Indigenous Languages Spoken in Argentina: A Survey of NLP and Speech Resources

Belu Ticona[1,2], Fernando Carranza[3,4], Viviana Cotik[2,5]
[1]George Mason University, United States
[2] Departamento de Computación, FCEyN, Universidad de Buenos Aires (UBA), Argentina
[3] Departamento de Letras, FFyL, UBA, Argentina
[4] Instituto de Filología y Literaturas Hispánicas "Dr. Amado Alonso", UBA, Argentina
[5] Instituto de Investigación en Ciencias de la Computación (ICC), CONICET-UBA, Argentina
Correspondence: mticona@dc.uba.ar

Abstract

Argentina has a diverse, yet little-known, Indigenous language heritage. Most of these languages are at risk of disappearing, resulting in a significant loss of world heritage and cultural knowledge. Currently, no unified information on speakers and computational tools is available for these languages. In this work, we present a systematization of the Indigenous languages spoken in Argentina, along with national demographic data on the country's Indigenous population. The languages are classified into seven families: Mapuche, Tupí-Guaraní, Guaycurú, Quechua, Mataco-Mataguaya, Aymara, and Chon. We also provide an introductory survey of the computational resources available for these languages, whether or not they are specifically developed for Argentine varieties.

1 Introduction

By the end of this century, about half of all languages spoken in the world are in danger of disappearing, according to UNESCO (Moseley, 2010). Since language is a key part of the identity and culture of speakers, the development of technology may help sustain and promote linguistic diversity and maintain cultural heritage.

Developing technology for endangered languages has the potential to help communities preserve and revitalize their cultural and linguistic heritage, enhance digital communication, increase access to information, and improve education in their native languages, among other benefits. In this context, the natural language processing (NLP) community has been putting efforts into developing computational resources for languages around the world, including Indigenous languages from Latin America (Tonja et al., 2024), and under-served languages for specific countries and regions (Aji et al., 2022; Ramponi, 2024; Adebara and Abdul-Mageed, 2022; Blaschke et al., 2023).

However, technical and ethical challenges emerge in adapting common NLP practices and techniques when working with Indigenous languages and their speaker communities (Mager et al., 2018; Bird, 2020; Liu et al., 2022; Schwartz, 2022; Mager et al., 2023; Bird, 2024). For instance, it is crucial to ensure that these technologies align with the specific needs and priorities of the communities they aim to support.

In Argentina, according to the National Institute of Statistics and Censuses (INDEC, 2024), there are 58 Indigenous Peoples and approximately 1.3 million Indigenous descendants, from which only 29.3% are speakers of an Indigenous language. In contrast to other countries in the region, the establishment of a narrative of European descendance has historically shaped the sociopolitical and cultural agenda, marginalizing the Indigenous population and creating a lack of social awareness regarding the cultural diversity of the country (Quijada, 2004; Adamovsky, 2012). This is evident in the fact that there is only one published linguistic survey of Indigenous languages in Argentina (Censabella, 1999)[1]. The lack of enough information on Indigenous languages, and the lack of reliable data on the number of speakers and sociolinguistic situation complicates the assessment of the state of computational resources available for these languages. To address this issue, we explore the status of Indigenous languages spoken in Argentina focusing on their prevalence, number of speakers, and available NLP resources, along with a discussion of the main trends that characterize them. This work could serve as a valuable resource for new research groups, helping them to quickly familiarize themselves with key topics, tools, and ongoing debates in the field.

We contribute by providing:

[1]Briefer, not specific, partial or unpublished work include Ciccone (2010), Censabella (2009) and Nercesian (2021), among others.

Proceedings of the 31st International Conference on Computational Linguistics, pages 8449–8461
January 19–24, 2025. ©2025 Association for Computational Linguistics

1. An overview of the linguistic diversity of Indigenous languages in Argentina,
2. A survey of computational resources and regional work done for these languages.

The rest of the paper is organized as follows. In Section 2 we present an overview of the Indigenous languages spoken in Argentina. In Section 3, we review the work done in the past for these languages. In Section 4, we discuss the general trends derived from our survey. In Sections 5 and 6, we provide conclusions and limitations of our work. In the Appendix, we explain the methodology employed to collect and select the papers, we provide additional information about the data sources used to create Figure 1, and two tables providing an overview of the available corpora and tasks studied for the Indigenous language families Mapuche, Tupí-Guaraní, Quechua, and Aymara.

2 Indigenous Languages in Argentina

Currently, there is no consensus on the precise list of languages spoken in Argentina, as demonstrated through by comparison of sources such as Censabella 1999, 2009; Ciccone 2010 and Nercesian 2021. This uncertainty is due to the complex situation of each language and its speakers, which covers language endangerment, as well as different situations regarding the amount of available documentation, standardization, use contexts, and the availability of a written form, among others. Furthermore, speakers of some of these languages often exhibit a negative attitude towards their linguistic heritage, complicating its study (Carrió, 2014).

Based on the works of Censabella (1999) and Ciccone (2010)[2], we consider the following language families: Mapuche, Tupí-Guaraní, Guaycurú, Quechua, Mataco-Mataguaya, Aymara, and Chon. When relevant, we also include notes on the peculiarities of Argentine varieties. In Figure 1 we present an overview of the Indigenous language families reviewed in this section along with their demographic data and geographical location specific to Argentina. More information about data sources to build the figure is available in Appendix (6). It is important to note that the available data is not sufficiently reliable due to a methodological issue. In the National Census, only individuals who self-identified as members of an Indigenous community were asked whether they spoke the Indigenous language of their community, without inquiring which is this language. For this reason, speakers of Indigenous languages who do not self-identified as members of an Indigenous community or who speak a language of another community were not considered.

From the **Mapuche** family, Mapudungún or Mapuzungún[3] is the most spoken Indigenous language in Argentina and Chile. According to Viegas Barros (1999), the varieties spoken in both countries differ mainly in pronunciation and vocabulary.

The **Tupí-Guaraní** family is primarily spoken in the south of the Amazon. In Argentina, it comprises Ava Guaraní (also known as Chiriguano), Tapiete (also considered an Ava Guaraní variety (Dietrich, 1986)), Mbyá Guaraní and two varieties of Guaraní: Corrientes Guaraní, the Argentinian variety spoken in the Corrientes province, and Paraguayan Guaraní, primarily spoken by Paraguayan immigrants specially in Buenos Aires, Misiones and Formosa provinces (Ciccone, 2010). The peculiarities of the Corrientes variety, as compared with the Paraguayan one, seem to lie in the amount of Spanish loanwords, the pronunciation of some phonemes, the use of *ta* instead of *piko* as an interrogative morpheme, its set of evidential particles and the use of *mã* as intensifier (Cerno, 2010, 2013).

The **Guaycurú** family includes Toba (or Qom, as their speakers call it), Mocoví, and Pilagá. These languages are spoken in the north of Argentina, mainly in the provinces of Formosa and Chaco.

The **Quechua** family covers different varieties spoken in Argentina, Bolivia, Colombia, Chile, Ecuador and Perú (Sichra 2009, p. 22). This makes it the most extended Indigenous language of South America, both in number of speakers (6,276,834 according to Moseley 2010, p. 101) and in countries where it is spoken (6 countries, Sichra 2009, p. 76). In Argentina, two Quechua varieties are identified: Cusco-Bolivian Quechua primarily spoken by Bolivian and Peruvian immigrants, and Santiago del Estero Quichua, which is mostly spoken by the local population of Santiago del Estero, a province in Argentina (Juanatey 2020, p. 24). According to Adelaar (1995) and Juanatey (2020), the particularities of the Santiago del Estero Quichua include the loss of the proto-quechua semivowel /w/ between vowels, the changes in plural morphol-

[2] See also Nercesian (2021).

[3] See Díaz-Fernández (2006) for details on the Glossonyms for this language family.

ogy and verbal inflection, and some specific lexical choices. In some literature, a Kolla Quechua variety is included (e.g., Censabella 1999). Kollas are an heterogeneous Indigenous community. We are not aware of any work addressing whether the Kolla Quechua shows peculiarities that justify treating it as a different variety.

The **Mataco-Mataguaya** family is spoken in Paraguay and Argentina. In Argentina, it includes the Wichí language, spoken in Formosa, the Nivaclé or Churupí, spoken in Salta and Formosa, and Chorote, spoken in Salta.

The **Aymara** family comprises different Aymara varieties. In Argentina, the most extended is considered to be the Central Aymara variety (Ciccone, 2010). We are not aware of any work addressing specifically the peculiarities of the Central Aymara variety spoken in Argentina.

Finally, the **Chon** family consists of several languages that were spoken in Patagonia. The majority are now extinct or in severe danger. One of the surviving languages (according to current knowledge, it might be the last) is Tehuelche, which is spoken in the Santa Cruz province (Censabella 1999).

3 Survey of computational resources

In this section, we present our survey of NLP research related to the Indigenous language families spoken in Argentina. It is worth noting that most of the resources we found were not developed specifically for Argentine varieties. Due to the lack of resources for Argentine varieties, and for the sake of completeness, we decided to include research on varieties spoken not only in Argentina but also in other countries. Therefore, some language varieties mentioned in this section may differ from those listed in Section 3. The protocol followed to collect and select the papers from which the available resources and tasks are outlined is described in Appendix (6). Tables 1 and 2 (also in Appendix -6-) summarize the corpora and tasks found for each of the considered language families.

3.1 Mapuche Family

Despite being the language of the largest Indigenous Peoples in Argentina and Chile, there is only one public large corpus for Mapundungún language (Duan et al., 2020). This corpus is a clean and detailed version of the recordings collected during the AVENUE project (Levin et al., 2002). As a result of this collaboration, rule-based MT systems were developed for Mapudungún-Spanish, as well as a spelling checker for Mapudungún (Monson et al., 2004, 2008; Llitjós et al., 2005; Monson et al., 2006). None of these resources are publicly available anymore. Related work can also be found in Pendas et al. (2023). Regarding the educational perspective, Ahumada et al. (2022) designed some tools for educational purposes, including an orthography detector and converter, a morphological analyzer, and an informal translator.

3.2 Tupí-Guaraní Family

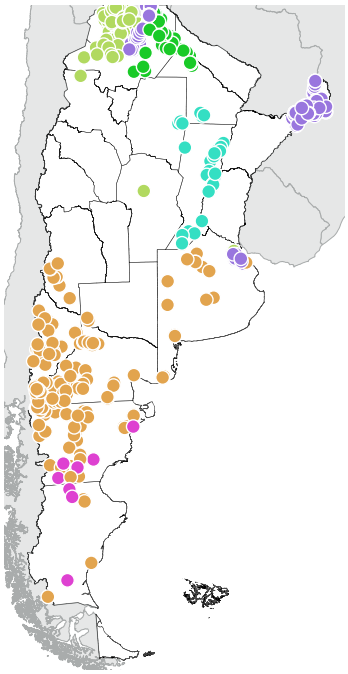
Among the Tupí-Guaraní languages, Paraguayan-Guaraní has been consistently covered over time by different NLP researchers, resulting in multiple monolingual corpora and even a parallel corpus in Guaraní-Spanish. However, much work remains to be done for other languages of the family. The first computational research initiatives in Guaraní were developed as isolated projects in different groups and countries. These include the work on data collection for sentiment analysis (Ríos et al., 2014; Agüero-Torales et al., 2021), database collection of historical texts (Cordova et al., 2019; The Langas Project), and the adaptation of Universal Dependencies guidelines for annotating Mbyá Guaraní (Thomas 2019).

In recent years, a research collaboration among multiple groups and countries gave birth to Jojajovai, the first medium-sized corpus of the language family (Chiruzzo et al., 2022). Previous works detail the challenges of developing a Guaraní corpus, suggesting ideas to diversify the type of content (Chiruzzo et al., 2020; Góngora et al., 2021). Góngora et al. (2022) used the Jojajovai dataset to enrich MT systems with pre-trained word embeddings. Torales and Matías (2022) conducted exhaustive research studying topic modeling and sentiment analysis on text from social media in Guaraní and Jopará, a mixture of Guaraní and Spanish used in Paraguay. For information on Jopará, see Estigarribia 2015.

Recently, the research community has started to work on other Tupian languages[4]. For instance, the TuLar project[5] collects, documents, and develops computational and pedagogical materials for Indigenous communities in Brazil. Martín Rodríguez et al. (2022) released an online lexical database with more than 400 concepts, a morphological

[4]Tupí is a language family native to South America, that includes various languages spoken primarily in Brazil. Tupí-Guaraní is one of its major subfamilies.

[5]https://tular.clld.org/



Language Family Language	ISO-639-3	Population	Speaker (%)	Speaking Population	Vitality
Mapuche					
Mapudungún	arn	145,783	18	26,240	❗
Tupí-Guaraní					
Ava Guaraní (Chiriguano)	nhd	3306	45	1490	❗
Paraguayan Guaraní	grn-gug	135,232	40	54,092	ℹ
Corrientes Guaraní	grn				🛡
Mbyá Guaraní	gun	11,014	84	9,251	🛡
Tapiete	tpj	654	39	255	❗
Guaycurú					
Toba (Qom)	tob	80,124	49	39,260	❗
Mocoví	moc	18,231	24	4,375	❗
Pilagá	plg	6,169	89	5,489	❗
Quechua					
Kolla Quechua	que	69,121	41	28,339	No data
Cusco-Bolivian Quechua	que-quz	52,154	51	26,598	ℹ
Santiago del Estero Quichua	que-qus				🛡
Mataco-Mataguaya					
Wichí	wlv/mzh	69,080	73	50,428	🛡
Chorote	crq/crt	3,238	75	2,428	🛡/❗
Nivaclé	cag	878	75	658	🛡
Aymara					
Central Aymara	aym-ayr	19,247	51	9,815	🛡
Chon					
Tehuelche	teh	17,420	8	1,393	❗

Figure 1: Geographical and demographic description of Indigenous languages spoken in Argentina by more than 1000 speakers (Instituto Nacional de Estadística y Censos, 2024). The data focuses exclusively on Argentina. To the left, the distribution of the Indigenous communities homonymous to Indigenous languages. To the right, Indigenous languages spoken in Argentina, grouped by family and identified by their ISO 639-3 code (macrolanguage identifier, and microlanguage in cases when more than one variety is mentioned). For each case, the population and the corresponding percentage of speakers are provided, with the percentage reflecting the proportion of the community that speaks the Indigenous language they consider to be the language of their community, as reported in the Argentine INDEC census. All data was obtained from the INDEC census (INDEC, 2024). *Speaking population* shows the estimated number of speakers of the language based on the Indigenous Population and the previously described percentage. *Vitality* shows the level of endangerment of the language according to Ethnologue (Eberhard et al., 2023): 🛡, ❗ and ℹ denote stability, endangerment and institutional status of the language, respectively. When there is no available data to distinguish between varieties associated with a given language, the data is placed in the middle of the lines. Aymara is not shown in the map, because there is no community data from Aymara in the INDEC data.

database with 51 languages, and a dependency treebank with 9 languages.

3.3 Quechua Family

The Quechuan languages are the most studied by the NLP community, mainly performed by Peruvian researchers. There are projects for creating speech corpora and monolingual text corpora, which cover only Cusco-Bolivian Quechua varieties (Cardenas et al., 2018; Zevallos et al., 2022b; Paccotacya-Yanque et al., 2022; Zevallos et al., 2022c).

There have been considerable studies conducted on Cusco-Bolivian Quechua, which cover common NLP tasks, such as language identification (Linares and Oncevay, 2017), machine translation (Ortega et al., 2020; Oncevay, 2021a; Alvarez-Crespo et al., 2023), corpora alignment (Ortega and Pillaipakkamnatt, 2018), lexical database construction (Melgarejo et al., 2022), and the creation of resources, such as data augmentation (Zevallos et al., 2022a). Other efforts have been made in evaluating and applying linguistic tools for Quechua languages, such as a morphological analyzer (Himoro and Pareja-Lora, 2022), and the use of an automatic grammar generator for the study of gerunds in Quechua and Spanish (Rodrigo et al., 2021). Only one resource specifically created for the Santiago del Estero Quichua was found in our survey: Porta (2010a). This study presents a transducer that aims at identifying the morphological structure of the language.

3.4 Other families

For the rest of the families considered in this paper, we could not find many resources. From the Guaycurú Family, we only found a work on spoken language identification for Qom (Garber, 2022)

and a description of its morphology using a linear context-free grammar (Porta 2010b). No special resources were found for Pilagá or Mocoví. To the best of our knowledge, no specific resources were developed for the Mataco-Mataguaya family, besides the fact that Nivaclé was taken into account in the language identification model presented in Kargaran et al. (2023). Regarding Aymara, the few existing works relied on data available through the shared task of the AmericasNLP workshop, exploring Spanish-Aymara machine translation (Tan, 2023; Oncevay, 2021b). Finally, regarding the Chon Family, Domingo and Manchado (2018) present a publicly available corpus on Techuelche, the only computational resource we found for this family.

4 Discussion: Trends and Challenges

The scarce resources available were produced in Argentina's neighboring countries. In this survey, we identified only a few research groups working steadily over time on Indigenous languages spoken in Argentina and its surroundings. Among the handful exceptions, we highlight the Peruvian academic community which has developed most of the work done for the Quechua language family. For other families, most of the available work has been done by a unique research group (e.g. NLP Group of UdelaR[6], in the case of Chiruzzo's work for the Guaraní family in Uruguay) or a particular initiative (e.g. the AVENUE301 project for Mapudugún). Besides these cases, most of the work identified has been conducted primarily by South American researchers working in the diaspora. It is worth mentioning that, in general, research groups from South America have limited access to computing resources and funding.

Local languages and variants have yet to be incorporated into emerging academic initiatives. A lot of work has been done for the AmericasNLP workshop on Indigenous languages (see 2020-2024 proceedings). This shows the positive impact of these challenges on the field. Nevertheless, there is almost no work conducted on the Argentinian local varieties. As seen in Section 2, the peculiarities of Argentinian varieties are scarcely studied for Corrientes Guaraní and Santiago del Estero Quichua, less known for Mapudungún and almost ignored for Aymara. For this reason, it is difficult to assess how effectively resources developed in other countries might work for local varieties.

More focus on written languages, while indigenous languages are traditionally oral. Additionally, we found that the academic community tends to highlight technical challenges encountered when adopting approaches commonly used for standardized languages. Most of the surveyed work uses techniques developed for written languages, while most Indigenous Peoples use their languages predominantly in a spoken form. Since modern approaches rely on data availability in written forms and computing power, the technical challenges are typically framed from the perspective of data scarcity (e.g. lack of parallel data, and lack of orthographic normalization, among others).

Inclusion of Indigenous People. Finally, we would like to point out that only a few exceptions consider an evaluation based on the needs and usage of Indigenous people. For example, Ahumada et al. (2022) provide detailed feedback given by the Indigenous descendants, as well as in-depth case study on usability.

5 Conclusions

In this paper, we survey existing computational resources for the most spoken Indigenous languages in Argentina. To better comprehend the Indigenous language diversity, we present an overview of demographic data for the Indigenous Peoples most present in the country. Among the seven Indigenous language families considered in this work, we find that most NLP applications and resources are developed for the Quechua, Tupí-Guaraní, and Mapuche families, often in varieties different from those spoken in Argentina. In contrast, resources available for the other language families are quite scarce.

6 Limitations

The authors of this work are culturally situated in academic contexts of hard access for indigenous identities in Argentina. Only one of us self-identifies as an indigenous descendant. We acknowledge that to work in this area an interdisciplinary approach is needed with members of the Indigenous communities being part of it.

Acknowledgements

Belu Ticona is partially funded by the generous support of the US National Science Foundation under grants CNS-2234895 and IIS-2327143.

[6]Universidad de la República, Uruguay.

References

Ezequiel Agustin Adamovsky. 2012. *El Color de la Nación Argentina: conflictos y negociaciones por la definición de un ethnos nacional, de la crisis al Bicentenario*. De Gruyter; Jahrbuch für Geschichte Lateinamerikas.

Ife Adebara and Muhammad Abdul-Mageed. 2022. Towards Afrocentric NLP for African Languages. pages 3814–3847.

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

Willem 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.

Willem FH Adelaar. 1995. Raíces lingüísticas del quichua de Santiago del Estero. *Actas de las segundas jornadas de lingüística aborigen*, 15:25–50.

Željko Agić and Ivan Vulić. 2019. JW300: A Wide-Coverage Parallel Corpus for Low-Resource Languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210. Association for Computational Linguistics.

Marvin Agüero-Torales, David Vilares, and Antonio López-Herrera. 2021. On the logistical difficulties and findings of Jopara Sentiment Analysis. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 95–102. Association for Computational Linguistics.

Nouman Ahmed, Natalia Flechas Manrique, and Antonije Petrović. 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, Toronto, Canada. Association for Computational Linguistics.

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

Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, et al. 2022. One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. *arXiv preprint arXiv:2203.13357*.

Honorio Apaza Alanoca, Brisayda Aruhuanca Chahuares, Kewin Aroquipa Caceres, and Josimar Chire Saire. 2023. Neural Machine Translation for Aymara to Spanish. In *Intelligent Systems and Applications*, pages 290–298, Cham. Springer International Publishing.

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.

Kenneth Beesley. 2003. *Finite-State Morphological Analysis and Generation for Aymara*. Association for Computational Linguistics.

Steven Bird. 2020. Decolonising Speech and Language Technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519. International Committee on Computational Linguistics.

Steven Bird. 2024. Must NLP be extractive? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14915–14929, Bangkok, Thailand. Association for Computational Linguistics.

Verena Blaschke, Hinrich Schütze, and Barbara Plank. 2023. A survey of corpora for Germanic low-resource languages and dialects. *arXiv preprint arXiv:2304.09805*.

Ronald Cardenas, Rodolfo Zevallos, Reynaldo Baquerizo, and Luis Camacho. 2018. Siminchik: A Speech Corpus for Preservation of Southern Quechua.

Cintia Carrió. 2014. Lenguas en argentina. notas sobre algunos desafíos. In Laura Kornfeld, editor, *De lenguas, ficciones y patrias*, pages 149–184. Universidad Nacional de General Sarmiento, Buenos Aires.

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. Association for Computational Linguistics.

Marisa Censabella. 1999. *Las Lenguas Indígenas de La Argentina: Una Mirada Actual*. Eudeba, Buenos Aires.

Marisa Inés Censabella. 2009. Chaco ampliado. In Inge Sichra, editor, *Atlas sociolingüístico de pueblos indígenas en América Latina*, pages 143–237. UNICEF.

Leonardo Cerno. 2010. Evidencias de diferenciación dialectal del guaraní correntino. *Cadernos de Etnolingüística*, 2(3).

Leonardo Cerno. 2013. *El Guaraní Correntino. Fonología, Gramática, Textos*. Peter Lang, Frankfurt.

Andrés Chandía. 2022-06. A Mapudüngun FST Morphological Analyser and its Web Interface. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6540–6547. European Language Resources Association.

Luis Chiruzzo, Marvin Agüero-Torales, Gustavo Giménez-Lugo, Aldo Alvarez, Yliana Rodríguez, Santiago Góngora, and Thamar Solorio. 2023. Overview of GUA-SPA at IberLEF 2023: Guarani-Spanish Code Switching Analysis.

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. European Language Resources Association.

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. European Language Resources Association.

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.

Johanna Cordova, Capucine Boidin, César Itier, Marie-Anne Moreaux, and Damien Nouvel. 2019. Processing Quechua and Guarani Historical Texts Query Expansion at Character and Word Level for Information Retrieval. In *Information Management and Big Data*, Communications in Computer and Information Science, pages 198–211. Springer International Publishing.

Ona De Gibert, Raúl Vázquez, Mikko Aulamo, Yves Scherrer, Sami Virpioja, and Jörg 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, Toronto, Canada. Association for Computational Linguistics.

Antonio Díaz-Fernández. 2006. Glosónimos aplicados a la lengua mapuche/glossonyms applied to the mapuche language. *Anclajes*, 10(10):95–111.

Wolf Dietrich. 1986. *El idioma chiriguano: gramática, textos, vocabulario*. Ediciones de cultura Hispánica, Madrid.

Javier Domingo and Dora Manchado. 2018. Usos cotidianos del tehuelche (aonekko 'a'ien) – homenaje a dora manchado. endangered languages archive.

Mingjun Duan, Carlos Fasola, Sai Krishna Rallabandi, Rodolfo Vega, Antonios Anastasopoulos, Lori Levin, and Alan W Black. 2020. A Resource for Computational Experiments on Mapudungun. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2872–2877. European Language Resources Association.

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

Abteen Ebrahimi, Manuel Mager, Shruti Rijhwani, Enora Rice, Arturo Oncevay, Claudia Baltazar, María Cortés, Cynthia Montaño, John E. Ortega, Rolando Coto-solano, Hilaria Cruz, Alexis Palmer, and Katharina Kann. 2023. Findings of the AmericasNLP 2023 Shared Task on Machine Translation into Indigenous Languages. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 206–219. Association for Computational Linguistics.

Bruno Estigarribia. 2015. Guaraní-Spanish Jopara mixing in a Paraguayan novel: Does it reflect a third language, a language variety, or true codeswitching? *Journal of Language Contact*, 8(2):183–222.

Leandro Martín Garber. 2022. *Sistema de identificación de idioma (LID) para grabaciones de entornos naturales bilingües en comunidades qom*. Ph.D. thesis, Universidad de Buenos Aires.

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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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

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, Marseille, France. European Language Resources Association.

Instituto Nacional de Estadística y Censos. 2024. Censo Nacional de Población, Hogares y Viviendas 2022: población indígena o descendiente de pueblos indígenas u originarios. Buenos Aires, Argentina.

Tommi Jauhiainen, H. Jauhiainen, and Krister Lindén. 2023. Tuning HeLI-OTS for Guarani-Spanish Code Switching Analysis.

Mayra Juanatey. 2020. *Relaciones entre eventos y referencialidad en quichua santiagueño: de la gramática al discurso*. Ph.D. thesis, Universidad de Buenos Aires, Buenos Aires.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schütze. 2023. Glotlid: Language identification for low-resource languages. *arXiv preprint arXiv:2310.16248*.

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. 2017. A Low-Resourced Peruvian Language Identification Model. In *CEUR Workshop Proceedings. CEUR-WS*, pages 57–63.

Zoey Liu, Crystal Richardson, Richard Hatcher, and Emily Prud'hommeaux. 2022. 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.

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

Manuel Mager, Elisabeth Mager, Katharina Kann, and Ngoc Thang Vu. 2023. Ethical Considerations for Machine Translation of Indigenous Languages: Giving a Voice to the Speakers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4871–4897. Association for Computational Linguistics.

Lorena Martín Rodríguez, Tatiana Merzhevich, Wellington Silva, Tiago Tresoldi, Carolina Aragon, and Fabrício Gerardi. 2022. *Tupian Language Resources: Data, Tools, Analyses*.

Nelsi Melgarejo, Rodolfo Zevallos, Hector Gomez, and John E. Ortega. 2022. 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.

Christian Monson, Lori Levin, Rodolfo Vega, Ralf Brown, Ariadna Font Llitjos, Alon Lavie, Jaime G Carbonell, Eliseo Cañulef, and Rosendo Huisca. 2004. Data Collection and Analysis of Mapudungun Morphology for Spelling Correction.

Christian Monson, Ariadna Font Llitjós, Vamshi Ambati, Lorraine Levin, Alon Lavie, Alison Alvarez, Roberto Aranovitch, Jaime G Carbonell, Robert Frederking, Erik Peterson, et al. 2008. Linguistic structure and bilingual informants help induce machine translation of lesser-resourced languages.

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. In *Strategies for developing machine translation for minority languages*.

Christopher Moseley, editor. 2010. *Atlas of the World's Languages in Danger*. UNESCO, Paris. 3rd edition.

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. 2021a. 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.

Arturo Oncevay. 2021b. 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. 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. Association for Machine Translation in the Americas.

John E. Ortega, Richard Castro Mamani, and Kyunghyun Cho. 2020. 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.

Begoña Pendas, Andres 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. Association for Computational Linguistics.

Andrés Osvaldo Porta. 2010a. Un parser para la morfología del quichua santiagueño con PC-KIMMO. In Víctor M. Castel and y Liliana Cubo de Severino, editors, *La renovación de la palabra en el bicentenario de la Argentina. Los colores de la mirada lingüística*. Editorial FFyL, UNCuyo, Mendoza.

Andrés Osvaldo Porta. 2010b. The use of formal language models in the typology of the morphology of amerindian languages. In *Proceedings of the ACL 2010 Student Research Workshop*, pages 109–114.

Mónica Quijada. 2004. *De Mitos Nacionales, Definiciones Cívicas y Clasificaciones Grupales. Los Índígenas en la Construcción Nacional Argentina, siglos XIX y XX.* Calidoscopio latinoamericano coord. Waldo Ansaldi, 425-450, Buenos Aires.

Alan Ramponi. 2024. Language varieties of italy: Technology challenges and opportunities. *Transactions of the Association for Computational Linguistics*, 12:19–38.

Andrea Rodrigo, Maximiliano Duran, and María Yanina Nalli. 2021. 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.

Adolfo A. Ríos, Pedro J. Amarilla, and Gustavo A. Giménez Lugo. 2014. Sentiment Categorization on a Creole Language with Lexicon-Based and Machine Learning Techniques. In *2014 Brazilian Conference on Intelligent Systems*, pages 37–43.

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

Inge Sichra, editor. 2009. *Atlas sociolingüístico de pueblos indígenas en América Latina*. UNICEF.

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.

Nllb Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. https://arxiv.org/abs/2207.04672v3.

The Langas Project. Langas (2012-), corpora diacrónicos de lenguas generales en línea (xvi-xix). Last visited 2024-15-01.

Guillaume Thomas. 2019. Universal Dependencies for Mbyá Guaraní. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 70–77. Association for Computational Linguistics.

Jörg Tiedemann. 2012. 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).

Atnafu Lambebo Tonja, Fazlourrahman Balouchzahi, Sabur Butt, Olga Kolesnikova, Hector Ceballos, Alexander Gelbukh, and Thamar Solorio. 2024. NLP Progress in Indigenous Latin American Languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 6972–6987.

Agüero Torales and Marvin Matías. 2022. *Machine Learning Approaches for Topic and Sentiment Analysis in Multilingual Opinions and Low-Resource Languages: From English to Guarani*. Universidad de Granada.

Raúl Vázquez, Yves Scherrer, Sami Virpioja, and Jörg Tiedemann. 2021. The helsinki submission to the americasnlp shared task. In *Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 255–264. The Association for Computational Linguistics.

J. Pedro Viegas Barros. 1999. Aspectos fonéticos y fonológicos de la dialectología del mapudungun en la Argentina. In *Actas de las III Jornadas de Etnolingüítica*, pages 141–149, Rosario. Universidad Nacional de Rosario, Facultad de Humanidades y Artes, Escuela de Antropología.

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

Rodolfo Zevallos, Luis Camacho, and Nelsi Melgarejo. 2022b. 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. 2022c. 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.

Appendix: Data sources, figures and tables

In this section we provide additional information about the methodology employed to collect and select the papers, about the data sources used to create Figure 1 (in Section 2), and also two tables providing an overview of the available corpora and tasks studied for the Indigenous language families Mapuche, Tupí-Guaraní, Quechua, and Aymara.

A. Employed protocol

In order to collect and select the papers from which the available resources and tasks were outlined, we gathered information from the Proceedings of the most relevant venues in the field: the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), the Association for Computational Linguistics (ACL) (main conference and workshops -such as Use of Computational Methods in the Study of Endangered Languages -ComputEL-, Workshop on Technologies for MT of Low Resource Languages (LoResMT), and Workshop on Deep Learning Approaches for Low-Resource NLP-), Language Resources and Evaluation Conference (LREC), Conference on Computational Linguistics (COLING), Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), and papers referred by selected papers from these sources.

B. Data sources

The right part of Figure 1, was created as follows. Language families and languages were based on those described by the literature (Adelaar 2012, 2010; Censabella 2009, 1999; Nercesian 2021; Ciccone 2010). Those languages were mapped to Ethnologue (Eberhard et al., 2023)[7], from where the ISO 639-3 code (i.e. macrolanguage identifier) was obtained. In cases when more than one variety is mentioned, the microlanguage identifier is provided in the *ISO 639-3* column. Data regarding Indigenous population and % of speakers was obtained from the Argentine National Census Data (Instituto Nacional de Estadística y Censos, 2024)[8]. These numbers only reflect the number of people self-identifying as part of an Indigenous community and, among them, the number of people who consider themselves to speak the Indigenous language of that community. For this reason, the data on speaking population may not fully represent the number of Indigenous language speakers: some communities have lost their ancestral languages and speak other Indigenous languages, while individuals outside the Indigenous population may still speak Indigenous languages (Ciccone 2010, Censabella 2009, pg. 159-169). The estimated number of speakers is based on the data from the previous two columns. Finally, vitality, the level of endangerment of the language, was completed according to Ethnologue web page. It is important to mention that not all information about Indigenous Peoples' communities provided by the national census, can be mapped to ISO languages straightforwardly. For instance, there is no special ISO code for the variety spoken by the Kolla community, considered in the census. Inversely, there is no Aymara community, among the communities considered in the census, which does not imply there is no speaker of this language, which is included in the ISO codes. In order to calculate number of speakers from the INDEC data, we only used as reference those Indigenous communities whose names are identical to the language name, except for the number of speakers of the Ava Guaraní (Chiriguano), where we summed, following (Ciccone, 2010), the number of people who self-identified as member of Chané and Isoceño groups. For some languages, there is no specific reference community. For instance, there is no distinction among Paraguayan Guaraní and Corrientes Guaraní or among Santiago del Estero Quichua and Cusco-Bolivian Quechua). In those cases, we vertically center aligned the available data (population, speaker ratio and speaking population).

Information regarding the number of speakers of languages in different countries can also be obtained from UNESCO's World Atlas of Languages (WAL) (Moseley, 2010)[9]. Nevertheless, we showed data from the national census (INDEC), which is the primary source for Argentina (and that differs from the information provided by UNESCO's WAL). The status of languages regarding their danger of disappearing can also be found in Glottolog (Hammarström et al., 2023) -which also

[7] https://www.ethnologue.com/

[8] The results were taken from https://censo.gob.ar/index.php/datos_definitivos_total_pais/, especially from tables 8 and 9.

[9] https://en.wal.unesco.org

shows all the varieties of the language families- and UNESCO's WAL. Finally, additional (and different) information regarding Indigenous languages spoken in Argentina can be seen in the Observatorio de los Derechos de los pueblos Indígenas y Campesinos[10].

C. Available corpora and tasks

Next, we present two tables (Tables 1 and 2), that provide an overview of the available corpora and tasks studied for the Indigenous language families Mapuche, Tupí-Guaraní, Quechua, and Aymara.

[10] https://www.soc.unicen.edu.ar/observatorio/index.php/22-articulos/106-unas-700-000-personas-mantienen-vivas-15-lenguas-indigenas-en-argentina

Family	Variety	Area	Task	Size & Description	Paper
Mapuche	Mapudungún	T	machine translation	384k sentences, medical domain	Pendas et al. (2023)
	Mapudungún	S	speech recognition, speech synthesis and machine translation	142 hs, transcribed audio, medical domain	Duan et al. (2020)
Tupí-Guaraní	Paraguayan Guaraní	T	machine translation	30k sentences, parallel Spanish-Guaraní data collected from news, folktales, articles, biographies (*)	Chiruzzo et al. (2022, 2020)
	Multiple varieties	T	multiple tasks	dependency treebanks, morphological and lexical datasets	Martín Rodríguez et al. (2022)
Quechua	South	T	machine translation	127k sentences, Spanish-Quechua parallel data, legal domain, biblical domain (*)	De Gibert et al. (2023); Ebrahimi et al. (2023); Ahmed et al. (2023); Agić and Vulić (2019), OPUS Corpus (Tiedemann 2012)
	Chanca, Collao	T	NER, POS tagging	384k sentences, multiple domains as religion, education, health, narrative, social (*)	Zevallos et al. (2022c)
	South., Central, North., Amazon	T	POS tagging	29k words, lexical resources for the development of a Quechua *wordnet* (*)	Melgarejo et al. (2022)
	Central, South	S	speech recognition, language identification, text-to-speech	220 hs, transcribed audio	Zevallos et al. (2022b)
	Collao	S	emotion recognition	15 hs, raw audio (*)	Paccotacya-Yanque et al. (2022)
	Chanca, Collao	S	speech recognition	97 hs, raw audio from radio shows	Cardenas et al. (2018)
Aymara	Unspecified	T	machine translation	900 Aymara-English pairs, lexical dataset from dictionaries	Tan (2023)
	Unspecified	T	machine translation	25k sentences aprox., multiples sources (legal, bliblical) (*)	De Gibert et al. (2023)
	Central	T	machine translation	3k sentences., multiples domains (news, health, informal and formal register) (*)	Team et al. (2022)
	Unspecified	T	machine translation	6.5k sentences, multiple domain (*)	Tiedemann (2012)

Table 1: Overview of available corpora for the Indigenous language families Mapuche, Tupí-Guaraní, Quechua and Aymara. South, North, and Amazon stand for Southern, Northern, and Amazonian, respectively. S and T stand for Speech and Text respectively. An asterisk (*) in the *Size & Description* column indicates that the resource is publicly available. The table shows research on languages spoken in Argentina, but for varieties spoken in other countries. Therefore, some language varieties mentioned in this section differ from those listed in Section 2.

Family	Task	Paper
Mapuche	MT (text)	Pendas et al. (2023); Levin et al. (2002); Duan et al. (2020); Monson et al. (2004, 2008); Llitjós et al. (2005); Monson et al. (2006)
	linguistic tools	Chandía (2022-06)
	educational tools (text)	Ahumada et al. (2022)
Tupí-Guaraní	language identification (text)	Cavalin et al. (2023)
	sentiment analysis	Ríos et al. (2014); Agüero-Torales et al. (2021)
	MT (text)	Góngora et al. (2022)
	code switching (text)	Chiruzzo et al. (2023); Jauhiainen et al. (2023); Torales and Matías (2022)
	topic modelling (text)	Torales and Matías (2022)
	educational tools (text)	Martín Rodríguez et al. (2022)
Quechua	NER	Zevallos et al. (2022c)
	POS tagging	Zevallos et al. (2022c)
	language identification (speech)	Paccotacya-Yanque et al. (2022); Linares and Oncevay (2017)
	MT (text)	Tiedemann (2012), Work on AmericasNLP shared task (Ahmed et al. 2023; Vázquez et al. 2021
	emotion recognition	Paccotacya-Yanque et al. (2022)
	speech recognition	Zevallos et al. (2022b)
	text-to-speech	Zevallos et al. (2022b)
	morphological analysis	Porta (2010a,b)
Aymara	MT (text)	Tan (2023); Alanoca et al. (2023); Oncevay (2021b)
	language identification (text)	Linares and Oncevay (2017)
	linguistic tools	Himoro and Pareja-Lora (2022); Beesley (2003)

Table 2: Overview of tasks studied for the Indigenous language families Mapuche, Tupí-Guaraní, Quechua, and Aymara. NER, POS, and MT stand for named entity recognition, part of speech, and machine translation respectively.