AmazoniaNLP: A Survey of Extreme Low-Resource Languages in the Peruvian–Brazilian Amazon

Rodolfo Zevallos[1] **Fabrício Carraro**[2] **John Ortega**[3]
[1] Universitat Pompeu Fabra, Spain
[2] Barcelona Supercomputing Center, Spain
[3] Northeastern University, The United States
rodolfojoel.zevallos@upf.edu, fabricio.carraro@bsc.es, j.ortega@northeastern.edu

Abstract

The Amazon basin along the Peru–Brazil border hosts extraordinary linguistic diversity, including many Indigenous languages whose speaker communities span national frontiers. Despite sustained documentation work, most remain extremely low-resource languages (ELRLs) for Natural Language Processing (NLP): reusable corpora are scarce, orthographies vary across countries and institutions, and basic tools such as tokenizers, taggers, and morphological analyzers are largely unavailable. We present a resource-oriented survey of five Indigenous languages of the Western Amazon—Matsés, Amahuaca, Kashinawa, Ticuna, and Kukama-Kukamiria—aimed at supporting more realistic NLP and speech work in extreme low-resource settings. Using a systematic search across academic venues, language archives, and public code/model repositories, we identify and cross-check available materials spanning lexical resources, text corpora, linguistic annotation, and speech collections. For each item we record practical reuse information, including the relevant task or modality, source location, and any stated access, licensing, or usage conditions. Our findings show strong cross-language asymmetries and fragmentation: most materials concentrate in documentation artifacts and lexicons, while standardized datasets with clear access and reuse conditions suitable for training and evaluation remain rare. We conclude with concrete recommendations to improve discoverability, normalize orthographic variation, and prioritize resource creation that maximizes interoperability across tools and benchmarks.

Keywords: extreme low-resource languages, Amazonian languages, language documentation

1. Introduction

Of the approximately 7,000 languages spoken worldwide, only a small fraction receives sustained computational attention (Eberhard et al., 2024). This imbalance motivates the study of ELRLs, which lack sufficient digital data to enable the effective application of data-driven NLP methodologies (Magueresse et al., 2020).

Recent work demonstrates that resource availability is not binary, but stratified. Joshi et al. (2020) show that most languages fall into categories with little or no NLP representation, introducing a taxonomy that distinguishes between moderate and ELRLs. In the latter, languages lack parallel corpora, annotated datasets, and even standardized orthographies, rendering conventional NLP pipelines inapplicable (Ranathunga et al., 2023; Mager et al., 2023).

This asymmetry is acute in Indigenous Amazonian languages. The UNESCO Atlas of the World's Languages in Danger classifies many as vulnerable or endangered (Moseley and Nicolas, 2010). These languages are typologically rich—featuring polysynthesis, complex morphology, and evidentiality—making them valuable for linguistic theory and cross-linguistic research. Yet their digital presence remains minimal, effectively excluding them from the web-scale corpora that underpin contemporary systems (Blasi et al., 2022).

Large-scale paradigms such as massively multilingual Machine Translation (MT), Automatic Speech Recognition (ASR), Text-to-Speech synthesis (TTS), and foundation models (Bommasani et al., 2021) presuppose data volumes that simply do not exist for these communities.

This study surveys the computational landscape for five Amazonian languages: Matsés (Fleck, 2024; Instituto Socioambiental (ISA)), Amahuaca (Chapman University (The Voice of Wilkinson), 2023), Kashinawa (Huni Kuĩ) (DoBeS (Documentation of Endangered Languages); Project, n.d.a), Ticuna (IBGE, 2022; Project, n.d.c), and Kukama-Kukamiria (Project, n.d.b). Spanning the Panoan and Tupian families and the Ticuna isolate, these languages vary in vitality but share a condition of ELRLs. While descriptive grammars exist, systematic NLP resources remain rare and unevenly distributed, with isolated advances such as an Amahuaca UD treebank, Kukama-Kukamiria UMR annotation, and task-specific multilingual speech checkpoints.

Treating all ELRLs as a homogeneous category ignores crucial factors such as dialectal fragmentation and restricted annotation capacity (Shahid et al., 2025). Therefore, a regionally grounded and linguistically informed perspective is required to advance research in these ELRLs.

This work presents a systematic analysis of the

Proceedings of SIGUL 2026 Joint Workshop with ELE, EURALI, and DCLRL @ LREC 2026, pages 280–287
11–12 May 2026.

computational landscape for these five languages with three objectives:

- **Assess digital availability:** cataloging corpora, lexicons, speech resources, and grammars from documentation repositories and NLP venues.
- **Summarize structurally relevant properties:** identifying how morphological complexity, tonal systems, and alignment patterns affect resource design, tokenization, and speech–text alignment in ELRL settings.
- **Identify barriers and pathways:** situating the discussion within broader critiques of multilingual representation in NLP (Joshi et al., 2020) and proposing concrete next steps for resource creation.

Given that foundation models (Bommasani et al., 2021) presuppose data conditions not met in these contexts, this study focuses on mapping existing resources and argues that foundational resource engineering is the most productive near-term direction under current data conditions. By foundational resource engineering, we mean the conversion of documentary materials into reusable computational assets: orthographic normalization across national traditions, digitization and normalization of lexica, conversion of ELAN/Toolbox materials into aligned speech–text corpora, explicit metadata and access conditions, and basic evaluation splits and baselines. This does not exclude transfer-based modeling; rather, it identifies the infrastructure on which such modeling depends.

2. Amazonian Languages: Overview

In this section, we provide a concise demographic and linguistic overview of the focal languages, emphasizing their typological diversity and current speaker populations.

Matsés (Mayoruna, ISO 639-3: `mcf`; Glottocode: `mats1244`) is a Panoan language spoken in the Yaquerana River basin on the Peru–Brazil border, with ∼3,300 speakers (Ministerio de Cultura de Perú, 2024c; Eberhard et al., 2024). It is ergative-absolutive, highly synthetic, with complex verbal morphology and a flexible SOV order (Fleck, 2003).

Amahuaca (ISO 639-3: `amc`; Glottocode: `amah1246`) is spoken in Peru and Acre (Brazil) by ∼500 people (Ministerio de Cultura de Perú, 2024a). It exhibits split-ergative alignment and flexible syntax depending on discourse focus (Clem, 2019, 2023).

Kashinawa (Huni Kuĩ, ISO 639-3: `cbs`; Glottocode: `cash1254`) is spoken along rivers in Peru and Brazil by over 13,000 speakers (Ministerio de Cultura de Perú, 2024b; Instituto Socioambiental, 2024). It is agglutinative, predominantly SOV, with split alignment: ergative for nouns, nominative-accusative for pronouns (Camargo, 2002).

Ticuna (ISO 639-3: `tca`; Glottocode: `ticu1245`) is a language isolate spoken in Brazil, Colombia, and Peru, with ∼65,000 speakers (Eberhard et al., 2024). It has complex tonal phonology, variable word order, and a rich system of nominal classification (Bertet, 2021; Montes Rodríguez, 2018).

Kukama-Kukamiria (ISO 639-3: `cod`; Glottocode: `coca1259`) is a Tupi-Guarani language spoken in the Peruvian Amazon, with ∼1,500 fluent speakers among ∼20,000 ethnic Kukama (Vallejos, 2016). It has stable SVO order, simplified morphology, and sociolectal differences in grammatical particles based on speaker gender (Vallejos, 2018).

3. Methodology

This survey covers NLP and speech resources for five Amazonian languages spoken across the Peru–Brazil border: Matsés, Amahuaca, Kashinawa, Ticuna, and Kukama-Kukamiria. They were selected because they represent the major language families present in the Western Amazon transborder region (three Panoan languages, one Tupí-Guaraní language, and one isolate), span a wide range of speaker population sizes, and differ in their degree of prior computational attention, thereby enabling a comparative analysis of resource availability across varying conditions.

We queried the ACL Anthology, IEEE Xplore, ACM Digital Library, Scopus, and SpringerLink, supplemented by grey literature from arXiv, Google Scholar, and Peruvian and Brazilian institutional repositories. Forward and backward citation chasing was applied to key references.

We searched DoBeS/TLA, the California Language Archive, DoReCo, SIL Language and Culture Archives, OLAC, and public code repositories (e.g., GitHub, Hugging Face) for reusable digital artifacts: corpora, lexica, time-aligned annotations (ELAN/Toolbox), conversion scripts, and pretrained model checkpoints.

Queries combined language names, ISO 639-3 codes, and orthographic variants (e.g., “Cashinahua,” “Kaxinawá,” “Huni Kuĩ”) with NLP and task-specific terms (e.g., “low-resource,” “MT,” “ASR,” “TTS,” “morphological analysis,” “Part-of-Speech (POS) tagging”).

Selection followed a three-stage process (title, abstract, full text). We included items that contribute reusable computational resources, annotated data, or modeling approaches for at least one target language, including theses and technical reports with usable digital material. Items mentioning the target languages only tangentially or focusing

exclusively on high-resource languages were excluded.

For each included item we recorded the resource type, relevant modality or task, source location (e.g., archive, repository, or publication), and, where explicitly reported in the source, coarse information about access conditions and scale (e.g., reported hours of audio or sentence counts). We also compiled short typological and orthographic notes for each language where these features directly affect tokenization, segmentation, or speech–text alignment. Searches were finalized in February 2026.

4. Language Profiles

The five languages analyzed in this study exhibit substantial typological variation along the transboundary Amazonian continuum between Peru and Brazil. Matsés, Amahuaca, and Kashinawa, belonging to the Panoan family (Fleck, 2013), share a predominantly ergative alignment system and a highly synthetic verbal morphology, characterized by extensive marking of tense, aspect, and mood (TAM) categories, as well as grammatical relations on the verb (Fleck, 2013; Clem, 2019; Camargo, 2002). This profile contrasts with Ticuna, a language isolate whose phonological organization is distinguished by a structurally complex tonal system (Bertet, 2021), and with Kukama-Kukamiria, a member of the Tupí-Guaraní family, which exhibits a relatively stable basic SVO word order, typologically unusual in the Amazonian context and associated with historical processes of language contact (Vallejos, 2016).

Despite these structural differences, all of them face a common challenge: high morphological variability, which increases lexical sparsity (Arnett and Bergen, 2025; Zevallos and Bel, 2023; Mager et al., 2020; Ortega et al., 2020; Ortega and Pillaipakkamnatt, 2018), and a lack of orthographic standardization, which hinders the direct application of modern NLP pipelines (Mager et al., 2018; Oliveira, 2024).

4.1. Descriptive and Lexicographical Foundations

The foundation of computational work for these languages lies in their limited linguistic documentation. Reference grammars exist for Matsés (Fleck, 2003), Amahuaca (Sparing-Chávez, 2012), and Kukama-Kukamiria (Vallejos, 2016), providing the formal descriptions necessary for rule-based systems. In the case of Ticuna, documentation has focused on specific regional varieties, such as the morphosyntactic foundation for Colombian Ticuna (Montes Rodríguez, 2004) and the semantic analysis of demonstratives in Peruvian Cushillococha Ticuna (Skilton, 2019).

Lexical resources are equally heterogeneous. While Matsés and Kukama-Kukamiria benefit from bilingual dictionaries (Fleck et al., 2012; Vallejos-Yopán and Amías, 2015), Kashinawa relies on early foundational vocabularies (Abreu, 1914) and more recent digitized lexicons distributed under restrictive licenses (DataScientia Foundation, 2023). More broadly, licensing conditions across the surveyed resources are highly heterogeneous: many archival materials lack any explicit license statement, while others restrict reuse to non-commercial or research-only purposes. This inconsistency complicates dataset aggregation and hinders reproducibility, underscoring the need for transparent and standardized licensing practices in ELRL resource creation.

4.2. Annotated Corpora and Benchmarks

The transition from descriptive materials, such as grammars, dictionaries, and text collections, to computationally exploitable resources for NLP continues to represent a significant challenge. Currently, Amahuaca stands out as a pioneer with the creation of the first dependency treebank under the Universal Dependencies (UD) standard (Angulo et al., 2025), a milestone that has not yet been achieved by the other languages in this study. In contrast, Kukama-Kukamiria has focused on deep semantic modeling through the Uniform Meaning Representation (UMR) framework, producing the first "gold-standard" graph-based dataset for an Amazonian language (Van Gysel et al., 2021; Wein, 2025). However, for languages such as Kashinawa and Ticuna, resources remain largely archived in documentation projects like DoBeS/TLA and CLA, requiring substantial effort to convert time-aligned recordings into structured text corpora (Skilton, 2021) (DoBeS (Documentation of Endangered Languages), 2017).

4.3. Speech Technologies and Downstream Applications

The scarcity of large-scale text corpora has prompted a shift toward speech-centered applications. Meta's Massively Multilingual Speech (MMS) project provides pretrained ASR, TTS, and LID resources (Pratap et al., 2023). According to Meta's official language coverage overview, all five focal languages are covered by MMS-ASR, while Ticuna and Kashinawa also have TTS and LID support; Matsés, Amahuaca, and Kukama–Kukamiria are ASR-only at time of access (Meta AI (facebook), 2023). This offers an immediate entry point for voice-based data augmentation, particularly for Ticuna and Kashinawa, where audio documentation is more abundant than transcribed text (Skilton, 2021) (Reiter, 2024). More recently, Meta's Omnilingual

ASR project (Omnilingual ASR Team, 2025) has extended multilingual speech recognition coverage to over 1,600 languages and emphasizes extensibility to previously unserved languages. We cite it here as recent context for speech technology in ELRL settings, while retaining MMS as the concrete system for which we verified task coverage of the five focal languages.

For higher-level tasks such as Machine Translation (MT), the landscape is significantly more limited. Parallel data is restricted to narrow-domain texts, mainly the Universal Declaration of Human Rights (UDHR) and religious translations, which serve as minimal test sets but are insufficient on their own for training robust neural systems. Transfer learning from related or even typologically distant languages, as well as synthetic data generation techniques, represent promising strategies to overcome this scarcity (Mager et al., 2018, 2023), though their effectiveness for the morphologically complex languages considered here remains largely unexplored.

5. Discussion and Findings

The analysis of the five focal languages reveals a landscape of significant contrasts regarding resource availability and NLP readiness. Building on broader discussions of resource inequality and ELRL stratification in NLP (Joshi et al., 2020; Magueresse et al., 2020), we discuss the key findings derived from the literature review and linguistic profiles.

5.1. Resource Availability Disparity

A marked hierarchy exists in the availability of deep annotation tools across the region. Amahuaca has reached a milestone with the creation of a treebank under the Universal Dependencies (UD) standard (Angulo et al., 2025), while Kukama-Kukamiria is the clearest case of deeper semantic annotation through the Uniform Meaning Representation (UMR) framework (Van Gysel et al., 2021; Bonn et al., 2024). Recent work has also begun to test the downstream utility of UMR for translation (Wein, 2025). Languages with larger speaker populations, such as Ticuna, still lack standardized syntactic resources of this caliber. This disparity suggests that resource development is driven more by targeted documentation projects and collaborations between theoretical and computational linguists than by raw population size.

We highlight UD and UMR milestones because they represent the highest-complexity annotation achievements for these languages to date. However, for practical NLP applications, monolingual text corpora and parallel corpora remain the most critically needed resources. Future resource-creation efforts should prioritize these foundational data types alongside deeper linguistic annotation.

5.2. The Challenge of Orthographic and Dialectal Variation

A recurring finding is the negative impact of orthographic non-standardization, particularly in cross-border languages like Matsés and Kashinawa. The divergence in writing practices between Peru and Brazil acts as a barrier to data interoperability. As seen in the Matsés case, extreme wordform sparsity is exacerbated by these variations, limiting the effectiveness of pretrained models unless explicit orthographic normalization is implemented (Mager et al., 2018; Oliveira, 2024).

5.3. Potential of Speech Technologies

Several of these languages, especially Ticuna and Kashinawa, have substantial audio archives but limited normalized transcriptions—exemplified by Ticuna materials in the California Language Archive totaling about 1,396 hours of recordings (1,227 hours audio), but only about 33 hours with transcriptions (Skilton, 2021). This asymmetry suggests that the most viable path for digital inclusion is through speech-centric approaches. The inclusion of all focal languages in Meta's MMS project (Pratap et al., 2023) provides a technological foundation that could mitigate the absence of large-scale parallel text corpora. Methodologically, re-speaking and oral translation techniques developed for unwritten African languages in the BULB project (Adda et al., 2016) offer a directly transferable paradigm for bootstrapping speech–text alignment from field recordings in similar low-resource settings.

5.4. Future Directions and Global Integration

To overcome the digital isolation of these languages, strategic priorities must focus on:

- **Normalization and Consolidation:** Developing automated orthographic conversion tools to unify data collected across Peru and Brazil, thereby maximizing available corpora.
- **Visibility in International Fora:** Amazonian languages should be made visible in under-resourced-language venues and infrastructures such as SIGUL, LT4All, and the LRE Map, and, where suitable datasets can be prepared, in evaluation campaigns such as AmericasNLP and IWSLT.
- **Multimodal Annotation:** Leveraging existing audio archives to generate automatic tran-

Language	Foundational resources	Scale indicator	NLP-ready asset	MMS tasks
Matsés	Grammar; bilingual dictionary	No directly comparable public corpus count was reported in the sources retained for this survey.	None identified	ASR
Amahuaca	Grammar; bilingual dictionary	UD corpus: 202 manually annotated sentences (1,028 word-level units; 1,928 morph-level units) (Angulo et al., 2025)	UD treebank	ASR
Kashinawa	Foundational lexica; archival recordings	TLA/DoBeS project page lists 286 archive bundles (146 open, 138 restricted, 2 registered), including 180 WAV bundles and 88 ELAN bundles (DoBeS (Documentation of Endangered Languages), 2017)	DoReCo-aligned speech corpus (Reiter, 2024)	ASR, TTS, LID
Ticuna	Grammar; bilingual dictionary; documentary collections	California Language Archive guide reports >1,396 hours of recordings, only 33 hours with transcriptions (Skilton, 2021)	None identified	ASR, TTS, LID
Kukama–Kukamiria	Grammar; bilingual dictionary	UMR release: 105 sentence-level annotations across 2 documents; 86 of these also have document-level annotation (Bonn et al., 2024)	UMR	ASR

Note: MMS task support in the last column is taken from Meta's official coverage table rather than inferred only from the general MMS paper (Meta AI (facebook), 2023).

Table 1: Overview of the main resource types identified for the five focal languages. The *Scale indicator* column reports one explicit public quantity per language when such metadata were available in the cited source.

scriptions using zero-shot techniques or cross-lingual transfer from typologically related languages (e.g., transferring knowledge from Matsés to Amahuaca).

- **Standardization Expansion:** Replicating the success of UD in Amahuaca and UMR in Kukama across other regional languages to enable global benchmarking.
- **Community-Centered Prioritization:** Resource development should be aligned with community-identified needs rather than inferred only from benchmark availability. In many settings, bilingualism in Spanish or Portuguese may reduce the immediate utility of some end-to-end translation scenarios, while low-cost tools such as keyboards, orthography converters, searchable lexica, and transcription support may offer more immediate value. Because we do not present direct needs-assessment data for all five languages, these priorities should be treated as hypotheses to be validated through participatory design with speaker communities.

6. Conclusion

This survey has mapped the computational landscape of five transborder Amazonian languages: Matsés, Amahuaca, Kashinawa, Ticuna, and Kukama-Kukamiria. Our analysis reveals a critical disconnect between the wealth of existing linguistic documentation and the scarcity of NLP-ready datasets. While the inclusion of these languages in global initiatives like Meta's MMS project provides a foundational entry point, the lack of standardized text corpora and the prevalence of orthographic divergence across the Peru–Brazil border remain significant bottlenecks.

Moving forward, we argue that the most impactful contributions in the near term will stem from targeted resource engineering—the systematic conversion of existing documentation into computationally reusable formats such as standardized lexica, orthographic normalization tools, morphological analyzers, and aligned speech–text datasets—rather than novel modeling. Transforming archival materials into structured lexica and aligning speech–text datasets offer a pragmatic path toward digital inclusion. Furthermore, integrating these languages into evaluation campaigns such as AmericasNLP and IWSLT, and into under-resourced-language venues and infrastructures such as SIGUL, LT4All, and the LRE Map, is essential for fostering visibility and discoverability. Ultimately, sustainable progress in Amazonian NLP must be anchored in clear access and reuse conditions, together with long-term collaborations with speaker communities, ensuring that technological advancements are both technically robust and socially responsible.

7. Ethics Statement

This is a survey of publicly described resources and does not involve primary data collection from speaker communities. We have endeavored to respect the access and licensing conditions of all materials cited.

8. Limitations

This survey is limited to resources and tools discoverable through the academic venues, archives, and repositories described in Section 3 as of February 2026. Relevant materials hosted on institutional or community-internal platforms, or available only in unpublished form, may have been missed. We do not evaluate the quality or usability of the resources cataloged, and licensing information is reported as found at the time of access.

9. Acknowledgements

We thank the speaker communities and language workers associated with the languages surveyed for making documentation and educational materials available and for ongoing efforts to sustain these languages. We also thank the maintainers and staff of documentation archives and repositories consulted in this study for curating and providing access to materials and metadata. Finally, we thank the SIGUL 2026 reviewers for constructive feedback that improved the camera-ready version.

10. References

João Capistrano de Abreu. 1914. *Rã-txa hu-ni-ku-ĩ: a lingua dos caxinauás do Rio Ibuaçu, affluente do Muru (Prefeitura de Tarauacá)*. Typographia Leuzinger, Rio de Janeiro.

Gilles Adda, Sebastian Stüker, Martine Adda-Decker, Odette Ambouroue, Laurent Besacier, David Blachon, Hélène Bonneau-Maynard, Pierre Godard, Fatima Hamlaoui, Dmitry Idiatov, Guy-Noël Kouarata, Lori Lamel, Emmanuel-Moselly Makasso, Annie Rialland, Mark Van de Velde, François Yvon, and Sabine Zerbian. 2016. Breaking the unwritten language barrier: The BULB project. In *Procedia Computer Science*, volume 81, pages 8–14. Elsevier. SLTU 2016.

Candy Angulo, Pilar Valenzuela, and Roberto Zariquiey. 2025. Universal Dependencies for Amahuaca. In *Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 150–154, Honolulu, Hawaii, USA. Association for Computational Linguistics.

Catherine Arnett and Benjamin Bergen. 2025. Why do language models perform worse for morphologically complex languages? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6607–6623, Abu Dhabi, UAE. Association for Computational Linguistics.

Denis Bertet. 2021. Tikuna: a ten-toneme language in amazonia. *Amerindia*, 43:55–101.

Damian 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, Dublin, Ireland. Association for Computational Linguistics.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Julia Bonn, Matthew J. Buchholz, Jayeol Chun, Andrew Cowell, William Croft, Lukas Denk, Sijia Ge, Jan Hajič, Kenneth Lai, James H. Martin, Skatje Myers, Alexis Palmer, Martha Palmer, Claire Benet Post, James Pustejovsky, Kristine Stenzel, Haibo Sun, Zdeňka Urešová, Rosa Vallejos, Jens E. L. Van Gysel, Meagan Vigus, Nianwen Xue, and Jin Zhao. 2024. Building a broad infrastructure for uniform meaning representations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2537–2547, Torino, Italia. ELRA and ICCL.

Eliane Camargo. 2002. Ergatividade cindida em caxinawá. In Ana Suelly Arruda Câmara Cabral and Aryon Dall'Igna Rodrigues, editors, *Línguas indígenas brasileiras: fonologia, gramática e história: Atas do I Encontro Internacional do Grupo de Trabalho sobre Línguas Indígenas da ANPOLL*. EDUFPA, Belém.

Chapman University (The Voice of Wilkinson). 2023. A historic event for the amahuaca people of peru. States that 330 people speak Amahuaca. Accessed 2026-02-14.

Emily Clem. 2019. Amahuaca ergative as agreement with multiple heads. *Natural Language & Linguistic Theory*, 37(3):785–823.

Emily Clem. 2023. The expression of time in amahuaca switch-reference clauses. *Languages*, 8(2):134.

DoBeS (Documentation of Endangered Languages). Cashinahua: Language description page. Mentions a Cashinahua ethnic community of around six thousand and discusses sociolinguistic tendencies. Accessed 2026-02-14.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2024. *Ethnologue: Languages of the World*, 27 edition. SIL International, Dallas, Texas.

David W. Fleck. 2024. Why is matses an onomatopoetic language? *Anthropological Linguistics*, 64(1–2):1–35. Issue dated Spring/Summer 2022; published on Project MUSE in 2024. Accessed 2026-02-15.

David William Fleck. 2003. *A grammar of Matses*. Ph.D. thesis.

David William Fleck. 2013. *Panoan languages and linguistics.(Anthropological papers of the American Museum of Natural History, no. 99)*. American Museum of Natural History.

IBGE. 2022. Censo demográfico 2022: resultados sobre línguas indígenas (tikuna). Accessed 2026-02-14.

Instituto Socioambiental. 2024. Povos indígenas no brasil: Huni kuin (kaxinawá). Accessed: 2026-02-19.

Instituto Socioambiental (ISA). Matsés (indigenous peoples in brazil). Accessed 2026-02-14.

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, Online. Association for Computational Linguistics.

Manuel Mager, Özlem Çetinoğlu, and Katharina Kann. 2020. Tackling the low-resource challenge for canonical segmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5237–5250, Online. Association for Computational Linguistics.

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, Toronto, Canada. Association for Computational Linguistics.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*.

Ministerio de Cultura de Perú. 2024a. Base de datos de pueblos indígenas u originarios: Amahuaca. Accessed: 2026-02-19.

Ministerio de Cultura de Perú. 2024b. Base de datos de pueblos indígenas u originarios: Kashinawa. Accessed: 2026-02-19.

Ministerio de Cultura de Perú. 2024c. Base de datos de pueblos indígenas u originarios: Matsés. Accessed: 2026-02-19.

María Emilia Montes Rodríguez. 2004. *Morfosintaxis de la lengua Tikuna (Amazonía Colombiana)*. Number 15 in Lenguas aborígenes de Colombia: Descripciones. Universidad de los Andes, Bogotá.

María Emilia Montes Rodríguez. 2018. Género, clasificación y nombres ligados en tikuna (amazonia colombiana). *Revista Brasileira de Linguística Antropológica*, 6(1). Portal metadata labels the issue as v.6 n.1 (2014); published online 2018-12-21.

Christopher Moseley and Alexandre Nicolas, editors. 2010. *Atlas of the World's Languages in Danger*, 3rd edition. Memory of Peoples. UNESCO, Paris, France.

Sanderson Castro Soares de Oliveira. 2024. Contribuições para a sociolinguística do kaxinawá, uma língua pluricêntrica. *Raído*, 18(46):378–397.

Omnilingual ASR Team. 2025. Omnilingual ASR: Open-source multilingual speech recognition for 1600+ languages.

John E Ortega, Richard Castro Mamani, and Kyunghyun Cho. 2020. Neural machine translation with a polysynthetic low resource language. *Machine Translation*, 34(4):325–346.

John E Ortega and Krishnan Pillaipakkamnatt. 2018. Using morphemes from agglutinative languages like quechua and finnish to aid in low-resource translation. *Technologies for MT of Low Resource Languages (LoResMT 2018)*, page 1.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. Scaling speech technology to 1,000+ languages. Accessed 2026-02-14.

Endangered Languages Project. n.d.a. Cashinahua (elp language page). Accessed 2026-02-14.

Endangered Languages Project. n.d.b. Cocama-cocamilla (elp language page). Accessed 2026-02-14.

Endangered Languages Project. n.d.c. Ticuna (elp language page). Accessed 2026-02-14.

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.

Farhana Shahid, Mona Elswah, and Aditya Vashistha. 2025. Think outside the data: Colonial biases and systemic issues in automated moderation pipelines for low-resource languages. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(3):2331–2344.

Amalia Skilton. 2021. Ticuna (tca) language documentation: A guide to materials in the california language archive. *Language Documentation and Conservation*, 15:153–189.

Amalia Horan Skilton. 2019. *Spatial and Non-spatial Deixis in Cushillococha Ticuna*. Phd dissertation, University of California, Berkeley.

Margarethe Sparing-Chávez. 2012. Aspects of amahuaca grammar: An endangered language of the amazon basin. *Dallas: SIL International*.

Rosa Vallejos. 2016. *A grammar of Kukama-Kukamiria: A language from the Amazon*, volume 13. Brill.

Rosa Vallejos. 2018. Kukama–kukamiria. *International Journal of American Linguistics*, 84(S1):S129–S147.

Jens E. L. Van Gysel, Meagan Vigus, Lukas Denk, Andrew Cowell, Rosa Vallejos, Tim O'Gorman, and William Croft. 2021. Theoretical and practical issues in the semantic annotation of four indigenous languages. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 12–22, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shira Wein. 2025. Can uniform meaning representation help GPT-4 translate from indigenous languages? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 278–285, Vienna, Austria. Association for Computational Linguistics.

Rodolfo Zevallos and Nuria Bel. 2023. Hints on the data for language modeling of synthetic languages with transformers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12508–12522, Toronto, Canada. Association for Computational Linguistics.

11. Language Resource References

DataScientia Foundation. 2023. *Cashinahua UKC Lexicon*. LMF-XML; small seed lexicon. License shown as CC BY-NC-SA. Accessed 2026-02-14.

DoBeS (Documentation of Endangered Languages). 2017. *Cashinahua: Documentation project page*. Accessed 2026-02-14.

Fleck, David William and Bëso, Fernando Shoque Uaquí and Huanán, Daniel Manquid Jiménez. 2012. *Diccionario matsés-castellano: con índice alfabético castellano-matsés e índice semántico castellano-matsés*. Tierra Nueva.

Meta AI (facebook). 2023. *MMS language coverage overview (ASR/TTS/LID)*. Accessed 2026-02-20.

Reiter, Sabine. 2024. *Cashinahua DoReCo dataset*. In Seifart, Frank and Paschen, Ludger and Stave, Matthew (eds.), Language Documentation Reference Corpus (DoReCo) 2.0. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Accessed 2026-02-14.

Rosa Vallejos-Yopán and Rosa Amías. 2015. *Diccionario Kukama-Kukamiria / Castellano*. FORMABIAP, Iquitos, Perú. 9.7MB.