Some of examples of relevant work done through computational resources for indigenous languages spoken in Argentina are (Ahumada et al.) in language learning and (Kellert and Zaman) in health care education.

In recent years, the natural language processing (NLP) community has put efforts to increase endangered and under-resource language research (Magueresse et al.; Hedderich et al.; Ranathunga et al., 2023). However, most efforts have been done from a *machine-centric* perspective, in order to tackle the technical challenges of working with limited data, without considering speaker perspective (Liu et al.), (Ramponi 2022), multiculturalism (Hershcovich et al.) and decolonization of NLP (Bird; Schwartz).

In this work, we aim to introduce the research opportunities and technical challenges of working with indigenous languages spoken in Argentina. Our contributions to the NLP research for South

American languages are following:

1. An overview of the linguistic diversity of spoken languages in Argentina and its relevance to the region
2. A survey of computational resources and regional work done for these languages
3. Identified technical challenges and opportunities of working with the under-resourced indigenous languages spoken in Argentina

The rest of the paper is organized as follows. In Section 2, we offer an overview of the demolinguistic situation of Argentina, focusing on indigenous languages. In Section 3, we review the computational resources and work done for 2 families of indigenous languages: Quechua and Mapuche families. In Section 4, we discuss some of the challenges on working with indigenous languages presents for NLP. Finally, in Section 5 we draw some final conclusions and prospects.

2 Language Diversity in Argentina

Before evaluating the landscape of potential resources for the languages spoken in Argentina, a snapshot is needed regarding what languages are spoken in the territory and how many people spoke them. A first challenge to be faced is that we lack a complete demolinguistic census in order to have rigorous information on this issue (Bein 2021). Besides this, it is known that the language diversity of Argentina covers at least Argentinian Spanish[1], Argentinian Sign Language, indigenous languages and languages of immigrant communities, such as German, Italian, English and the like (see Bein 2021 for details). In this paper, we deal only with indigenous languages.

The classifications in the literature differ with respect to which indigenous languages exist in Argentina and which are (still) spoken (see Censabella 1999a, Censabella 2009, Ciccone 2010, Nercesian 2021). In part, this is due to the fact that these languages show different situations concerning their standardization, their vitality, the sociolinguistic attitude of speakers towards the languages, their documentation, their denomination (sometimes two or more denominations for the same language coexist, sometimes, different languages share the same name), their delimitation (it is not always clear whether a language is a variety of another or an autonomous language) and the amount of information on the languages.

Furthermore, according to our best knowledge, there is no reliable source regarding the number of speakers of each language in Argentina. The national census and the Supplementary Survey of Indigenous People conducted in Argentina asked people if they self-identified as part of a indigenous community, but not if they speak the language[2], and if they understand or speak an indigenous language, but there is no information about the language they speak[3]. Other sources of information are UNESCO's World Atlas of Languages[4], Glottolog (Hammarström et al., 2023) –a catalogue of the world's languages, language families and dialects– and Ethnologue (Eberhard et al., 2023) –a catalog of the 'metadata' of languages, that contains information about how languages are used around the world, who uses them, where and for what purpose– [5], that provides information about languages spoken in different countries, their level and endangerment and number of speakers. Finally, a research group provides information regarding the number of speakers of indigeneous languages in Argentina[6]. As there is no definition about how to consider a speaker, about if the studies are considering immigrants or not, there is no coincidence in the number of speakers informed by each study.

Table 1 shows a possible systematization of the languages spoken in Argentina with some data on number of speakers according to UNESCO's WALS, the ISO codes, the indigenous population according to INDEC and the vitality according to Ethnologue. Languages for which it is unknown whether there are still living speakers or not, such as aonekko 'a'ish (tehuelche), cacán (diaguita), al-

[1] Argentinian Spanish is usually divided in a series of subvarieties (see Vidal de Battini 1964, Fontanella de Weinberg 2000, Villarino and Piñeiro Carreras). The variety spoken in the City of Buenos Aires of Argentina along with its surrounding areas and in Uruguay is often called Rioplatense Spanish.

[2] 2010 Argentinian Census, ECPI Supplementary Survey of Indigenous Peoples: https://www.indec.gob.ar/micro_sitios/webcenso/ECPI/index_ecpi.asp.

[3] 2004-2005 Argentinian ECPI: https://www.indec.gob.ar/micro_sitios/webcenso/ECPI/pueblos/ampliada_index_nacionales.asp?mode=00.

[4] WALS UNESCO: https://en.wal.unesco.org/countries/argentina/

[5] Ethnologue: https://www.ethnologue.com/.

[6] Observatorio de los Derechos de los pueblos Indígenas y Campesinos: https://www.soc.unicen.edu.ar/observatorio/index.php/22-articulos/106-unas-700-000-personas-mantienen-vivas-15-lenguas-indigenas-en-argentina.

Family Language	ISO	Indigenous Population (INDEC)	Number of Speakers (UNESCO)	Vitality (Ethnologue)
Aymara				
Aymara Central	ayr	4104	1,707	●
Chon				
Tehuelche	teh	1059	961	●
Guaycuruan				
Toba (o qom)	tob	69452	34,949	●
Mocoví	moc	15837	3,752	●
Pilagá	plg	15837	3,512	●
Mapuche				
Mapudungún	arn	11368	17,897	●
Mataco-Mataguaya				
Wichí	wlv	40036	29,066	●
Nivaclé	cag	553	266	●
Chorote	crq/crt	2613	1,711	●
Quechuan				
Bolivian and Peruvian Quechua	que	561	9,999,999	○
Kolla quechua	que	70505	-	○
Quichua Santiagueño	qus	-	99,999	●
Tupí Guaraní				
Ava Guaraní	gui	21807	8,943	●
Tapiete	tpj	484	282	●
Guaraní	grn	22059	8,178	○
Mbya Guaraní	gun	8223	99,999	●

Table 1: Indigenous languages spoken in Argentina. ISO 639-3 code, a three-letter identifier code for all known human languages[7], language name, family, indigenous population according to INDEC, number of speakers according to UNESCO, and levels of endangerment (LoE) according to Ethnologue are shown, where green symbol means stable, red symbol means endangered and the transparent symbol means that there is no data.

lentiac and millcayac (huarpe) (Nercesian 2021) are not shown. It is worth noting that the indigenous population does not necessarily reflect number of speakers, because some indigenous populations have lost their original languages and some people which do not self-recognize as part of the indigenous population do maintain an indigenous language (Ciccone 2010, Censabella 2009: 159-169). Another consideration to pinpoint is that the sources used for the table differ with respect to which languages or which varieties they take into account.

3 NLP for Indigenous languages

The study of indigenous languages from a computational perspective has gained popularity in the international research community, as part of the recent trend in low-resource NLP. Many established researchers of the field evinced the large gap between high and low resource language research (Tsvetkov; Ben; Joshi et al.; Blasi et al.), leading to the creation of various dedicated venues. For example, special interest groups such as SIGEL at ACL and SIGUL at ELRA/ISCA, workshops at top tier conferences such as ACL (LoResMT), EMNLP (ComputEL), NAACL (DeepLo), LREC (DCLRL), ICLR PML4DC, among others. Recently, NLP research community from Latin America and the Global South have created venues and events dedicated to regional challenges, such as AmericasNLP, KHIPU, RIIAA.

However, despite the effort on promoting NLP research in native American languages, there is still a huge language disparity on the NLP world that specially affects the Global South

3.1 Quechua Family

The Quechuan languages are the most studied by the NLP community, mainly performed by Peruvian researchers. There are projects for speech corpus creation and monolingual text corpus, which covers only Peruvian quechua languages. In 2018, the Siminchik corpus with 97 hours of raw audio recordings for Quechua Chanca and Quechua Collao (Cardenas et al.). In 2022, the multilingual Huqariq corpus for the development of automatic speech recognition, language identification and text-to-speech tools, with 220 hours of transcribed audios in Central Quechua (Glottolog quec1386) and Southern Quechua (Glottolog quec1389) (Zevallos et al., a). Same year, a 15 hours speech corpus was published with audio recordings in Quechua Collao for automatic emotion recognition (Paccotacya-Yanque et al., 2022). Among the mentioned projects, only the last one has their data publicly available under an open source licence. There exist other data resources that have been used in the AmericasNLP shared task, which are also for Peruvian Southern Quechua (Agić and Vulić; Tiedemann).

Among the monolingual text resources, we highlight the first large combined corpus for deep learning, publicly available in Hugging Face[8]. This dataset was used to train QuBERT, the first pretrained BERT model for Quechuan languages, which was tested for (1) named-entity recognition (NER) and (2) part-of-speech (POS) tagging, achieving similar results to other work on high-resource languages (Zevallos et al., b).

There are several work done for Peruvian

[8] https://huggingface.co/

Quechua which cover common NLP task and subfields, such as language identification (Linares and Oncevay-Marcos, 2017), data augmentation (Zevallos et al., 2022), multilingual neural machine translation (Ortega et al.; Oncevay, 2021; Alvarez-Crespo et al., 2023), corpora alignment (Ortega and Pillaipakkamnatt), lexical database construction (Melgarejo et al.). Other efforts has been made in evaluating and applying linguistic tools for Quechua languages, such as a morphological analyzer (Himoro and Pareja-Lora), the use of automatic grammar generator for the study of gerunds in Quechua and Spanish (Rodrigo et al.). No special resources for the varieties of Quechua spoken in Argentina were found in our survey.

3.2 Mapuche Family

For the Mapuche Family, most of the resources we found were developed in Chile and were aimed predominantly to machine translation tasks. For instance, (Pendas et al., 2023) present a Neural Machine Translation model based on active learning for Mapudungun. Active learning is an algorithm approach which actively selects informative data to learn from, leading to better results when applied to low-resources data.

In a similar fashion, in (Levin et al., 2002), the AVENUE-Mapudungun plan is presented, which consists in a conjoined project between the Ministry of Education in Chile and Carnegie Mellon University's Language Technologies Institute in the United States. The aim of the project is to build a a parallel corpus of Spanish and Mapudungun containing both written texts and transcribed speech. Plans are described for using this corpus for machine translation.

(Duan et al., 2019) describes a corpus of oral transcriptions in Mapudungun from 142 hours of conversations in the domain of medical treatment along with its translation into Spanish and some additional annotations. They also provide some provisional results on speech recognition, speech synthesis, and machine translation between Spanish and Mapudungun based on this corpus.

Another similar contribution is presented in (Ahumada et al., 2022). These authors afford three tools designed towards supporting educational activities of Mapuzugun: an orthography detector and converter, a morphological analyzer, and, again, an informal translator.

Mapudungun also counts with a digital Corpus of Historical Mapudungun, which includes many of the earliest writings in the Mapudungun language from 1606 to 1930[9].

Again, as in the case of Quechua Family resources, Argentina's contributions are, to the best of our knowledge, insufficient, if not non-existent.

4 Discussion

One of the main challenges faced for the development of NLP applications for indigenous languages is the low availability of resources. Annotated corpora is fundamental for training machine learning systems. Nevertheless, most indigenous languages are used in spoken form and there is scarce usage of them on the Internet. The oral predominance of these languages (over the written) collaborates with the non-standardized spelling, which hinders the performance of NLP tasks, leading to sub-optimal results. Furthermore, as it also happens with medium-resource languages, there is low domain diversity.

Also, for generating resources, such as corpora, there have to be annotators, interest of the government and financial support for generating it.

Besides, these languages have rich morphology and dialectal variety. This indicates that tools and resources developed for one variety should probably be also generated for other varieties of the same language, as it happens with different varieties of Spanish.

See Mager et al. (2023), to refer to a very complete list of challenges faced for the development of NLP applications.

As follows from our survey, more work is to be done on indigenous languages of Argentina. First, as observed in Section 2, there is no agreement regarding which indigenous languages are spoken in Argentina. Second, there is no demolinguistic reliable information of Argentina and, the data available do not establish a clear definition of what qualifies an individual as a speaker. Should it be based on having the language as a mother tongue? It is often accepted that speakers of indigenous languages in Argentina are in a diglossia situation, i.e. a situation of asymmetric bilinguism where some language is used in more prestigious context and the other is relegated to informal contexts (Ferguson 1959). However, more studies on what kind of proficiency on the language the speakers have and in which context do they use the indigenous

[9] See http://www.amc-resources.lel.ed.ac.uk/CHM/.

language are needed. Third, Argentina seems to be in disadvantadge when compared to other countries sharing some indigenous langugages. For instance, in Chile some efforts have been made in order to build resources to assist learning in Mapudungun. Though Argentina counts with legislation that intends to promote the use of indigenous languages in education, the so-called Educación Intercultural Bilingüe (EIB)[10], we are not aware of any existing computational resources pursuing similar goals to those we found in Chile.

5 Final remarks

There is no consensus about the indigeneous languages that exist and that are still spoken in Argentina. We provide a table summarizing information about languages, number of speakers, and vitality provided by the main sources. Of the about 14 indigenous languages spoken in Argentina, we found NLP applications and resources for 4 of them (in variants different than those spoken in Argentina). We describe those applications for Quechua and Mapuche families. Finally, we discuss the situation and some challenges of the development of NLP applications for indigenous languages in Argentina. These could serve as research opportunities. We are finishing a survey of NLP development for Guaraní and Aymara.

References

The #BenderRule: On Naming the Languages We Study and Why It Matters.

Willem Adelaar. 2010. South america. In Christopher Moseley, editor, *Atlas of the world's languages in danger*, pages 86–94.

Willem FH Adelaar. 2012. Historical overview: Descriptive and comparative research on south american indian languages. In Verónica Grondona and Lyle Campbell, editors, *The indigenous languages of South America: A comprehensive guide*. De Gruyter.

Željko Agić and Ivan Vulić. 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. Association for Computational Linguistics.

Cristian Ahumada, Claudio Gutierrez, and Antonios Anastasopoulos. Educational Tools for Mapuzugun. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 183–196. Association for Computational Linguistics.

Cristian Ahumada, Claudio Gutierrez, and Antonios Anastasopoulos. 2022. Educational tools for mapuzugun. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 183–196. Association for Computational Linguistics.

Abraham Alvarez-Crespo, Diego Miranda-Salazar, and Willy Ugarte. 2023. Model for real-time subtitling from spanish to quechua based on cascade speech translation. In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence (ICAART 2023)*, volume 3, pages 837–844. SCITEPRESS – Science and Technology Publications.

Roberto Bein. 2021. Las políticas lingüísticas. In *La lingüística. Una introducción a sus principales preguntas*, pages 407–431. Eudeba, Ciudad de Buenos Aires.

Steven Bird. Decolonising Speech and Language Technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519. International Committee on Computational Linguistics.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 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. Association for Computational Linguistics.

Ronald Cardenas, Rodolfo Zevallos, Reynaldo Baquerizo, and Luis Camacho. Siminchik: A Speech Corpus for Preservation of Southern Quechua.

Marisa Censabella. 1999a. *Las lenguas indígenas de la Argentina*. Eudeba, Buenos Aires.

Marisa Censabella. 1999b. Las lenguas indígenas de la Argentina: Una mirada actual.

Marisa Inés Censabella. 2009. Chaco ampliado. pages 143–237.

Florencia Ciccone. 2010. Aportes al conocimiento de las lenguas indígenas en Argentina y su tratamiento desde la EIB. Material elaborated for the Ministerio de Educación de la Nación.

Robert Dixon and Alexandra. Aikhenvald. 2006. *The Amazonian Languages*. Cambridge University Press.

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*.

[10] https://www.argentina.gob.ar/nivelesymodalidades/modalidad-de-educacion-intercultural-bilingue

David M Eberhard, Gary Francis Simons, and Charles D Fenning. 2023. Ethnologue: Languages of the world.

Charles A Ferguson. 1959. Diglossia. *word*, 15(2):325–340.

Beatriz Fontanella de Weinberg, editor. 2000. *El español de la Argentina y sus variedades regionales*. Edicial, Buenos Aires.

Barbara Grimes, editor. 1996. *Ethnologue. Languages of the World*. Summer Institute of Linguistics, Dallas, Texas.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2023. glottolog/glottolog: Glottolog database 4.8.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568. Association for Computational Linguistics.

Daniel Hershcovich, Stella Frank, Heather Lent, prefix=de useprefix=true family=Lhoneux, given=Miryam, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. Challenges and Strategies in Cross-Cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013. Association for Computational Linguistics.

Marcelo Yuji Himoro and Antonio Pareja-Lora. 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. European Language Resources Association.

IOM, 2022. Recent migration movements in South America, 2022 Annual Report. International Organization for Migration and Specialized Forum on Migration of MERCOSUR and Associated States".

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 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. Association for Computational Linguistics.

Olga Kellert and Mahmud Zaman. Use of NLP in the Context of Belief states of Ethnic Minorities in Latin America. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 1–5. Association for Computational Linguistics.

Lorraine Levin, Rodolfo M Vega, Jaime G Carbonell, Ralf D Brown, Alon Lavie, Eliseo Cañulef, and Carolina Huenchullan. 2002. Data collection and language technologies for mapudungun. In *International Workshop on Resources and Tools in Field Linguistics*, Las Palmas, Spain.

Alexandra Espichán Linares and Arturo Oncevay-Marcos. 2017. A low-resourced peruvian language identification model. In *CEUR Workshop Proceedings. CEUR-WS*.

Zoey Liu, Crystal Richardson, Richard Hatcher, and Emily Prud'hommeaux. Not always about you: Prioritizing community needs when developing endangered language technology. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3933–3944. Association for Computational Linguistics.

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

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. Low-resource Languages: A Review of Past Work and Future Challenges.

Nelsi Melgarejo, Rodolfo Zevallos, Hector Gomez, and John E. Ortega. WordNet-QU: Development of a Lexical Database for Quechua Varieties. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4429–4433. International Committee on Computational Linguistics.

Verónica Nercesian. 2021. Las lenguas del mundo. In *La lingüística. Una introducción a sus principales preguntas*, pages 77–106. Eudeba, Ciudad de Buenos Aires.

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. Association for Computational Linguistics.

John Ortega and Krishnan Pillaipakkamnatt. 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. Association for Machine Translation in the Americas.

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

Rosa YG Paccotacya-Yanque, Candy A Huanca-Anquise, Judith Escalante-Calcina, Wilber R Ramos-Lovón, and Álvaro E Cuno-Parari. 2022. A speech corpus of quechua collao for automatic dimensional emotion recognition. *Scientific Data*, 9(1):778.

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.

Alan Ramponi. 2022. Nlp for language varieties of italy: Challenges and the path forward. *arXiv preprint arXiv:2209.09757*.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37.

Andrea Rodrigo, Maximiliano Duran, and María Yanina Nalli. Approach to the Automatic Treatment of Gerunds in Spanish and Quechua: A Pedagogical Application. In *Formalizing Natural Languages: Applications to Natural Language Processing and Digital Humanities*, Communications in Computer and Information Science, pages 135–146. Springer International Publishing.

Lane Schwartz. 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: Short Papers)*, pages 724–731. Association for Computational Linguistics.

Jörg Tiedemann. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218. European Language Resources Association (ELRA).

Yulia Tsvetkov. Opportunities and Challenges in Working with Low-Resource Languages.

UNESCO, 2022. 2022. State of the art of indigenous languages in research.

United Nation, Agenda 2030, 2017. 2017. Indigenous peoples' rights and the 2030 agenda.

Berta Vidal de Battini. 1964. *El español de la Argentina. Estudio destinado a los maestros de las escuelas primarias*. Consejo Nacional de Educación, Buenos Aires.

Julio Villarino and Julia Piñeiro Carreras. Los aglomerados urbanos y la diversidad poblacional: aportes para una futura actualización de la regionalización dialectal del español en Argentina. *RASAL Lingüística*, 2023:59–97.

Rodolfo Zevallos, Nuria Bel, Guillermo Cámbara, Mireia Farrús, and Jordi Luque. 2022. Data augmentation for low-resource quechua asr improvement. *arXiv preprint arXiv:2207.06872*.

Rodolfo Zevallos, Luis Camacho, and Nelsi Melgarejo. a. Huqariq: A Multilingual Speech Corpus of Native Languages of Peru forSpeech Recognition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5029–5034. European Language Resources Association.

Rodolfo Zevallos, John Ortega, William Chen, Richard Castro, Núria Bel, Cesar Toshio, Renzo Venturas, Aradiel, and Hilario Nelsi Melgarejo. b. 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. Association for Computational Linguistics.