Community-Driven Data Practices for Advancing Ethical and Equitable AI in Low-Resource Language Contexts

Charles Nimo
Georgia Institute of Technology
Atlanta, Georgia, USA
nimo@gatech.edu

Shuheng Liu
Georgia Institute of Technology
Atlanta, Georgia, USA
sliu775@gatech.edu

Amy Z. Chen
Georgia Institute of Technology
Atlanta, Georgia, USA
amychen@gatech.edu

Ramaravind Kommiya Mothilal
University of Toronto
Toronto, Ontario, Canada
ram.mothilal@mail.utoronto.ca

Michael L. Best
Georgia Institute of Technology
Atlanta, Georgia, USA
mikeb@gatech.edu

Abstract

This workshop examines the challenges and ethical concerns surrounding AI data collection, particularly for low-resource languages in the Global South. It critiques traditional data annotation models for perpetuating inequities and exploitative labor practices while advocating for community-driven approaches that integrate local cultural contexts. By empowering local contributors as co-creators, the workshop seeks to develop sustainable and ethically sound data practices that address historical imbalances. Aligned with ACM COMPASS 2025's theme, this workshop will foster dialogue among researchers, practitioners, and community representatives to generate actionable recommendations, promote equitable and participatory practices, and build resilient interdisciplinary partnerships for a more inclusive future in AI development.

CCS Concepts

• **Human-centered computing** → **Collaborative content creation**; • **Computing methodologies** → *Language resources.*

Keywords

Community-driven data collection, Participatory design, Low-resource languages, Ethical AI, Natural language processing

ACM Reference Format:
Charles Nimo, Shuheng Liu, Amy Z. Chen, Ramaravind Kommiya Mothilal, and Michael L. Best. 2025. Community-Driven Data Practices for Advancing Ethical and Equitable AI in Low-Resource Language Contexts. In *ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies (COMPASS '25), July 22–25, 2025, Toronto, ON, Canada.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3715335.3737684

1 Introduction

In today's rapidly evolving landscape of low-resource language technologies, rethinking data collection through a community-driven perspective is not only timely but essential. Web scraping, such as the Common Crawl project[1], has become one of the most prevailing ways for language data collection, but for low-resource languages, especially those that are not even used by their native speakers online, data collection has become increasingly more difficult.

In contrast, more researchers have turned to the community-driven data collection methods. Such methods often result in the collection of high-quality data [2] and allow communities speaking the language to be directly benefited from contributing to the data collection process [1]. Moreover, innovative frameworks such as equitable licensing and participatory governance [4] not only promote the redistribution of benefits but also foster local ownership and capacity building, paving the way for AI systems that are both culturally relevant and ethically sound. Nevertheless, the current language data collection in the Global South, where most speakers of low-resource languages reside, suffers from issues such as precarious employment, substandard wages, and conditions that undermine both financial stability and psychological well-being for data annotators [3]. How to develop a framework for sustainable and ethical community-driven low-resource language data collection remains an open discussion.

Aligned within ACM COMPASS 2025's theme of exploring how *"technological practices emerge as solutions, but also contributors to the complexities they claim to address,"*[2] this workshop serves as a critical platform for dialogue surrounding the community-driven low-resource language data collection process. It is designed to bring together researchers, practitioners, and community representatives to share insights, discuss challenges, and chart pathways toward a more inclusive and equitable future in low-resource language technology development. We hope to achieve the following goals with this workshop:

- Encourage community-driven practices in low-resource language data collection by centering the voices and expertise of local communities, particularly from the Global South and speakers of low-resource languages.
- Discuss the ethical challenges and labor concerns, and highlight innovative frameworks that empower communities rather than exploit them.
- Create a platform for researchers, practitioners, and community representatives to exchange ideas, share case studies, and develop actionable pathways for ethical data practices.

COMPASS '25, Toronto, ON, Canada

ACM ISBN 979-8-4007-1484-9/2025/07
https://doi.org/10.1145/3715335.3737684

[1] https://commoncrawl.org
[2] https://compass.acm.org/call-for-papers/

- Engage with how technology solutions can be grounded in a deeper and situated understanding of local contexts, avoiding the reinforcement of existing inequities.

2 Workshop Format

We propose a 3.5 hour (half-day) hybrid workshop, given that many relevant participants may be located in the Global South or otherwise face barriers to travel. We will have organizers present both in person and virtually. To make this workshop an inclusive and accessible experience, we plan to have a 360-degree webcam in the in-person room and (when interacting as a whole group) project online participants in person, display real-time captioning, and have an in-person organizer responsible for managing the hybrid set-up and participation, including monitoring incoming chats.

2.1 Activities and Tentative Schedule

We plan to include a variety of sessions and activities over the course of the workshop to encourage active interaction and idea generation, and allow for opportunities for participants of different levels of familiarity with the topic to learn from and discuss with experts and each other.

- **Icebreaker and Idea Jam:** We will begin our workshop with a brief interactive session to spark dialogue and set a collaborative tone around the workshop's core themes. We will present a series of short, real-world case studies about AI technologies in the context of low-resource languages. Each case study will be followed by exploratory and analytical questions, open discussion, and collective reflection on data practices. This will help lay the groundwork for informed questions and comments during the subsequent panel session.
- **Panel discussion:** We will then host a panel discussion with several invited researchers and practitioners with extensive experience in fostering equitable language technology development, particularly for low-resourced languages and in partnership with communities. Invited experts will be asked to give a brief presentation on their work as related to the workshop theme (prospectively, three speakers giving 10-minute talks). Potential topics include promoting broader participation and cultural representation in low-resource language NLP and technology development, sharing case studies of community-driven data collection efforts, ethical labor practices for data collection and annotation, and decolonial approaches to community data governance. Participants will be provided sticky notes to jot down ideas and key points from the talks, relating to or building upon some of the ideas generated in the opening session.
- **Breakout session:** In the final breakout session, participants will be divided into small groups to explore ethical and inclusive considerations in designing an annotation scheme for an AI application. The groups will be offered a selection of diverse, real-world AI use cases to choose from—or they may propose their own AI application of interest. Guided by a set of structured tasks, each group will be expected to collaboratively analyze their chosen scenario, determine their target community, and plan an annotation task. By the end of the session, groups will synthesize their ideas into three core themes: (a) aspirations of ideal annotator characteristics along with the constraints that may influence final selection, (b) communication procedures with annotators, including decisions around hiring process and interface design, and (c) expectations from annotators and assessment methods. These group insights will then be shared in brief presentations, followed by an open discussion where participants can exchange feedback, iterate on their initial ideas, and reflect collectively on unresolved questions in engaging with diverse communities for the inclusive development of AI systems.

The table below outlines a tentative schedule for the workshop, for a total duration of 3.5 hours.

Duration	Activity
10 min	Opening remarks and workshop overview
20 min	Icebreaker and idea jam
60 min	Panel discussion and Q&A with invited speakers
15 min	Tea / coffee break
60 min	Breakout session
30 min	Synthesis session: Report back and reflect
15 min	Closing remarks and next steps

2.2 Invited Speakers (candidates)

We have reached out to several senior researchers in the field to prospectively invite them as speakers for the panel discussion, should the workshop be accepted. Speakers will be finalized during planning leading up to the workshop, and provided with more details about the workshop goals and format so that they may shape their talks accordingly.

- **Dr. Monojit Choudhury** is a professor of Natural Language Processing at Mohamed bin Zayed University of Artificial Intelligence. His research interests center around the convergence of Language Technology and Society, exploring pivotal inquiries such as the learning and (mis)representation of linguistic and cultural diversity by foundation models. His research delves into the impact of representational disparities on present and future of technology use, and their impact on linguistic and cultural dynamics in the real world. A focal objective is to devise fair and equitable language technologies, contributing to an inclusive future. Additionally, Dr. Choudhury is keen on leveraging generative AI for planetary-scale quantitative investigations into and modeling of culture.
- **Dr. Julia Kreutzer** is a Senior Research Scientist at Cohere Labs, where she conducts research on large language models, focusing on multilinguality, reinforcement learning, and evaluation. Previously, she worked at Google Research. She is passionate about advancing NLP for under-resourced languages and underrepresented groups. Her long-term goal is to foster an inclusive NLP ecosystem by lowering entry barriers for new researchers, supporting research in diverse language communities, and engaging the public in machine learning to deepen understanding of machine learning methods and their limitations.

- **Ned Cooper** is a PhD candidate at The Australian National University College of Engineering, Computing & Cybernetics, a Research Affiliate of the MINT Lab, and an incoming postdoctoral researcher at Cornell University. His research sits at the intersection of human-AI interaction and AI policy, specifically exploring how to empower communities to actively shape machine learning development.

3 Expected Outcomes

We hope for workshop participants to walk away with:

- **Strengthened relationships** with researchers and practitioners sharing an interest in community-driven language data practices and technology development, opening avenues for future interdisciplinary and cross-regional projects.
- **Increased understanding** of the systemic challenges in data collection for low-resource languages and substantive equitable engagement of community members in the process.
- **Co-generated ideas and resources** including guidelines, best practices, and open questions for research around community engagement, language data collection and annotation, and language technology development.

To encourage the continued building of this research community, we plan to document and share workshop outputs and resources with participants as well as more broadly:

- Create a shared digital repository of resources for participants (e.g., as a Google Drive folder) with relevant research papers and reports, speaker presentation slides (with permission), and collated notes from the workshop.
- Share contact information of participants with each other (with permission).
- Leverage multiple online channels such as social media posts and community blogs to share insights and learnings from the workshop with a broader audience.

4 Call for Participation

The collection of low-resource language data for AI technology development has always remained difficult, but the exponentially increasing need for data for the current paradigm of training large language models can further marginalize these languages. Even with some community-driven data collection methods developed, there are various ethical issues to be considered, given that many of these languages are spoken in the Global South, where such technology development might not always benefit, or could even be potentially harmful to the communities.

This workshop invites researchers with diverse experiences—including but not limited to community-based research, low-resource language technology development, data collection and analysis—as well as practitioners, developers, and community representatives to join us in discussing potential pathways for community-driven data practices for low-resource language technologies and the ethical challenges associated with them. Our goal is for this workshop to serve as a platform for sharing ideas and fostering collaboration. We will provide a link to a Google form for registration, but we also welcome anyone interested to walk in and participate in the discussion, whether attending in person or virtually.

5 Organizers' Biographies

Charles Nimo is a second-year Ph.D. student in Computer Science at Georgia Institute of Technology, co-advised by Dr. Michael Best and Dr. Irfan Essa. Currently, his research focuses on understanding and improving the cultural awareness and adaptability of large language models. He is also interested in the broader field of responsible AI and working with low-resource languages.

Shuheng Liu is a 4th year PhD student in Computer Science at the School of Interactive Computing in Georgia Institute of Technology. His research focuses on the technology development for low-resource Sino-Tibetan languages and community-based methods for data collection.

Amy Chen is a PhD student in Human-Centered Computing at the Georgia Institute of Technology. Her research examines the design and role of digital platform-based language technologies in supporting public health communication. She is part of several interdisciplinary community-engaged research coalitions addressing health disparities in the state of Georgia, and has experience working in India and East Africa.

Ramaravind Kommiya Mothilal is a PhD student in Information and a graduate fellow of the Data Sciences Institute at the University of Toronto. He is broadly interested in advancing responsible AI by emphasizing real-world considerations over abstract idealisms. His current research focuses on developing methodological frameworks to articulate and evaluate mundane choices and assumptions of AI practitioners.

Dr. Michael Best is Executive Director of the Institute for People and Technology (IPaT) and Professor with the Sam Nunn School of International Affairs and the School of Interactive Computing at Georgia Institute of Technology where he directs the Technologies and International Development Lab. His research explores the promise, and the peril, of information and communication technologies (ICTs) in social, economic, and political development, with a recent focus on low-resource languages. He hopes to create new forms for inclusive innovation, and has particularly sought to build partnerships with researchers and communities in Africa and Asia.

References

[1] Basil Abraham, Danish Goel, Divya Siddarth, Kalika Bali, Manu Chopra, Monojit Choudhury, Pratik Joshi, Preethi Jyoti, Sunayana Sitaram, and Vivek Seshadri. 2020. Crowdsourcing Speech Data for Low-Resource Languages from Low-Income Workers. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 2819–2826. https://aclanthology.org/2020.lrec-1.343/

[2] Harshita Diddee, Anurag Shukla, Tanuja Ganu, Vivek Seshadri, Sandipan Dandapat, Monojit Choudhury, and Kalika Bali. 2024. INMT-Lite: Accelerating Low-Resource Language Data Collection via Offline Interactive Neural Machine Translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italia, 9097–9109. https://aclanthology.org/2024.lrec-main.797/

[3] Chinasa T. Okolo and Marie Tano. [n. d.]. Moving toward truly responsible AI development in the global AI market. https://www.brookings.edu/articles/moving-toward-truly-responsible-ai-development-in-the-global-ai-market/

[4] Jenalea Rajab, Anuoluwapo Aremu, Everlyn Asiko Chimoto, Dale Dunbar, Graham Morrissey, Fadel Thior, Luandrie Potgieter, Jessico Ojo, Atnafu Lambebo Tonja, Maushami Chetty, Onyothi Nekoto, Pelonomi Moiloa, Jade Abbott, Vukosi Marivate, and Benjamin Rosman. 2025. The Esethu Framework: Reimagining Sustainable Dataset Governance and Curation for Low-Resource Languages.

doi:10.48550/arXiv.2502.15916 arXiv:2502.15916 [cs] version: 1.