CENTER FOR DEMOCRACY & TECHNOLOGY

Beyond English-Centric AI

Lessons on Community Participation from Non-English NLP Groups

October 2024

Authored by
Evani Radiya-Dixit, *CDT Summer Fellow, CDT AI Governance Lab*
Miranda Bogen, *Director, CDT AI Governance Lab*

Many leading language models are trained on nearly a thousand times more English text compared to text in other languages. These disparities in large language models have real-world impacts, especially for racialized and marginalized communities. For example, they have resulted in inaccurate medical advice in Hindi, led to wrongful arrest because of mistranslations in Arabic, and have been accused of fueling ethnic cleansing in Ethiopia due to poor moderation of speech that incites violence.

These harms reflect the English-centric nature of natural language processing (NLP) tools, which prominent tech companies often develop without centering or even involving non-English-speaking communities. In response, region- and language-specific research groups, such as Masakhane and AmericasNLP, have emerged to counter English-centric NLP by empowering their communities to both contribute to and benefit from NLP tools developed in their languages. Based on our research and conversations with these collectives, we outline promising practices that companies and research groups can adopt to broaden community participation in multilingual AI development.

The Role of Non-English NLP Groups in Multilingual AI

In recent years, tech companies and research groups have been advancing *multilingual language models*: large language models trained on text or speech from multiple languages rather than just one, which seems to enable them to learn more general rules of language. Multilingual models are already used in a variety of applications like content moderation, machine translation, transcription tools, and customer service chatbots.

Previously, CDT has written about the limits of multilingual language models in non-English languages. Historically and presently, non-English languages, especially regional and indigenous ones, have faced erasure by imperial powers. For example, colonial governments have imprisoned African authors for writing in their native tongue. This erasure has led to the hegemony of English as the default language for international media, politics, and business. Consequently, while many other languages have limited high-quality digital data and are considered low-resourced, there is a wealth of online content in English that prominent tech companies use to train, tune, and test language models. As a result, non-English languages are often under-represented or not supported in mainstream NLP technology built by companies, which, in turn, drives more English-centric NLP research and accelerates the disappearance of indigenous and other endangered languages.

Even when prominent tech companies — most of which are based in the United States — develop the multilingual capabilities of their language models, they often reinforce the hegemony of the English language by not dedicating sufficient attention or resources to involving the communities who speak the non-English languages their tools aim to support. These companies typically use a "one model, all languages" approach, illuminated by initiatives like Meta's No Language Left Behind project and Google's 1000 Languages Initiative. Most of these efforts rely on machine-translated text, overlook cultural context, and have limited engagement with local communities, though some NLP efforts such as Mozilla's Common Voice and Cohere's PRISM Alignment Project are exploring structures for stronger community participation in language model development.

NLP research groups have emerged as alternative spaces to counter the English-centric approach of U.S. tech companies and mainstream NLP spaces. Often housed in academic or grassroots settings, these groups aim to strengthen NLP research for languages within their own communities. Based on conversations with NLP researchers and comprehensive desk research, we identified eight such groups that exemplify these efforts, focused on Arabic, Indian, African, Indonesian, and indigenous American languages.

These groups demonstrate that multilingual AI can be developed differently from the mainstream approaches of prominent tech companies by engaging communities to create culturally relevant technology that integrates local knowledge and serves community needs. While some NLP groups are structured around a centralized academic NLP conference or university, others are organized in a more decentralized or distributed grassroots manner. Regardless of their structure, the groups we examined operate openly, enabling people to join through mailing lists or messaging apps like Slack, Discord, or WhatsApp. Common features of these NLP groups include celebrating linguistic diversity, offering open-source datasets and models, building community among NLP researchers, making NLP research accessible to non-experts, and using participatory methods to involve native speakers and language experts in the development of NLP tools.

Group Name	Languages	Group Type	Goal
AI4Bharat	Indian languages	Research lab at university	Craft open-source datasets, tools, models, and applications for Indian languages
AmericasNLP	Indigenous languages of the Americas	Conference workshop	Promote indigenous American NLP and connect underrepresented groups with NLP communities
ARBML	Arabic language and its dialects	Research group	Enrich Arabic content with open-source ML projects and advance ML learning among Arabs
HausaNLP	Hausa and other African languages	Research group	Promote Hausa NLP by advancing language resources, research, and collaboration
IndoNLP	Indonesian languages	Research community	Advance Indonesian NLP research through new data resources and open-source projects
Masakhane	African languages	Grassroots community	Strengthen NLP research in African languages, for Africans, by Africans
NorthAfricanNLP	North African languages and dialects	Affinity group with events at conferences	Build a North African NLP research community and increase visibility of North Africans in NLP
SIGARAB	Arabic language and its dialects	Special interest group with workshops and conferences	Provide a forum for researchers to share and discuss their Arabic NLP work

Table 1. Overview of Non-English NLP Groups

Best Practices for Community Participation in NLP

Based on our research of non-English NLP groups, we identify three best practices for meaningfully involving communities throughout the lifecycle of multilingual AI development. These practices help address the limitations observed in multilingual models when community participation is not adequately centered, from problem definition to dataset creation to model and application development.

1. Meaningfully involve diverse communities in tailoring NLP datasets to local cultures and geographies

Many non-English NLP groups recruit native speakers and language experts to create datasets that reflect local cultures, which are then used to develop NLP models and applications. In this process, NLP groups integrate culturally and geographically relevant topics, which helps reduce the risk of imposing English-based assumptions on other languages, a major challenge in developing multilingual AI given the abundance of online content in English. These NLP groups also involve diverse community members to capture different language practices, such as slang, local idioms, and regional expressions, which helps prevent imposing dominant ways of using language, another challenge in language model development.

To create datasets, NLP groups typically use one or both of two approaches: translating existing datasets from a high-resource language, or generating entirely new data. When using the former approach, NLP groups often translate and supplement an English dataset by integrating culturally relevant topics, or translate from a closely related non-English language that shares similarities in vocabulary and cultural context.

For instance, when the IndoNLP group created NusaX, a human-translated benchmark dataset for 10 low-resource Indonesian languages, they hired bilingual speakers to translate an existing Indonesian sentiment analysis dataset to the 10 local languages (Indonesian itself is a high-resource language in Indonesia, but other local languages like Ngaju and Sundanese are considered low-resource). By using Indonesian as a base for expansion, IndoNLP ensures that topics and entities in the data –– such as people, organizations, and locations –– are culturally relevant to the local languages, a characteristic that is often lacking when translating English datasets. Additionally, some local languages share lexical similarities with Indonesian in vocabulary and grammatical structure, leading to better translation performance compared to translating from English.

Even when translating English datasets, extra steps can be taken to improve the cultural relevance of the data and resulting models, like replacing Western locations, brands, and cuisines with local entities and using colloquial translation to capture the linguistic patterns of everyday speakers, such as slang and code-mixing, where speakers switch between languages in a conversation.

When generating entirely new data, NLP groups often incorporate culturally relevant topics into their conversation scenarios to elicit participant responses that reflect local culture. The AI4Bharat group illustrates this practice with its IndicVoices speech dataset for 22 Indian languages, part of which has already been transcribed into written form. AI4Bharat tailored the dataset to local geographies at multiple levels of granularity by designing general, state-specific, and district-specific roleplay scenarios to enrich conversations between two participants with cultural context. For example, state-specific roleplays involved interactions such as "Kashmiri artisan and local Kashmiri discussing the impact of industries on handcrafted items," while district-specific roleplays explored more localized contexts like "Rice dealer and customer discussing types of rice native to Palakkad." AI4Bharat also contextualized the dataset to rural and urban geographies by conducting a pilot in a rural district where scenarios for data creation were iteratively revised. For example, the scenario of "hailing a cab" was modified to "arranging for transport for cattle or food grains" to better reflect the experiences of rural participants.

Several NLP groups prioritize including diverse community members in data collection and translation efforts to ensure that datasets reflect the heterogeneity of their communities. For example, to create IndicVoices, AI4Bharat included speakers of various ages, genders, professions, and educational backgrounds. Instead of using random sampling, AI4Bharat used sampling by group, where they worked to ensure the dataset included sufficient representation from each sub-group and was not dominated by majority groups. Some data was also collected on an 8 KHz telephone channel to ensure the representation of low-income users in India who may not have smartphones. AI4Bharat recruited these diverse community members through local partnerships with universities, data collection agencies, foundations dedicated to regional language preservation, and social sector professionals with connections at the grassroots level. AI4Bharat's efforts to diversify community participation are particularly impactful, as its tools are used by the National Programme on Technology Enhanced Learning to subtitle higher education videos, and by the Supreme Courts of India and Bangladesh to translate judicial documents.

Finally, several NLP groups follow ethical practices that recognize participation as a form of labor. For example, when developing IndicVoices, AI4Bharat ensured informed consent and allowed participants to opt out of any tasks they found uncomfortable. Participants were compensated according to the daily wages in their districts, and AI4Bharat prioritized their well-being, providing refreshments to create a welcoming environment.

To be sure, while the approaches described above are strong practices for localizing datasets with community input, they can be resource-intensive. One less costly alternative for dataset creation is integrating semi-automatic methods, where large language models are used to generate initial datasets that are then refined and annotated manually by experts and native speakers. Semi-automatic methods can be useful when funds or access to such individuals are limited, and are better than fully automatic methods that are unlikely to incorporate important contextual knowledge. Another alternative some consider is crowdsourcing, where data is collected from a large group of dispersed people, exemplified by IndoNLP's NusaCrowd. However, crowdsourcing in this context can be challenging, as platforms like

Amazon Mechanical Turk typically have few low-resource language speakers. Moreover, crowdsourcing often relies on unpaid or low-wage labor, may overlook minority voices or those most impacted by technology, and typically follows a top-down design process where communities are consulted but not involved in long-term partnerships.

Promising practices for companies and research groups:

- Involve native speakers and language experts in creating datasets that are culturally relevant and geographically specific. Recruit diverse participants through local networks in a decentralized manner, and involve them throughout the lifecycle of dataset creation, from content creation to annotations to quality checks.
- Adopt an iterative feedback process with participants to design and revise questions, instructions, and scenarios for new data acquisition.
- Most importantly, recognize participation in dataset creation as labor –– data workers do ghost work that is often invisible and downplayed, yet essential for maintaining and improving NLP systems. Therefore, it is crucial to ensure ongoing consent throughout the data collection process as well as fair compensation for participants.

2. Create ways for communities to benefit from NLP tools built in their languages

While some efforts to develop multilingual AI can lead to tokenism and exploitation of annotators, translators, moderators, and native speakers, many non-English NLP groups have developed mechanisms for communities to experience the benefits of the NLP tools built in their languages. These efforts are crucial for advancing community participation as justice, where relationships with participants are long-term and mutually beneficial, with involvement in design decisions and throughout the lifecycle of AI development.

First, to foster empowering rather than exploitative participation, several NLP groups establish shared tasks, where community members collaborate to build datasets or use a common dataset to build models that tackle a particular problem. For example, participants might build a model to detect propaganda in Arabic to combat its spread in media. Participants often publish a paper on their work and engage with other contributors at conference workshops. Shared tasks foster open participation and collaboration, in contrast to approaches where companies design and develop private datasets or build models internally without community input. These tasks also increase the visibility of non-English datasets to the NLP research community and incentivize the creation of models to address specific problems. Moreover, NLP groups usually launch and revisit shared tasks annually, enabling opportunities for longer-term engagement rather than one-time consultations for NLP development.

Ideally, communities are involved in designing a shared task to address a specific problem they face, rather than just participating in a predetermined task. For example, speakers of oral German dialects indicated a desire for NLP tools, such as virtual assistants, that respond to audio input rather than text, which could prompt a shared task to build these speech-based tools. AmericasNLP's new shared task, which emerged from researchers working with Mayan communities in Mexico, focuses on automatically creating materials for teaching native

American languages, addressing the critical need for educational resources to support the learning of endangered indigenous languages. The SIGARAB group also organizes Arabic-specific shared tasks oriented towards community needs like financial NLP to manage banking data in Arabic-speaking regions and annotation of news articles about Gaza to uncover media bias. The datasets and models generated from shared tasks can be especially impactful when used in real-world tools –– for example, the HausaNLP group collaborated with the start-up Cohere, which used their AfriSenti shared task dataset for sentiment analysis to build the multilingual Aya model, covering 51 low-resource languages including Hausa.

Such NLP groups often foster accessible research to help people without formal training learn about NLP and contribute to projects, by offering tutorials, creating beginner-friendly tools, and addressing barriers to accessing conferences. For example, the ARBML group ensures that its Arabic NLP tools are easily accessible via different interfaces, allowing people to test models for Arabic-English translation directly in their browser. Additionally, ARBML shares its source code on GitHub and provides tutorials that enable novice researchers to replicate experiments and gain a clear understanding of how to do each task. HausaNLP is similarly committed to making AI and NLP education more accessible. One of their founders established the Arewa Data Science Academy, which offers free data science and machine learning training to underserved students in Nigeria, many of whom join HausaNLP afterward. Making NLP research accessible also includes addressing barriers that prevent attendance at academic conferences like visa issues and prohibitive costs. For example, the NorthAfricanNLP group offers conference fee waivers to help North African students and early-career researchers attend academic conferences and have greater visibility among the research community.

Lastly, some NLP groups foster flexibility in contributions throughout the lifecycle of NLP development. Groups that enable deeper levels of participation encourage participants who generate data to have input into other parts of the lifecycle, such as the problem definition and model development phases. Masakhane, a grassroots NLP community for African languages, exemplifies this practice by not imposing fixed roles on participants. For example, someone who initially joins as a translator to help with dataset creation might later become a junior language technologist and guide the creation of NLP models when equipped with tools, tutorials, and mentoring. Maskahane enables this fluidity by sharing agendas and meeting notes openly, allowing for participants to be involved throughout the lifecycle and to democratically vote on agendas. Such flexibility in contributions allows for people in different roles to learn from each other –– model developers can gain insights about the data from translators, while translators can learn NLP skills from developers. Masakhane's community-driven approach is especially impactful through its partnership with Lelapa AI, a leading company in multilingual AI on the African continent, which provides transcription and content analysis products in Afrikaans, isiZulu and Sesotho for people and businesses.

Promising practices for companies and research groups:

- Establish shared tasks that reflect community needs and interests.
- Develop educational materials to make NLP research more accessible to non-experts.
- Create opportunities for community involvement throughout the lifecycle of NLP development — not just in data collection, but also in the design and development of applications, to ensure that the tools serve the people whose data is being collected.

3. Create channels for ownership and authorship by NLP contributors

In contrast to Western transactional approaches to data sharing and restrictive authorship models that only reward certain kinds of participation like data analysis and manuscript writing, several non-English NLP groups prioritize data ownership, use inclusive authorship models, and provide clear guidelines for contributions. These efforts enable more ethical and substantive participation, shifting power to communities in exercising their agency and shaping NLP development.

This extends to data refusal, where communities say "no" to how their data is collected or used and challenge the authority of data collectors. For instance, the Māori community has advocated for indigenous data sovereignty to protect their data from exploitation. By resisting Western open-source practices, the Māori aim to prevent highly resourced tech companies from co-opting and monopolizing their datasets in ways that do not primarily benefit the Māori people.

Masakhane exemplifies an inclusive approach to participation with its non-traditional authorship model that recognizes not only contributions to the results, analysis, and writing, but also those in the form of data, evaluation, lived experiences, research strategies, and coordination of participant activities. Additionally, in Masakhane's effort to collect benchmarks for African languages, contributors submit benchmarks as GitHub Pull Requests, ensuring they can be contacted and their ownership is maintained.

IndoNLP also supports data ownership and inclusive authorship in its Indonesian NLP crowdsourcing effort, NusaCrowd. IndoNLP did not copy or store crowdsourced datasets, but rather maintained the control and ownership with the original contributors. Dataset contributors could become co-authors on the NusaCrowd publication based on a transparent scoring guideline that evaluated the value of their contributions. To encourage diverse NLP datasets, this guideline gave higher scores to datasets that were publicly available, manually validated, and developed for rare NLP tasks and local languages. IndoNLP was also open to other forms of contributions, subject to open discussion in its Slack and WhatsApp community channels.

Promising practices for companies and research groups:

- Co-design guidelines with communities on how they can participate in diverse ways and be recognized, including through authorship.
- Ensure that inclusive authorship does not replace fair compensation for contributions, such as data creation, that constitute labor.
- Preserve ownership of dataset or model contributions, and respect community decisions to decline collaboration or resist inclusion in datasets.

Reconciling Participation with Scale and Speed

The localized approach to community participation that these NLP collectives embrace may seem incompatible with the globalized operation of AI systems—but it does not need to be.

In the tech industry today, scaling tends to centralize power, whereas participation aims to distribute power more equitably. When implemented at scale, participation often becomes a means-to-an-end for improving AI systems for profit, rather than being valued as an end itself for empowering communities. But despite the tension between scale and participation, they are not mutually exclusive.

First, participation can support scaling efforts. Several NLP groups have developed tools across many languages and geographies. For example, Masakhane collected machine translation benchmarks for 32 African languages, and AI4Bharat built a speech dataset from 16,237 speakers covering 145 districts and 22 Indian languages. One key difference from traditional scaling is that these groups often *scale from below* rather than from above. Their bottom-up approach enables scaling, when desired, to be grounded in local community interests by operating through distributed or grassroots networks.

Of course, as participation scales, it requires more time and resources, especially to build partnerships with the growing number of communities involved. This can create a tension between the depth of participation and the appetite for rapid AI development, especially when working to address timely issues that may require immediate action. For example, when addressing online hate speech related to ongoing violence, should a social media platform prioritize deeply involving communities to create language- and context-specific moderation tools, even if it delays deploying less effective moderation tools that may help reduce violent events? Or should the platform swiftly develop and deploy tools that may be partially effective but involve less comprehensive participation, and then gather and incorporate feedback afterward?

While speed and participation are often at odds, social infrastructures that support community engagement at scale can help address this tension. For example, community advisory boards can facilitate the designation of community representatives. Regular community forums can enable communities to collectively identify and act on issues posed by an AI system. Networks of grassroots organizations can mobilize community members

to contest or provide feedback on an AI system. These infrastructures would help enable community decision-making and timely changes to AI systems and related policies, addressing issues like moderation over-enforcement, such as the ban of the Arabic word "shaheed," and under-enforcement, such as the failure to detect violent threats in Amharic.

By supporting locally rooted NLP efforts and establishing infrastructures that allow for meaningful community participation, practitioners can create AI systems that are better attuned to diverse cultures and languages. Participation can help redistribute power, enabling communities to benefit from the technologies built in their own languages, rather than concentrating power with those who design and deploy them.

Find more from CDT's AI Governance Lab at *cdt.org*

Acknowledgements

Thank you to Luis Chiruzzo, Nedjma Ousidhoum, and Shamsuddeen Hassan Muhammad for insightful conversations about non-English NLP groups, and to Gabriel Nicholas, Mona Elswah, and Stephen Yang for valuable feedback on this brief.

*The **Center for Democracy & Technology** (CDT) is the leading nonpartisan, nonprofit organization fighting to advance civil rights and civil liberties in the digital age. We shape technology policy, governance, and design with a focus on equity and democratic values. Established in 1994, CDT has been a trusted advocate for digital rights since the earliest days of the internet. The organization is headquartered in Washington, D.C. and has a Europe Office in Brussels, Belgium.*