Language Technologies as if People Mattered: Centering Communities in Language Technology Development

Nina Markl, Lauren Hall-Lew, Catherine Lai
University of Essex, University of Edinburgh
Colchester, Edinburgh
nina.markl@essex.ac.uk, {lhlew, c.lai}@ed.ac.uk

Abstract

In this position paper we argue that researchers interested in language and/or language technologies should attend to challenges of linguistic and algorithmic injustice together with language communities. We put forward that this can be done by drawing together diverse scholarly and experiential insights, building strong interdisciplinary teams, and paying close attention to the wider social, cultural and historical contexts of both language communities and the technologies we aim to develop.

Keywords: algorithmic bias, sociolinguistics, ethics

1. Introduction

In the last decade, speech and language technologies have seen unprecedented "successes" across the board. Performance of a wide range of applications has apparently increased steadily, as measured in established benchmarks. Many tools have found widespread adoption through integration in consumer and business computing, and speech and language technologies have become a focal point in the interest (and hype) surrounding "artificial intelligence".

As a result, technologies that researchers have known in some form for a long time, like automatic speech recognition (ASR), speech synthesis (TTS) and (large) language models (LLMs) are being deployed (and developed) in novel social contexts. These changes in context, rather than (just) the technologies themselves, raise a number of ethical, technical and legal questions such as:

- How should we develop language technologies that work for everyone?
- Who should be developing language
- What risks and impacts should we accept in the development of language technologies?

2. Continuing the conversation

These questions are not (only) technical but social and normative: they are about what we *should* do rather than (just) what we *can* do. They are, of course, not *new* questions. However, unlike many technical questions, they cannot be definitively resolved but need to be engaged with on a continuous basis. They are particularly important at this juncture. The most prominent and widely available language technologies at this moment are highly resource-intensive models (in terms of data, hardware, energy, labour) controlled by (large) commercial developers (Tacheva and Ramasubramanian, 2023; Whittaker, 2021). Despite growing access to speech technologies for more and more languages (e.g., Zhang et al., 2023), we do not see growing participation, agency, or ownership by language communities in the development and deployment process (Mahelona et al., 2023; Schwartz, 2022).

This paper is intended as an introduction to language, identity, and injustice in the context of language technologies, and an invitation to all members of this research community to recognise their own responsibility in grappling with complex ethical, technical and legal questions and deferring to language communities. It has been heartening to see these big issues move towards the center of language technology research in recent years, including a long-overdue discussion of coloniality in language technology development (Held et al., 2023; Mahelona et al., 2023; Schwartz, 2022; Bird, 2020), algorithmic bias and harm (Wenzel et al., 2023; Bender et al., 2021; Koenecke et al., 2020), and the double-edged sword of "diversity and inclusion" (Helm et al., 2024; Hoffmann, 2021). Here, we want to add to this conversation and propose some practical steps to sustain a (renewed) focus on interdisciplinarity and language communities in language technology development.

3. Language, identity, and injustice

Using language (regardless of modality) is a fundamentally social process. How we use language depends on many different layers of context: the people we are engaging with and our relationship to them, the physical and social setting of the interaction, our linguistic backgrounds, and our embodiment, among others. In this way, language use

LREC-COLING 2024, pages 10085–10099
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

is closely tied to social identity. Beyond the individual, language communities (and by extension, their languages) are also embedded in a web of power relations. Below we discuss some ways of conceptualising the role of these contexts, in particular as related to identity and justice, before bringing them into conversation with technology.

3.1. Language and identity

Since at least the 1960s, research in the field of "variationist" sociolinguistics has been documenting that language variation is socially stratified along axes like class, race, and ethnicity (Tagliamonte, 2011)[1]. While the specific linguistic variables of interest differ and much of the foundational work was conducted on variation in English in the United States (following Labov, 1966), broad social patterns have been found to hold across a wide range of linguistic and social contexts. For example, that speakers with a lower socioeconomic status use stigmatised variants more frequently than speakers with higher socioeconomic status, and that there are relatively fewer markers of regional and ethnic variability among higher-status speakers, and that this lack of variation itself marks high status (e.g., Arabic (Hassemer and Garrido, 2020); Chinese (Dong and Blommaert, 2009); English (Romaine, 1980)).

The "local context" in which speech occurs has also been found to have a regular effect on patterns of linguistic variation. Ethnographic research has uncovered the importance of "local" categories and "local" meaning which can account for differences within social groups. For example, while stigmatised variants are typically more frequent in the speech of men than women (Labov, 1990), Hazen (2008) found that Appalachian West Virginian women were more likely than men to drop their G's (i.e., to produce the *-ing* suffix with an alveolar nasal than a velar nasal). In contrast to every other study on English *-ing*, in Appalachian West Virginia the women are more "confident and unashamed" of their stigmatised regional dialect than the men are. In other words, binary gender has a different relationship to linguistic variation in this "local" context than in the context-free generalisations we might otherwise make. The same point was made by Haeri (1994), who found that a stigmatised variant of Cairene Arabic was used more often among women than men, in part because it was the men who had more access to formal education.

As such examples grew in number, the field of sociolinguistics shifted from recognising that linguistic variants are *correlated* with social categories to theorising that these categories are *constructed through* these linguistic variants (Bucholtz and Hall, 2005; Eckert, 2008, 2012). Socially meaningful linguistic variation is not the "incidental fallout" (Eckert, 2012) of a broader social structure, but rather one of the ways in which we build, maintain and challenge social structure(s). Importantly, the social meanings attached to any linguistic variable are not fixed. They only index social categories indirectly, and their meaning depends on speaker, speech situation, and hearer (Eckert, 2008). For example, the exact same vowel quality in the exact same speech community can index youthfulness, effeminacy, flamboyance, trendiness, regional identity, or all or none of these, depending on the time and context in which it is spoken and heard (Hall-Lew et al., 2021). As we use language we can draw on these meanings to construct social identity, and express stances by combining different linguistic variables into styles (e.g., Podesva, 2007; Zimman, 2017).

3.2. Language, power and justice

Some of the patterns of variation discussed above are noticed by speakers. Over time, correlations between particular social groups and (their) particular ways of using language become associated in speakers' minds (Irvine and Gal, 2000; Campbell-Kibler, 2010). This knowledge about language variation becomes very deeply embedded in our sense of how the world is and should be, what (and who) is "normal" or "different" (Craft et al., 2020; Rosa and Burdick, 2016; Irvine and Gal, 2000). The way this is achieved is, in part, through the way ideologies about language inform language *management* within a social context, that is how institutions and collectives decide which (kinds of) language(s) to use in particular social contexts.

Two such beliefs (or ideologies) which are particularly relevant to language technologies are that languages can and should be "standardised" (Lippi-Green, 2012; Milroy, 2001; Spolsky, 2003), and that languages are clearly delineated objects (Otheguy et al., 2015; Schneider, 2019). Like the standardisation of objects, measurements and tools (Bowker and Star, 2000), language standardisation is also not a neutral, but a political process (Milroy, 2001). Standardising languages involves selecting a variety (as there are always several different styles or varieties to choose from) and codifying (a written form of) this variety in dictionaries and grammars (Johnson, 2013). Crucially, the choices involved in this process are guided implicitly and explicitly by language ideologies as well as pre-existing power structures (Spolsky, 2003; Ricento, 2000; Shohamy, 2006). To further spread a standard and entrench its status, it is

[1] See Tagliamonte (2015) and Eckert (2012) for the history of variationist sociolinguistics, and Heller and McElhinny (2017) for a broader history of linguistics.

adopted in a variety of domains, including education and government. It is in part because of this official association with nation states and codification in dictionaries and grammars that we tend to perceive languages as clearly delineated objects even though they are marked by significant variation (Schneider, 2019; Otheguy et al., 2015; Irvine and Gal, 2000).

3.3. Algorithmic injustice

Given the connection between identity and language variation, worse language technology performance for a particular language variety or linguistic features often means worse performance for a particular group of people. This is especially problematic because language technologies tend to work better for the (high status) varieties of high-status speakers (e.g., Koenecke et al., 2020; Markl, 2022a). Empirically grounded understanding of different varieties can be used to audit language technologies, discover and mitigate performance differences, and build systems specifically for different varieties (e.g., Blodgett, 2021; Martin, 2022; Wassink et al., 2022; Choe et al., 2022).

However, there are limitations to this quantitative approach of measuring "bias". Birhane et al. (2022b) highlight that a focus on (quantitative measures of) unequal outcomes allows researchers to ignore users' lived experiences with algorithmic systems, and reproduces Western approaches to ethics and fairness. Birhane (2021) argues instead for a "relational ethics" approach to what she terms "algorithmic injustice". This approach, rather than privileging the hegemonic Western rationalism, draws on "relationality" as theorised and practised by different schools of thought (including Afrofeminism and complexity science) (Birhane, 2021). At the heart of this perspective lies a focus on "interdependence, relationships and connectedness", and a rejection of the rationalist quest for "timeless and absolute knowledge" predicated on a "rational, static, self-contained, and self-sufficient subject" (Birhane, 2021, 3). Instead, a relational ethics approach to algorithmic bias (and injustice), urges both breadth and depth of perspective. Away from abstracted metrics, it encourages us to consider the broader deployment and development contexts of a system, and the specific ways it interacts with people (Birhane, 2021). Feminist science and technology studies have long pointed out the fundamental impossibility of the kind of disembodied objectivity (implicitly assumed or explicitly asserted) in rationalist science (Haraway, 1988) and, more recently, machine learning (Talat et al., 2021).

Recent work has argued for the importance of incorporating the social meaning of linguistic variation in the design of language technologies which we want to promote justice (Sutton et al., 2019; Nguyen et al., 2021; Nee et al., 2021; Blodgett, 2021). Understanding the complex situated and relational nature of both people and their language varieties is crucial here. For example, as discussed above, linking language variation and macro-level social categories like race, gender, and class can help us audit language technologies for algorithmic bias. However, this very same linkage risks stereotyping (potential or actual) language (technology) users and glosses over a huge diversity in language use within social groups. Language is also, in a neutral sense of the term, ideological. The meanings we attach to linguistic variants and varieties are embedded within broader ideological frameworks and socio-cultural and historical contexts that we often take for granted both as researchers and everyday users of language. Which varieties and variants we develop for is always a political and ideological choice, even if we're not aware of it. While researchers may be constrained by wider social structures (e.g., funding incentive structures, data availability etc.), these constraints too are the result of pre-existing social and linguistic hierarchies (Markl, 2022b; Hanna and Park, 2020).

4. Social contexts of development and deployment

With the proliferation of language technologies in consumer and business computing, we have seen a rapid change in how they are being developed and deployed. These changes lead to important debates between and among developers and users regarding diversity and inclusion, bias and fairness, and sovereignty and responsibility.

4.1. New deployment contexts: Language technologies for all?

The biggest improvements in speech technologies, whether we measure them in terms of accuracy, efficiency, or affordability, have benefited only a small number of language communities. Languages like English, Spanish and Mandarin are often described as "high-resource" languages (Joshi et al., 2020; Bird, 2022). These "resources" are typically understood to be language datasets. However, the communities who speak these languages, and the nation states which are associated with these languages, are also rich in wealth and geopolitical power, in many cases as a direct result of violent colonial expansion (Heller and McElhinny, 2017). In part because language technology development is so costly, in terms of data, labour, and money, language communities which are smaller, minoritised, or "under-resourced" have historically been sidelined.

Of course, as discussed above, "English", "Spanish" and "Mandarin" are not monoliths. Each of these languages is comprised of a large number of varieties spoken in different regions and by different people. As a result, we see large performance disparities for different language communities even for "high-resource languages". It is the standard varieties of these languages which tend to be richest in resources, prestige, and, as a result, best-supported by language technologies (Markl, 2022a; Koenecke et al., 2020). In this way, linguistic hierarchies within a particular larger language community are reproduced in language technologies and speakers of marginalised varieties are less likely to enjoy any of their benefits and more likely to be negatively affected.

The boundaries between different languages are furthermore more porous than we often assume as the majority of people around the world use multiple different languages. Linguistic practices like code-switching (Heller, 1988) or translanguaging (Otheguy et al., 2015) whereby speakers effortlessly weave together words from what might be considered different languages (like Spanish and English) are extremely common but also very stigmatised in many "monoglot" societies such as the United States (Flores and Rosa, 2015; Silverstein, 1996).'[2] They are also poorly supported by language technologies (Doğruöz et al., 2021).

The dominance of a small number of "high-resource" language varieties (and their speakers) directly leads to the marginalisation of smaller language communities. In colonial contexts specifically, indigenous communities have often been violently suppressed, including in their use of their language(s) (Chiblow and Meighan, 2021; Charity Hudley et al., 2020; Kroskrity, 2021). Over time, discriminatory (legal or social) "rules" on how and where languages (and other cultural practices) should be used can lead to the loss of language varieties (and other cultural practices). Furthermore, communities often shift to languages which they perceive to be more (economically or socially) valuable.

Language technologies are often positioned as "saviours" in such contexts of language endangerment. For instance, in 2019 UNESCO organised (in partnership with ELRA) the "Language Technologies for All" conference where the demise of languages not supported by language technologies was framed as inevitable: "Languages that miss the opportunity to adopt Language Technologies will be less and less used, while languages that benefit from cross-lingual technologies such as Machine Translation will be more and more used" (ELRA, 2019, cited in Bird, 2020). While this impulse is often well-intentioned, it arguably reproduces a kind of "tech-solutionism" or "tech-chauvinism" (Broussard, 2019; Greene, 2021). Without a doubt, language technologies can be useful for minoritised and "under-resourced" communities in some contexts, they might also negatively impact communities and their languages. Whether and how technological interventions in precarious linguistic ecologies are ultimately successful depends on many factors (Bird, 2020). The impulse to apply the same standard to all languages, regardless of their historical, cultural and sociolinguistic context, and understand them all in the same way is an extension of the colonial approach to linguistic research and documentation (Deumert and Storch, 2018; Heller and McElhinny, 2017; Kuhn et al., 2020; Schwartz, 2022). Helm et al. (2024) use the term "language modelling bias" for "linguistic or cultural inaccuracies in the way a language is processed or represented" because the technology is (fundamentally) designed with a different social, cultural and linguistic context in mind. It is this intrinsic technical bias that is very difficult to resolve without reimagining the process from scratch.

Thinking carefully about the wider historical and political context of the language community, their language(s), and their needs and desires, is, in our opinion, an absolutely crucial first step. Ideally, this consideration should go far beyond "participatory design" (Sloane et al., 2022), and involve serious commitments to communities' sovereignty of their data and perhaps even the technologies themselves as discussed below.

4.2. Language technologies by whom?

While industry has always been a major driver of innovation, its influence across machine learning domains is perhaps now greater than ever (Birhane et al., 2022a; Rikap, 2022). This is in part because of the resources required to "beat" the current state of the art: ever larger datasets and computing resources (Whittaker, 2021). Data requirements have been dramatically changing the way datasets are compiled for about a decade now (e.g., Crawford and Paglen, 2021; Denton et al., 2021; Paullada et al., 2021).

Expanding language technologies to groups, be they understood as "language communities", "user groups", or "markets", also affects the data compilation process, in particular if these "new" com-

[2] Within the borders of the United States exist a very large number of languages and language communities, many of which predate the United States. Silverstein (1996) uses the term "monoglot" to describe societies which despite their obvious lived plurilingualism are characterised by a very strong commitment and to one monolingual standard variety (such as Standard English in the United States).

munities have previously not been considered in language technology development. As alluded to above, data compilation can form part of a larger project that Birhane (2020) terms “algorithmic colonisation”. Academic institutions and large technology corporations operating from the Global North seek in this way to extract (language) resources to develop tools, services, and research which, ultimately, benefit them at least as much as they benefit the communities they're supposedly serving, both in terms of financial and cultural capital. As Hoffmann (2021) highlights, discourses of “inclusive” and “ethical” development can be used by technology corporations (and academic institutions) to position themselves as responsible and “doing good” (see also Green, 2019).[3] But, as Fuller Medina argues: “language data is patrimony” (2022, 2). Fuller Medina is talking about one specific sociolinguistic corpus which contains “disappearing cultural heritage” (the “Older Recordings of Belizean varieties of Spanish”), but since linguistic corpora often feature folklore or personal recollections of a particular time and place, her point is relevant to many datasets of “naturalistic” language use. To honour this patrimony (and the language communities) she calls for “repatriation [of linguistic data]” (Fuller Medina, 2022, 19). This framing raises important questions regarding the “ownership” of not just data but language varieties more broadly, which are particularly acute in language technology development.

One interesting case study here is the response by Māori speakers to efforts to create proprietary or open-source Māori ASR systems (Coffey, 2021; Mahelona et al., 2023). Having compiled a transcribed speech dataset with Māori speakers, the Māori media company Te Hiku resisted requests to sell or license it to non-Māori developers (Coffey, 2021). Instead they trained their own system (building on open-source architectures) to transcribe their own radio archive for the Māori community (Coffey, 2021), a project they have since expanded into the Papa Reo project[4]. As one Te Hiko employee put it: “They suppressed our languages and physically beat it out of our grandparents. [...] And now they want to sell our language back to us as a service” (Coffey, 2021). Importantly, these questions of data sovereignty and who should *own* language data are not limited to explicitly for-profit contexts. The recently released open-source multilingual ASR model Whisper (Open AI) (Radford et al., 2022) was trained on over a thousand hours of Māori speech data. As Mahelona et al. (2023) (Papa Reo) note it is not clear where exactly this data was drawn from as Radford et al. (2022) provide no detailed description, but like Google USM (Zhang et al., 2023), Whisper is trained on data from the web. While it is therefore likely not drawing on the *same* datasets Te Hiko compiled and tried to safeguard, it does represent language technology development without (meaningful) engagement or consent of the Māori community.

4.3. Language technologies at what cost?

As pointed out by Crawford (2022) and Tacheva and Ramasubramanian (2023), machine learning is *extractive*, requiring ever-large amounts of resources: energy, data, minerals, labour. The significant harms caused by mining and manufacturing, and the huge carbon footprints associated with training and deploying deep learning based language technologies are slowly being recognised (Hershcovich et al., 2022; Schwartz et al., 2020). For example, Hershcovich et al. (2022) call for transparent and accurate reporting of carbon emissions and energy use associated with natural language processing experiments. They highlight that this kind of reporting is currently absent from much of the research, and argue that these should be reported alongside other impacts and ethical considerations (Hershcovich et al., 2022).

Language technology development also comes at significant human costs. Perhaps contrary to the public perception, the development of language technologies is not just conducted by (relatively) highly-paid engineers and researchers in universities and technology firms (located, predominately, in the Global North). Even in the age of unsupervised model training, most language technologies require human annotation at some point in their development cycle. Across machine learning domains, this annotation work is generally precarious and underpaid but ultimately crucial “cultural work” (Irani, 2013) outsourced to workers in the Global South (Gray and Suri, 2019). For example, detection of “toxic” (i.e., undesirable) text requires datasets with manually labelled examples. Like social media content moderation (Perrigio, 2022), annotating such “toxic” text can be extremely disturbing. In recent months, Kenyan employees of data annotation company Sama[5] which was contracted by Open AI, have alleged “exploitative conditions” (Rowe, 2023) and called for an investigation by the Kenyan government (Perrigio, 2023). They say that the task of reviewing graphic

[3]Furthermore, Sadowski (2019) argues, in modern capitalism, data is not *like* capital, but rather it *is* capital as it is essential to (especially (AI) technology) production.

[4]https://papareo.nz/

[5]Sama has since stated that they will no longer work on content moderation or natural language processing (Perrigio, 2023).

descriptions of violence without (what they deem) adequate preparation or support, has caused serious harm to their mental and physical health (Perrigio, 2023). The dataset these workers annotated was eventually used to limit the amount of "toxic" content generated by ChatGPT (Perrigio, 2023). The impacts on workers involved in the development language technologies should be a central ethical concern.

The deployment of language technologies, like other machine learning technologies, also affects the (economic) value assigned to some linguistic work, like translation (do Carmo, 2020). More broadly, as (Levy, 2022) argues in the context of long-haul truck drivers in the US, the expansion of "AI" in the workplace often translates to a deterioration of working conditions and much reduced worker autonomy due to increased ML-facilitated surveillance. Furthermore, the spread of proprietary language technologies to workplaces has as-yet poorly understood privacy and security implications, leading some workplaces to ban employees from using them (Naidu and Lange, 2023).

5. Attending to challenges together: slowly and carefully

Much of the research on language technologies is focused on *solving* carefully formulated *problems* through technical innovation. While this approach has proven extremely successful on a plethora of small and large challenges, and continues to lead to great innovations and improvements, it does not apply to all types of problems.

Some problems, we argue, we need to *attend to* even if we cannot "solve" them quickly or alone. Attending to a problem is about noticing and caring, paying deliberate attention. Big questions like "what are the impacts of language technologies on individuals and communities?", "what is language data and who can lay claim to it?", "how can we foster linguistic diversity?", cannot be answered definitively. Problems of "algorithmic bias", "data bias", "language endangerment", "linguistic discrimination", cannot be solved definitively – at least not without radical social change. But this indeterminacy need not be the end of the path. Instead, it can be a starting point. It is an invitation for persistent engagement with these issues and collaboration across and beyond disciplines. As Tsing argues: "Collaboration means working across difference, which leads to contamination. Without collaboration we all die." (2015, 28). Tsing is talking about collaboration between and within species (including plants and humans) in the face of ecological disturbance but it is not difficult to see how her point translates to social and technological change and the challenges they raise (Tsing, 2015, 160).

5.1. Attention, not (quick) solutions

The problem of "bias in computing", as initially discussed almost 30 years ago by Friedman and Nissenbaum (1996), is one such challenge. There are ways to mitigate biases in machine learning (e.g., Mehrabi et al., 2021). However in addition to potentially having significant technical limitations (Gonen and Goldberg, 2019), these approaches always fail to address the underlying causes of "bias" in the first place. As Hoffmann puts it: "[E]fforts to achieve fairness and combat algorithmic discrimination fail to address the very hierarchical logic that produces advantaged and disadvantaged subjects in the first place. Instead, these efforts have tended to admit, but place beyond the scope of analysis important structural and social concerns relevant to the realization of data justice" (2019, 901). Auditing algorithmic systems and documenting algorithmic bias can push developers to adjust system behaviours for instance by changing training data (Buolamwini and Gebru, 2018; Raji and Buolamwini, 2019). However, this auditing paradigm does assume that public pressure or governance mechanisms internal or external to the developing organisation can affect such change – an assumption that does not always hold (Metcalf et al., 2021). More deeply, the kinds of biases in the technical designs of algorithmic systems, such as the ones identified by Helm et al. (2024) often cannot be easily addressed. As a growing area of scholarship points out, algorithmic bias in speech and language technologies are not just a matter or data sparsity for some language varieties. More fundamentally, language technologies tend to presume a written standard and monolingual speakers. Designing useful language technologies for contexts in which there either is no standard written form or it is not seen as culturally appropriate (Deumert, 2010), or for (the global majority of) communities whose linguistic repertoires include multiple "different" languages (Otheguy et al., 2015), requires a complete rethinking of our design process (Bird and Yibarbuk, 2024; Markl et al., 2023).

A parallel can be drawn here between the work to tackle algorithmic injustice (or bias, or discrimination) and linguistic injustice (or bias, or discrimination). Much of the work in sociolinguistics has been explicitly or implicitly motivated by a desire to prevent linguistic discrimination (Charity Hudley, 2013; Charity Hudley et al., 2020). Researchers have been documenting language difference and highlighting that this difference should not be understood as deficit (Charity Hudley, 2013; Henner and Robinson, 2021; Craft et al., 2020). For example, much work employing quantitative and qualitative methodologies and different theoretical frameworks has shown how linguistic difference and racial difference is co-constructed – and how this difference is then framed as a deficit in comparison to a white (linguistic) norm (Rosa and Flores, 2017; Rosa, 2018; Figueroa, 2023). As Henner and Robinson (2021) discuss, these norms are furthermore ableist in the way they position some ways of using language as disordered. The oppression of (a) language is not (just) about language, but about culture, history and identity. This is why language revitalisation efforts are complex and differ depending on the social and historical context of the language community (Chiblow and Meighan, 2021; Yamada, 2007; Smith, 2021).

Compiling evidence on discrimination has limited use. It can contribute to efforts to slowly change attitudes and change or dismantle oppressive institutions, but it is not a "quick" solution, since the problem, usually, is not that we don't know about the discrimination. Instead, linguistic discrimination requires our persistent attention in scholarship and teaching and requires us to reflect on our own biases as well (Mallinson and Charity Hudley, 2018). The same is true for algorithmic discrimination. As we approach (at least) thirty years of "bias" discussions in computing, we have amassed a wealth of evidence that racism, misogyny and ableism are reproduced in algorithmic systems due to "biases" in data and technology design, and that they can perpetuate harms regardless of biases in implementation (e.g., in policing, surveillance, automation) and development (e.g., exploitation of workers and environmental resources). And thanks to this evidence, many practices have changed, such as the adoption of documentation frameworks (Gebru et al., 2021; Mitchell et al., 2019; Bender and Friedman, 2018), routine testing for algorithmic bias in language technology development, a growing awareness of the social implications of this bias (Blodgett et al., 2020; Schwartz, 2022), and a move towards multilingual models. Nevertheless, there is still a lot of work to do. In particular, many of the fundamental logics of natural language processing (and AI more broadly), remain unchanged, such as a focus on scale, speed, novelty, efficiency, and universality (Birhane et al., 2022a; Tacheva and Ramasubramanian, 2023; Rikap, 2022; Ricaurte, 2022; Bird, 2022). Many, if not most, language communities are not well-served by these logics. The first step in figuring out a better process, is putting communities (back?) at the centre of language technology design in a meaningful way.

5.2. Shifting the centre of attention

If the development is lead by language communities, they can, firstly, decide themselves whether and how their language(s) should be used in technology development. They can furthermore retain sovereignty over their (language) data and any derived technologies. This would follow the perspective of, for example, Mahelona et al. (2023) who argue, indigenous language technology development should be led by indigenous language communities in ways which ensure that they retain control over both the technologies and the datasets they are trained on. Similar arguments are made by organisers of participatory projects like Masakhane NLP (Nekoto et al., 2020) who describe themselves as a "grassroots NLP community for Africa, by Africa" and have been working on a range of language technology tasks in a number of "low-resource" African languages, such as named entity recognition (Adelani et al., 2022). They have furthermore compiled datasets for speech synthesis (Meyer et al., 2022), and developed a pre-trained language model (Dossou et al., 2022). This approach of involving (and crediting) large numbers of community members, is a way of shifting the centre of attention to what are often considered the "margins" of language technology development: "under-resourced" varieties and "under-resourced" communities.

The Distributed AI Research Institute (DAIR)[6] represents a complimentary movement towards community-centred, distributed AI research which pushes against the increasing consolidation of AI research. As Mahelona et al. (2023) highlight, the kind of models and datasets used by Google or Open AI are difficult to recreate, store and use without access to the right kind of (considerable) computing power, storage and expertise even if they are "open-source". These discussions of course tie into broader debates on ethics of data sharing, especially in the Global South. Abebe et al. (2021) identify the same kind of "deficit narratives" we see applied to "low-resource" language varieties, applied to African societies more broadly. Folding African researchers, research institutions and governments into a global culture of (more or less open) "data sharing", is framed as a necessary aspect of "development", but as Abebe et al. (2021) highlight, "equitable data sharing" is challenging. It requires a nuanced understanding of the "data setting" (i.e., the context) (Loukissas, 2019), local norms and interests and infrastructures which enable access for data subjects (Abebe et al., 2021).

5.3. Attending together

As a community of researchers interested in language(s) and language technologies, we should attend to deep-rooted linguistic and algorithmic injustice. Practically, this requires us to take an ethi-

[6] https://www.dair-institute.org/

cal, moral or political position, as noted by Blodgett et al. (2020). While there are arguably no neutral decisions in science or technology development, the need for foundational principles is particularly clear here. For example, we might take the position (and, we, as the authors do) that language communities *should* retain a level of sovereignty over their language(s). We also condemn linguistic and algorithmic injustice, which we consider forms of discrimination rooted in racism, ableism, classism and misogyny among others. Starting from these ethical, moral and/or political principles, we can focus on the contexts of language technology development and deployment.

As researchers and educators we have some ability to influence how language technologies are developed. Going beyond noticing injustices requires, we believe, interdisciplinary perspectives. It also requires us to take ourselves and our work outside of the traditional research centres to learn from language communities themselves. Attending together should involve interdisciplinary teams of researchers looking at different aspects of language and language technologies from different vantage points, including linguists, computer scientists, philosophers, interaction designers, law scholars, and sociologists, as well as relevant experts in the deployment domain of the technology (e.g., teachers and pedagogues for tools used in education). Members of the language community should also be considered experts in their own languages. They understand the histories of their languages and community and are best-placed to build towards their futures. These kinds of collaborative processes are inevitably complicated and slow, and likely involve disagreement and discomfort. They are also an ideal – there are many barriers to building and maintaining collaborative projects across and beyond institutions. But even if ideals cannot always be realised, they can be useful starting points guiding us towards what we might want to aim to do. Similarly changing teaching practices or broadening curricula in education is slow and laborious, but ultimately a powerful way to affect research cultures (Charity Hudley and Mallinson, 2018; Raji et al., 2021). For interdisciplinarity to be sustainable and rewarding, we also need to foster inclusive events and spaces where different kinds of skills, interests, backgrounds and knowledges are valued and recognised. Collaborators across and outwith academia and industry are affected by different external pressures (e.g., publication norms) and constraints (e.g., financial, geographic), and likely have different core interests. Translating between these differences and allowing for fertile cross-contamination is hard, but ultimately worthwhile, work.

Many of the positive and negative impacts of language technologies emerge only within specific deployment contexts. It is therefore important to consider how the technologies (and the research) we are working on are actually used and experienced by people. Once we have that established, we can, again departing from some ethical principles unique to us, think about their impacts. For example, automatic speech recognition tools can greatly improve the accessibility of digital technologies and information for a wide range of people (Reitmaier et al., 2023; Pradhan et al., 2018). However, when embedded in voice user interfaces in the home, they could also be collecting sensitive information without the informed consent of their users (Lau et al., 2018; Rincón et al., 2021). Where automatic speech recognition tools are biased, any benefits might be completely negated for some user groups and might even exacerbate existing linguistic discrimination (Wenzel et al., 2023; Mengesha et al., 2021).

Changing the deployment contexts perhaps means changing everything. And while that's forever "beyond the scope" of any one research project, curriculum and career, it is something we should consider in how we conduct our work.

6. Conclusions

In this paper, we invite you to draw your attention to the persistent ethical and social challenges raised by language technologies. Developing and deploying language technologies "as if people mattered" (Schumacher, 1993), involves grappling with linguistic and algorithmic bias, injustice, and discrimination, and engaging with language communities. Rather than positioning ourselves as experts who can "solve problems", this requires a reflexive and receptive approach. Acting as experts and problem solvers comes naturally to researchers – after all that is what we have been trained to do. But deeply-rooted social inequities cannot be "solved" over night, or alone. It is through collaboration across and beyond academic disciplines, that the "interdependence, relationships and connectedness" (Birhane, 2021) of languages, language communities and language technologies becomes apparent. Layering many different perspectives, and many different contexts on top of each other both complicates and clarifies the picture. It allows us to uncover the logics and histories of technologies, appreciate the cultural significance and societal role of language varieties and listen to and honour the desires and needs of language communities. Starting from our own ethical and political commitments, we can use this patchwork of insights and interests to build more equitable futures.

7. Acknowledgements

NM was supported by the UKRI Centre for Doctoral Training in Natural Language Processing (grant EP/S022481/1) and the University of Edinburgh, School of Informatics and School of Philosophy, Psychology & Language Sciences and the Institute for Analytics and Data Science at the University of Essex. NM would also like to thank her collaborators across different disciplines for generously engaging with her work and introducing her to new (to her) perspectives and scholarship.

8. Bibliographical References

Rediet Abebe, Kehinde Aruleba, Abeba Birhane, Sara Kingsley, George Obaido, Sekou L. Remy, and Swathi Sadagopan. 2021. Narratives and counternarratives on data sharing in africa. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM.

David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Mboning Tchiaze Elvis, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo Lerato Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Oluwaseun Adeyemi, Gilles Quentin Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu, and Dietrich Klakow. 2022. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

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. 2022. Local languages, third spaces, and other high-resource scenarios. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Steven Bird and Dean Yibarbuk. 2024. Centering the speech community. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 826–839, St. Julian's, Malta. Association for Computational Linguistics.

Abeba Birhane. 2020. Algorithmic colonization of Africa. *SCRIPTed*, 17(2):389–409.

Abeba Birhane. 2021. Algorithmic injustice: a relational ethics approach. *Patterns*, 2(2):100205.

Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022a. The values encoded in machine learning research. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM.

Abeba Birhane, Elayne Ruane, Thomas Laurent, Matthew S. Brown, Johnathan Flowers, Anthony Ventresque, and Christopher L. Dancy. 2022b. The forgotten margins of AI ethics. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM.

Su Lin Blodgett. 2021. *Sociolinguistically Driven Approaches for Just Natural Language Processing*. Ph.D. thesis.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5454–5476. Association for Computational Linguistics.

Geoffrey C. Bowker and Susan Leigh Star. 2000. *Sorting Things Out: Classification and Its Consequences*. The MIT Press.

Meredith Broussard. 2019. *Artificial Unintelligence: How Computers Misunderstand the World*. The MIT Press.

Mary Bucholtz and Kira Hall. 2005. Identity and interaction: A sociocultural linguistic approach. *Discourse Studies*, 7(4-5):585–614.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st conference on fairness, accountability and transparency*, volume 81 of *Proceedings of machine learning research*, pages 77–91. PMLR.

Kathryn Campbell-Kibler. 2010. The sociolinguistic variant as a carrier of social meaning. *Language Variation and Change*, 22(3):423–441.

Anne H. Charity Hudley. 2013. Sociolinguistics and social activism. In *The Oxford Handbook of Sociolinguistics*, pages 812–832. Oxford University Press.

Anne H. Charity Hudley and Christine Mallinson. 2018. Dismantling "the master's tools". *American Speech*, 93(3-4):513–537.

Anne H. Charity Hudley, Christine Mallinson, and Mary Bucholtz. 2020. Toward racial justice in linguistics: Interdisciplinary insights into theorizing race in the discipline and diversifying the profession. *Language*, 96(4):e200–e235.

Susan Chiblow and Paul J. Meighan. 2021. Language is land, land is language: The importance of indigenous languages. *Human Geography*, 15(2):206–210.

June Choe, Yiran Chen, May Pik Yu Chan, Aini Li, Xin Gao, and Nicole Holliday. 2022. Language-specific effects on automatic speech recognition errors for world englishes. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7177–7186. International Committee on Computational Linguistics.

Donavyn Coffey. 2021. Māori are trying to save their language from big tech. *Wired*.

Justin T. Craft, Kelly E. Wright, Rachel Elizabeth Weissler, and Robin M. Queen. 2020. Language and Discrimination: Generating Meaning, Perceiving Identities, and Discriminating Outcomes. *Annual Review of Linguistics*, 6(1):389–407.

Kate Crawford. 2022. *Atlas of AI Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press.

Kate Crawford and Trevor Paglen. 2021. Excavating AI: the politics of images in machine learning training sets. *AI SOCIETY*.

Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, and Hilary Nicole. 2021. On the genealogy of machine learning datasets: A critical history of imagenet. *Big Data & Society*, 8(2):20539517211035955.

Ana Deumert. 2010. Imbodela zamakhumsha – reflections on standardization and destandardization. *Multilingua - Journal of Cross-Cultural and Interlanguage Communication*, 29(3-4):243–264.

Ana Deumert and Anne Storch. 2018. Language as world heritage?: Critical perspectives on language-as-archive. In Natsuko Akagawa and Laurajane Smith, editors, *Safeguarding Intangible Heritage*. Routledge.

Félix do Carmo. 2020. 'time is money' and the value of translation. *Translation Spaces*, 9(1):35–57.

A. Seza Doğruöz, Sunayana Sitaram, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2021. A survey of code-switching: Linguistic and social perspectives for language technologies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1654–1666, Online. Association for Computational Linguistics.

Jie Dong and Jan Blommaert. 2009. Space, scale and accents: Constructing migrant identity in beijing. 28(1):1–23.

Bonaventure F. P. Dossou, Atnafu Lambebo Tonja, Oreen Yousuf, Salomey Osei, Abigail Oppong, Iyanuoluwa Shode, Oluwabusayo Olufunke Awoyomi, and Chris Emezue. 2022. AfroLM: A self-active learning-based multilingual pretrained language model for 23 African languages. In *Proceedings of The Third Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pages 52–64, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Penelope Eckert. 2008. Variation and the indexical field. *Journal of Sociolinguistics*, 124:453–476.

Penelope Eckert. 2012. Three Waves of Variation Study: The Emergence of Meaning in the Study of Sociolinguistic Variation. *Annual Review of Anthropology*, pages 87–100.

Megan Figueroa. 2023. Language development, linguistic input, and linguistic racism.

Nelson Flores and Jonathan Rosa. 2015. Undoing appropriateness: Raciolinguistic ideologies and language diversity in education. *Harvard Educational Review*, 85(2):149–171.

Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on Information Systems*, 14(3):330–347.

Nicté Fuller Medina. 2022. Data is patrimony: on developing a decolonial model for access and repatriation of sociolinguistic data.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM*, 64(12):86–92.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North*, page 609–614. Association for Computational Linguistics.

Mary L. Gray and Siddharth Suri. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Houghton Mifflin Harcourt.

Ben Green. 2019. "good" isn't good enough. In *AI for Social Good workshop at NeurIPS (2019), Vancouver, Canada*.

Daniel Greene. 2021. *The Promise of Access: Technology, Inequality, and the Political Economy of Hope*. The MIT Press.

Niloofar Haeri. 1994. A linguistic innovation of women in cairo. *Language Variation and Change*, 6(1):87–112.

Lauren Hall-Lew, Amanda Cardoso, and Emma Davies. 2021. *Social Meaning and Sound Change*, page 27–53. Cambridge University Press.

Alex Hanna and Tina M. Park. 2020. Against scale: Provocations and resistances to scale thinking.

Donna Haraway. 1988. Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist Studies*, 14(3):575–599.

Jonas Hassemer and Maria Rosa Garrido. 2020. Language as a resource with fluctuating values: Arabic speakers in humanitarian and social work. *International Journal of the Sociology of Language*, 2020(264):137–161.

Kirk Hazen. 2008. (ing): A vernacular baseline for english in appalachia. *American Speech*, 83(2):116–140.

William Held, Camille Harris, Michael Best, and Diyi Yang. 2023. A material lens on coloniality in nlp.

Monica Heller, editor. 1988. *Codeswitching*. De Gruyter Mouton.

Monica Heller and Bonnie S. McElhinny. 2017. *Language, Capitalism, Colonialism: Toward a Critical History*. University of Toronto Press.

Paula Helm, Gábor Bella, Gertraud Koch, and Fausto Giunchiglia. 2024. Diversity and language technology: how language modeling bias causes epistemic injustice. *Ethics and Information Technology*, 26(1).

Jon Henner and Octavian Robinson. 2021. Unsettling languages, unruly bodyminds: Imaging a crip linguistics. *Preprint*.

Daniel Hershcovich, Nicolas Webersinke, Mathias Kraus, Julia Bingler, and Markus Leippold. 2022. Towards climate awareness in NLP research. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2480–2494, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Anna Lauren Hoffmann. 2019. Where fairness fails: Data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society*, 22(7):900–915.

Anna Lauren Hoffmann. 2021. Terms of inclusion: Data, discourse, violence. *New Media & Society*, 23(12):3539–3556.

Lilly Irani. 2013. The cultural work of microwork. *New Media Society*, 17(5):720–739.

J. T. Irvine and S. Gal. 2000. Language ideology and linguistic differentiation. In P. V. Kroskrity, editor, *Regimes of language: Ideologies, polities, and identities*, pages 35–84. School of American Research Press.

David Cassels Johnson. 2013. *Language policy*. Palgrave Macmillan.

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. pages 6282–6293.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and

Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

Paul V. Kroskrity. 2021. Covert linguistic racisms and the (re-)production of white supremacy. *Journal of Linguistic Anthropology*, 31(2):180–193.

Roland Kuhn, Fineen Davis, Alain Désilets, Eric Joanis, Anna Kazantseva, Rebecca Knowles, Patrick Littell, Delaney Lothian, Aidan Pine, Caroline Running Wolf, Eddie Santos, Darlene Stewart, Gilles Boulianne, Vishwa Gupta, Brian Maracle Owennatékha, Akwiratékha' Martin, Christopher Cox, Marie-Odile Junker, Olivia Sammons, Delasie Torkornoo, Nathan Thanyehténhas Brinklow, Sara Child, Benoît Farley, David Huggins-Daines, Daisy Rosenblum, and Heather Souter. 2020. The indigenous languages technology project at NRC canada: An empowerment-oriented approach to developing language software. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics.

William Labov. 1966. *The social stratification of English in New York City*. Center for Applied Linguistics, Washington.

William Labov. 1990. The intersection of sex and social class in the course of linguistic change. *Language Variation and Change*, 2(2):205–254.

Josephine Lau, Benjamin Zimmerman, and Florian Schaub. 2018. Alexa, are you listening? *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–31.

Karen Levy. 2022. *Data Driven*. Princeton University Press.

Rosina Lippi-Green. 2012. *English with an accent language, ideology, and discrimination in the United States*, 2nd ed.. edition. Routledge.

Yanni Alexander Loukissas. 2019. *All Data Are Local*. The MIT Press.

Keoni Mahelona, Gianna Leoni, Suzanne Duncan, and Miles Thompson. 2023. Openai's whisper is another case study in colonisation.

Christine Mallinson and Anne H. Charity Hudley. 2018. Turning the lens onto our own language: Engaging in critical reflexivity in the pursuit of social change. *Language in Society*, 47(3):361–364.

Nina Markl. 2022a. Language variation and algorithmic bias: Understanding algorithmic bias in british english automatic speech recognition. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 521–534, New York, NY, USA. Association for Computing Machinery.

Nina Markl. 2022b. Mind the data gap(s): Investigating power in speech and language datasets. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 1–12. Association for Computational Linguistics.

Nina Markl, Electra Wallington, Ondrej Klejch, Thomas Reitmaier, Gavin Bailey, Jennifer Pearson, Matt Jones, Simon Robinson, and Peter Bell. 2023. Automatic transcription and (de)standardisation. In *2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL 2023)*. ISCA.

Joshua L. Martin. 2022. *Automatic Speech Recognition Systems, Spoken Corpora, and African American Language: An Examination of Linguistic Bias and Morphsyntactic Features*. Ph.D. thesis.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35.

Zion Mengesha, Courtney Heldreth, Michal Lahav, Juliana Sublewski, and Elyse Tuennerman. 2021. "I don't Think These Devices are Very Culturally Sensitive."—Impact of Automated Speech Recognition Errors on African Americans. *Frontiers in Artificial Intelligence*, 4:725911.

Jacob Metcalf, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeleine Clare Elish. 2021. Algorithmic impact assessments and accountability. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, page 735–746. ACM.

Josh Meyer, David Adelani, Edresson Casanova, Alp Öktem, Daniel Whitenack, Julian Weber, Salomon KABONGO KABENAMUALU, Elizabeth Salesky, Iroro Orife, Colin Leong, Perez Ogayo, Chris Chinenye Emezue, Jonathan Mukiibi, Salomey Osei, Apelete AGBOLO, Victor Akinode, Bernard Opoku, Olanrewaju Samuel, Jesujoba Alabi, and Shamsuddeen Hassan Muhammad. 2022. BibleTTS: a large, high-fidelity, multilingual, and uniquely African speech corpus. In *Proc. Interspeech 2022*, pages 2383–2387.

James Milroy. 2001. Language ideologies and the consequences of standardization. *Journal of Sociolinguistics*, 5(4):530–555.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA. Association for Computing Machinery.

Coulter Martin Naidu, Richa and Jason Lange. 2023. Chatgpt fever spreads to us workplace, sounding alarm for some. *Reuters*.

Julia Nee, Genevieve Macfarlane Smith, Alicia Sheares, and Ishita Rustagi. 2021. Advancing social justice through linguistic justice: Strategies for building equity fluent NLP technology. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21, pages 1–9, New York, NY, USA. Association for Computing Machinery.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160. Association for Computational Linguistics.

Dong Nguyen, Laura Rosseel, and Jack Grieve. 2021. On learning and representing social meaning in NLP: A sociolinguistic perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 603–612. Association for Computational Linguistics.

Ricardo Otheguy, Ofelia García, and Wallis Reid. 2015. Clarifying translanguaging and deconstructing named languages: A perspective from linguistics. *Applied Linguistics Review*, 6(3):281–307.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336.

Billy Perrigio. 2022. Inside facebook's african sweatshop. *Time*.

Billy Perrigio. 2023. Openai used kenyan workers on less than $2 per hour to make chatgpt less toxic. *Time*.

Robert J. Podesva. 2007. Phonation type as a stylistic variable: The use of falsetto in constructing a persona. *Journal of Sociolinguistics*, 11(4):478–504.

Alisha Pradhan, Kanika Mehta, and Leah Findlater. 2018. "accessibility came by accident": Use of voice-controlled intelligent personal assistants by people with disabilities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable auditing. In *AIES '19: Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society*, AIES '19, pages 429–435. Association for Computing Machinery. Number of pages: 7 Place: Honolulu, HI, USA.

Inioluwa Deborah Raji, Morgan Klaus Scheuerman, and Razvan Amironesei. 2021. You can't sit with us: Exclusionary pedagogy in ai ethics education. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21. ACM.

Thomas Reitmaier, Electra Wallington, Ondřej Klejch, Nina Markl, Lea-Marie Lam-Yee-Mui, Jennifer Pearson, Matt Jones, Peter Bell, and Simon Robinson. 2023. Situating Automatic Speech Recognition Development within Communities of Under-heard Language Speakers. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23. Association for Computing Machinery.

Paola Ricaurte. 2022. Ethics for the majority world: Ai and the question of violence at scale. *Media, Culture Society*, 44(4):726–745.

Thomas Ricento. 2000. Historical and theoretical perspectives in language policy and planning. *Journal of Sociolinguistics*, 4(2):196–213.

Cecilia Rikap. 2022. The expansionary strategies of intellectual monopolies: Google and the digitalization of healthcare. *Economy and Society*, 52(1):110–136.

Cami Rincón, Os Keyes, and Corinne Cath. 2021. Speaking from experience. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–27.

Suzanne Romaine. 1980. Stylistic variation and evaluative reactions to speech: Problems in the investigation of linguistic attitudes in scotland. *Language and Speech*, 23(3):213–232.

Jonathan Rosa. 2018. *Looking Like a Language, Sounding Like a Race*. Oxford University Press, Incorporated.

Jonathan Rosa and Christa Burdick. 2016. Language ideologies. In Ofelia García, Nelson Flores, and Massimiliano Spotti, editors, *Oxford Handbook of Language and Society*, page 103–124. Oxford University Press.

Jonathan Rosa and Nelson Flores. 2017. Unsettling race and language: Toward a raciolinguistic perspective. *Language in Society*, 46(5):621–647.

Niamh Rowe. 2023. 'it's destroyed me completely': Kenyan moderators decry toll of training of ai models. *The Guardian*.

Jathan Sadowski. 2019. When data is capital: Datafication, accumulation, and extraction. *Big Data & Society*, 6(1):205395171882054.

Britta Schneider. 2019. Methodological nationalism in linguistics. *Language Sciences*, 76:101169.

E. F. Schumacher. 1993. *Small is Beautiful*. Vintage.

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)*. Association for Computational Linguistics.

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2020. Green ai. *Commun. ACM*, 63(12):54–63.

Elana Shohamy. 2006. *Language policy: hidden agendas and new approaches*. Routledge.

Michael Silverstein. 1996. Monoglot "standard" in america: Standardization and metaphors of linguistic hegemony. In Donald Lawrence Brenneis and Ronald K. S. Macaulay, editors, *The Matrix of Language*, pages 284–306. Westview Press.

Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. 2022. Participation is not a design fix for machine learning. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '22, New York, NY, USA. Association for Computing Machinery.

Linda Tuhiwai Smith. 2021. *Decolonizing Methodologies Research and Indigenous Peoples*. Bloomsbury Academic Professional.

Bernard Spolsky. 2003. *Language policy*. Cambridge University Press.

Selina Jeanne Sutton, Paul Foulkes, David Kirk, and Shaun Lawson. 2019. Voice as a design material. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM.

Jasmina Tacheva and Srividya Ramasubramanian. 2023. Ai empire: Unraveling the interlocking systems of oppression in generative ai's global order. *Big Data Society*, 10(2).

Sali A. Tagliamonte. 2011. *Variationist Sociolinguistics Change, Observation, Interpretation*. Wiley Sons, Incorporated, John.

Sali A. Tagliamonte. 2015. *Making Waves*. John Wiley & Sons, Inc.

Zeerak Talat, Smarika Lulz, Joachim Bingel, and Isabelle Augenstein. 2021. Disembodied Machine Learning: On the Illusion of Objectivity in NLP.

Anna Lowenhaupt Tsing. 2015. *The Mushroom at the End of the World: On the Possibility of Life in Capitalist Ruins*. Princeton University Press.

Alicia Beckford Wassink, Cady Gansen, and Isabel Bartholomew. 2022. Uneven success: automatic speech recognition and ethnicity-related dialects. *Speech Communication*, 140:50–70.

Kimi Wenzel, Nitya Devireddy, Cam Davison, and Geoff Kaufman. 2023. Can voice assistants be microaggressors? cross-race psychological responses to failures of automatic speech recognition. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Meredith Whittaker. 2021. The steep cost of capture. *Interactions*, 28(6):50–55.

Racquel-María Yamada. 2007. Collaborative linguistic fieldwork: Practical application of the empowerment model. *Language Documentation & Conservation*, 1:2.

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, Zhong Meng, Ke Hu, Andrew Rosenberg, Rohit Prabhavalkar, Daniel S. Park, Parisa Haghani, Jason Riesa, Ginger Perng, Hagen Soltau, Trevor Strohman, Bhuvana Ramabhadran, Tara Sainath, Pedro Moreno, Chung-Cheng Chiu, Johan Schalkwyk, Françoise Beaufays, and Yonghui Wu. 2023. Google usm: Scaling automatic speech recognition beyond 100 languages.

Lal Zimman. 2017. Gender as stylistic bricolage: Transmasculine voices and the relationship between fundamental frequency and /s/. *Language in Society*, 46(3):339–370.