Advancing social justice through linguistic justice: Strategies for building equity fluent NLP technology

Julia Nee
University of California, Berkeley
Berkeley, California, USA
jnee@berkeley.edu

Genevieve Macfarlane Smith
University of California, Berkeley
Berkeley, California, USA
genevieve.smith@haas.berkeley.edu

Alicia Sheares
University of California, Berkeley
Berkeley, California, USA
amsheares@berkeley.edu

Ishita Rustagi
University of California, Berkeley
Berkeley, California, USA
ishita.rustagi@berkeley.edu

ABSTRACT

Language and social reality are mutually reinforcing; as a result, natural language processing (NLP) presents a unique opportunity to shift social reality at scale, advancing *social* justice by promoting *linguistic* justice. We provide an overview of how language and bias are intertwined and implications for building NLP tools that actively advance equity and inclusion. Then, we present a framework for centering inclusion and social justice in NLP design at four overlapping layers of linguistic structure. The goal is to provide a foundation for adopting equity-centered principles in the creation of NLP tools that don't simply *mitigate* social biases, but actively *advance* inclusion and social justice through language. This work aims to be practical and builds from a partnership between researchers at the Center for Equity, Gender, and Leadership at the UC Berkeley Haas School of Business and leaders and practitioners at a large Silicon Valley tech firm. This framework can foster equity-centered thinking to lead to greater "equity fluent" NLP tools that have the potential to advance justice more broadly.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; • **Human-centered computing** → **Accessibility**.

KEYWORDS

equity-centered design, linguistic justice, language ideologies, discrimination, bias

ACM Reference Format:
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), October 5–9, 2021, –, NY, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3465416.3483301

EAAMO '21, October 5–9, 2021, –, NY, USA

ACM ISBN 978-1-4503-8553-4/21/10.
https://doi.org/10.1145/3465416.3483301

1 INTRODUCTION

When we as humans use language, we are able to communicate vast amounts of social and contextual information alongside what is literally said [28, 32]. When, for example, you pick up the phone and hear someone say, "Hello?" you are likely already forming a mental picture of whoever is on the other end of the line - their age, race, gender, and other features - even without being able to see them [14]. Because language invites us to make assumptions about others' identities, it can serve as a tool for propagating harmful bias and discrimination by proxy. At the same time, because language and social reality are mutually reinforcing (meaning our language reflects the world around us and influences how we think and what we do), language can also serve as a mechanism for advancing social justice.

NLP tools present a unique opportunity to address the biases that arise in human language because they can be trained to actively counter harmful biases [58]. At the same time, because NLP tools are typically trained on human language, often with the goal of producing language that is as close to naturally occurring human language as possible, unless the biases embedded in human language are explicitly addressed, NLP tools risk reproducing biases at scale.

In this paper, we provide an overview of previous work (building on [18]) illustrating the links between language and power as well as implications for NLP tools. We then present a framework designed to guide NLP researchers and developers to identify areas to promote linguistic justice at four layers of linguistic structure. We follow Aguilar Gil's [10] claim that - given how language is required to access other aspects of social and political life - linguistic rights are a prerequisite for human rights. We define linguistic justice, then, as the realization of equitable access to social and political life through any mother language. This includes equitable access to NLP tools and the opportunities they provide. Linguistic justice, under our definition, also entails the equal valuation of all language varieties.[1] Within our framework, we outline opportunities to center equity and inclusion in NLP research and development and to advance social justice through linguistic justice [10].

[1] Instead of referring to languages, dialects, and accents separately, we refer to "language varieties." All languages, dialects, and accents are equally capable of expressing complex concepts, and none is inherently better than another [39, 42]. We use "language variety" to highlight the equality of all human linguistic expression - both spoken and signed.

2 BACKGROUND

Language and reality are mutually reinforcing. As humans, we create and use language that reflects our own realities, experiences, and worldviews, but language can also subtly influence how we feel, think, and act [52]. If we use language with discriminatory categories or descriptions, we may be encouraged to think in discriminatory terms [61, 62]. Because language has the power to subtly shape the way that we conceptualize reality, it can be used as a tool to promote equity or to perpetuate bias. In NLP, whether an algorithm outputs "illegal alien" or "undocumented worker," for example, could impact how users of the NLP tool feel about the real individuals those terms describe. Ensuring that words and phrases that promote equity are output by NLP tools can have a positive impact on social justice.

Power and language are intertwined. People in power inform what words, phrases, and patterns of language use are legitimized and included in dominant narratives. For example, whether the term "illegal alien" or "undocumented worker" is used by political leaders and written in government documents influences which term may be seen as the "standard" and used in other contexts [64, 76]. If NLP tools are trained on language data from sources that over-represent certain voices (like government documents or online platforms), they may replicate dominant narratives [17]. The notion that algorithms, and thus NLP tools, are unbiased is inaccurate and indicative of "automation bias," or the over-reliance and over-acceptance of suggestions from automated systems [37]. When NLP systems output language that contains harmful bias, humans may internalize those biases and assume they are well-founded [17].

At a higher level, the language varieties that are used by those in power can also come to have more power or prestige than other language varieties. This results from *prestige transfer*, the process through which the prestige associated with people in power is transferred from the individuals themselves to things associated with them, such as their race, gender, or language [46]. Language is particularly susceptible to prestige transfer, as (unlike race or country of origin) it can be learned by those who wish to gain prestige.[2] Because of systemic racism in the US, "Standard" American English ("S"AE[3]) - but not equally expressive varieties like African American English (AAE[4]) - has been assigned value through prestige transfer. As Baker-Bell writes, the way that Black language is devalued, "reflects how Black lives are devalued in the world" [13, p. 2]. However, systemically *valuing* different language varieties can also result in valuing the users of those varieties.

Language can serve as a gatekeeper. Those in power often deem some ways of speaking as "appropriate" and others as "inappropriate" [34]. Individuals who have greater access to "appropriate" ways of communicating are given more advantages and opportunities. Defining what language variety is "appropriate" for a certain situation is not neutral [40]. For example, "S"AE is often used as a baseline in the professional world. This privileges White, middle-class "S"AE speakers, and disadvantages speakers of other varieties like AAE who must invest additional time and resources to acquire the prestige language variety [44]. English is not the only example: German, Mandarin and other languages have "standard" varieties that afford similar privileges to their speakers. Language standardization feeds into a narrative of progress that is covertly racist [40], as language standardization results in the marginalization and erasure of people who speak marginalized language varieties. When "standard" varieties are prioritized in NLP, further gatekeeping can result, as "standard" language users are given first access to and/or better service from new NLP tools. This further entrenches the connection between the varieties used in NLP development and the prestige they are assigned.

While race and other identities are protected categories, and discrimination against members of protected categories is illegal in the US, linguistic profiling - making judgments about individuals based on their speech - has been ruled legal in many cases, such as in hiring [14, 28, 44]. Because we often infer social categories such as race and gender from language, linguistic profiling can result in racial, gender, and other forms of discrimination by proxy. Linguistic profiling can put intense pressure on marginalized language speakers to change their ways of speaking and, in the workplace, can lead to stress and decreased morale [44]. If an NLP system requires or performs better when a user interacts with it using "S"AE, this places an inequitable burden on users of other language varieties.

Even if everyone speaks the same variety of "S"AE, White people - particularly men - enjoy extra privilege [11, 31, 34, 53, 54, 65]. For example, when students listened to identical audio lecture recordings, their evaluation of how understandable the speaker was varied depending on whether the same recording was paired with an image of an Asian speaker (less understandable) or a White speaker (more understandable) [65]. In another study, the same written text was judged differently whether the reader thought the writer was White or Black and a man or a woman [54].

These examples show that while all speakers of the same language variety may be able to *produce* the same linguistic outputs, audiences may *react* in different ways to the same linguistic stimuli depending on the social identities of the speaker. If we train NLP tools used for decision making (such as resumé screening or conversational analysis tools used in hiring) on previous decisions made by humans, the systems are susceptible to learn to discriminate as humans do.

How we talk about language varieties matters. As language and social reality are mutually reinforcing, how we talk about language impacts how that language is treated in the world. Consider, for example, the common practice of describing NLP tools as being developed for "low-resource," "zero-resource," or "resource-poor languages" (i.e. languages for which there is a relatively small set of language data available to NLP developers). While there may be

[2] As will be discussed in layer 4, marginalized speakers who use the "standard" or "prestige" forms may nevertheless be *perceived* as using nonstandard forms, making it impractical to seek linguistic justice through homogenous adoption of a singular language variety (i.e. monolingual language policies).

[3] We use *"Standard" American English* ("S"AE) to describe the language variety widely used in media, politics, and education in the United States. It is based largely on the English used by middle-class White men [53]. We use quotes around "Standard" to highlight the socially constructed nature of this variety's position and prestige. Other authors refer to this variety as *White Mainstream English* [13].

[4] We use *African American English* to describe the language varieties used by many Black Americans, though others may use terms including *African American Language, Black English, African American Vernacular English,* or *Ebonics* to describe these varieties [50].

fewer annotated data samples for some languages, framing those entire languages as "low-" or "zero-resource" minimizes the importance of the resources that *are* available in that language, whether spoken, written, or audio recorded. This reproduces a dominant discourse about what types of resources are valuable. Instead, we could identify these languages more precisely (e.g., "languages with growing annotated datasets"). Not only is such a label more accurate, it allows for a more equitable positioning between languages by creating space for a variety of language resources to be valued.

3 PROMOTING POSITIVE FRAMING AT FOUR LAYERS OF LINGUISTIC STRUCTURE

Having seen the links between language and power, and implications for NLP tools, the question remains: how might we advance diversity and inclusion - and ultimately social justice - through NLP? In this section, we present a framework outlining opportunities for NLP tools to advance social and linguistic justice across four layers of linguistic structure.

3.1 Layer 1: Words and phrases

Words and phrases convey more than their simple truth-conditional meanings; they also bring with them a network of associations accrued over time and space [17, 22]. Some of these associations are widely recognized as harmful (as with slurs), but others may be more opaque. In this layer, we present several opportunities for NLP tools to advance equity and inclusion in word choice.

NLP tools are increasingly able to flag language that may reinforce limiting societal expectations or advance stereotypes, and they can offer alternatives. For example, the NLP-powered tool Acrolinx helps companies create more inclusive technical documentation by flagging non-inclusive terms and suggesting alternatives [15]. Allybot, another NLP tool, integrates with Slack to monitor conversations and send inclusive recommendations for better word choices, such as flagging gendered language like "guys" and suggesting alternatives like "folks" or "team". NLP systems could be developed to flag definite phrases containing nouns derived from adjectives for identity categories (i.e. phrases with "the" as in "the Blacks"). Such definite phrases can become associated with stereotypes about the group they describe. NLP systems could offer replacement suggesions or, in the case of tools such as chatbots, provide outputs that are more precise by using the identity category as an adjective modifying a noun (e.g., "Black Americans").

Secondly, it is important to think about words and phrases in code itself. Ensuring that the language used in coding is not harmful is crucial to enhancing belonging among diverse groups of developers. For example, the terms "master" and "slave" used in coding may cause harm through their association with chattel slavery [66]. Certain terms may perpetuate harmful stereotypes and associated ideologies [42, 43]. We can replace such harmful language in code with neutral or positive language that conveys the same message (e.g., "primary / secondary"). A helpful resource is our Terminology Guide [9] which includes a list of terms with racist, colonialist, and otherwise problematic associations and presents a list of neutral and positive alternatives. Because language is constantly changing, and different words may evoke different associations for different people, identifying and remedying problematic associations should be an ongoing process.

Third, words and phrases in datasets - including data labels and the language data itself - are critical. Within data labelling, a good language practice in general is to be precise. For example, an imprecise demographic label like "non-White" privileges whiteness as the default and homogenizes the experiences of the people pushed into the category of "non-White". Similarly, if language data is labelled as "non-standard English," it privileges "S"AE as the default.

Within language data that NLP systems learn from, there can be harmful terms. It may be difficult or impossible to extricate all harmful terms from datasets, particularly given the importance of context in determining harm, but careful curation and documentation of datasets can help [17]. Using filters to flag and remove hate speech and toxic language from datasets can also help. However, these types of language can be hard to detect, as they are context dependent [2]. Words and phrases that are derogatory in one context may have been reclaimed and used for empowerment in other contexts. If filtering tools remove reclaimed terms, they risk filtering out marginalized perspectives. More work is needed to ensure context is incorporated in identifying hate speech. Also (as will be discussed in layer 4) determining what is or is not hate speech or toxic language is subjective, and judgments may be influenced by the developer's own positionality [18].

Relatedly, not all members of a given group may self-identify in the same way. For example, some self-identify as Latinx (a gender inclusive term) while others find that term to be "presumptuous" and self-identify as Latino/a or Hispanic [60]. Similarly, within the disability community, some people prefer "person-first" language (i.e. "person with a disability") while others prefer "identity-first" language (i.e. "disabled person"), and supporters of one framing may feel harmed by the other [23, 33, 48] For this reason, we emphasize that language is fluid and context-dependent. Our guiding questions help move towards dynamic and nuanced inclusive language practices within and through NLP.

Questions to utilize in developing NLP systems that advance equity and inclusion:

1.1 How might an NLP tool be built to help individuals or organizations advance their own use of words and phrases for equity and inclusion?
1.2 What terms are used within our code? How can harmful terms be replaced with neutral/positive alternatives?
1.3 Are terms used in data labels and in the data itself precise and accurate? Do they convey respect and acknowledge diversity? Have we included any data labels that make assumptions about the "standard" or "default" category? How might we allow for self-identification and respect differences in self-identification?
1.4 What might be some unintended consequences of the tools we are developing for well-intentioned purposes such as flagging hate speech? What are current limitations of NLP tools and where is additional research needed?

3.2 Layer 2: Organization of words and phrases

How we combine words and phrases into sentences and conversations also impacts our perceptions of social reality. We focus on

three opportunities here for NLP tools to improve language practices. First, an NLP tool could help identify passive language and suggest active language to more accurately assign agency to actors. Through active language, actions may be more perceptibly linked to the actors who carry them out, unlike passive language which can put the focus on the person or thing who is affected [21]. When speaking about enslavement of Africans, for example, people commonly use the passive voice (e.g., "Enslaved people were brought to the United States.") [59]. This framing minimizes the agency and responsibility of White enslavers and settler-colonials. Similarly, in reports of sexual violence, writers tend to use the passive voice, which highlights the role of the victim and minimizes the role of the assailant [21, 41, 55]. Active language more accurately frames who is responsible for an action, whether that is an individual responsible for a negative or positively framed action [38].[5] Secondly, in some cases, institutions are personified and assigned agency for actions that were carried out by individuals; this may allow those individuals not to be held accountable. Being precise about who is responsible ensures actions and their impacts are not obscured or hidden behind an institution. An NLP tool could flag potential instances of personification for writers to improve precision and accountability.

Finally, it is important to reflect on the organization of words and phrases in NLP datasets more broadly. NLP datasets may overemphasize loss and under-emphasize resilience of certain identities, thereby framing and reinforcing information from a deficit perspective. More specifically, using present tenses to describe ongoing actions or processes can help to accurately convey their impacts. Talking about Indigenous groups in the present tense helps strengthen an accurate narrative asserting that Indigenous people continue to thrive and have not been erased. Within NLP tool development, think critically about the dominant narratives in your datasets and consider including sources that celebrate how much has been maintained despite efforts by settler-colonials to interrupt traditions. Further, curating datasets that use language in ways that do not simply reinforce hegemonic narratives but portray agency and temporality more accurately can enable NLP tools that advance more accurate narratives.

Questions to utilize in developing NLP systems that advance equity and inclusion:

2.1 How might an NLP tool be built to help individuals or organizations utilize patterns of organizing words and phrases for equity and inclusion? For example, can a tool flag for human review (e.g., for journalists and writers) potential uses of the passive voice and personification of institutions?

2.2 How might we ensure that datasets include accurate data that does not replicate deficit-based narratives?

3.3 Layer 3: Patterns of language use over time

Patterns in how members of different groups are described repeatedly over time contribute to the formation of stereotypes based on these descriptions, even when those descriptions are erroneous [73]. Repeating words and phrases, and organizations of words and phrases, over time amplifies their effects. While this has traditionally resulted in the amplification of dominant narratives, careful curation of datasets and monitoring of NLP tools could help shift the narrative and amplify historically marginalized perspectives [17]. For example, repeatedly using gendered terms like "chairman" can reinforce an expectation that men prototypically fill that role. Because of patterns in use over time, even while using gender-neutral terms for different roles (such as "doctor" or "engineer"), people may assume a particular identity based on the role through associated stereotypes about who canonically fills such a role [36]. NLP tools could help to push back against such harmful patterns by outputting language that reflects more equitable narratives.

Repeated associations between words and different social identities are not always easy to spot. Wright [75] created a database of articles reporting on Black and White athletes and trained a support vector machine (SVM) to categorize athletes into racial categories based on how they were described in the news. The SVM was able to do so, and furthermore there was a lack of gradience in the SVM's certainty: the model was highly confident in its ability to correctly predict the race of the athlete just based on the lexical items that were used to describe the athlete. In the corpus, seemingly neutral words, like "athletic" and "wonderful," were not used equally to describe individuals of different races. Some were used more frequently for Black athletes ("athletic") and others for White athletes ("wonderful").

Wright's findings are linked to *indexicality*, which arises when two things are repeatedly juxtaposed over time, and by association one comes to indicate the other [67]. In Wright's study, "athletic" co-occurs frequently with Black athletes, and so it comes to be used by the SVM as a proxy to indicate blackness. Similar indexes form throughout language. White-sounding and Black-sounding names, for example, have been found to be more associated with positive and negative words respectively [24]. Linguistic features that are characteristic of AAE can serve as indexes to indicate blackness, or features more commonly used by women can index gender. This was shown in Amazon's resumé-screening tool. Not only did the tool penalize resumés containing the word "women's," but when direct mentions of "women's" were removed, the tool continued to disprefer female candidates [29]. The tool was found to favor candidates who used words such as "executed" or "captured" - words that were more commonly used in the male engineer's resumés [29]. While it may not be obvious to an observer that "executed" and "captured" are indicators of gender, because of their repeated use by men more than women, they have come to index masculinity. At the same time, NLP tools like Textio [5] - which flags gendered wording in job descriptions - have been employed to push back against these biases.

Furthermore, NLP systems such as translation tools can be updated to move unconscious biases to areas of more critical thinking, as Google Translate has begun doing [8]. Once the tool has generated its default translation in Spanish (for example), a rewriter algorithm detects whether the translation includes gendered features not present in the input. If it does, the translation is rewritten and users are given an option to select a feminine or masculine translation.

In building NLP tools, researchers often seek to include ever-greater amounts of data under the assumption that more data makes

[5]Not all passive language is harmful. However, as will be discussed in layer 3, the repeated use of passive language over time can add to potential harm.

the system more accurate [17]. However, if the data contain patterns reflecting human biases, more data will lead to the same biases. Even when tools are trained on carefully curated data, if they continue to learn through natural interaction with humans, they may replicate biases they are exposed to through humans. For example, Microsoft released a chatbot in 2016 that was trained on "modeled, cleaned, and filtered" public data; yet, in less than 24 hours, it was producing racist, transphobic, and other discriminatory language that it had learned through human interaction [72]. Instead of simply relying on greater quantities of data, it is important to be careful in how data is curated and whose voices are - or are not - represented (as will be discussed further in layer 4) [17].

Questions to utilize in developing NLP systems that advance equity and inclusion:

3.1 How might we develop NLP systems that flag language that advances harmful baises and offer recommendations?

3.2 In developing NLP tools that flag patterns of biased language, what narratives have we included in our dataset and how might that impact identity-based indexes that arise? Have we sought input from a diversity of experts, social scientists, and community members?

3.4 Layer 4: Power inequities

Not all groups have equal access to power and languages associated with power. For example, many language models draw data from the internet, but internet access and use varies greatly by gender, geography, and socio-economic status, resulting in skews in available language data. Some sites from which data is commonly scraped (e.g., Reddit, Wikipedia and Twitter) have particularly strong skews in who is represented. For example, 67% of Reddit users in the US are men and 70% are White [17]. On Twitter, marginalized individuals experience harassment - including pervasive abuse against women, and particularly Black women - which may lead to self-censorship [6]. The result is certain voices and opinions (e.g., those of White men) overrepresented online and, subsequently, in language models and datasets [17].

At the same time, there is great potential for positive change: by creating curated datasets that include an equitable representation of diverse language varieties and which seek to remove or otherwise mitigate harmful words and terms, NLP tools can become more inclusive and create a more equitable presence of linguistically diverse language resources. Increasing language diversity in NLP development could also help to solve misclassification issues through which tools often fail to identify speakers' language variety if it differs from the "standard" [18, 47]. Although language datasets often include dozens of languages, they tend to feature texts of "standard," monolingual language varieties [18]. While NLP tools are often meant to serve a wide public, if trained only on a subset of language data with an overrepresentation of one variety, they will not perform highly across all varieties. In an evaluation of four off-the-shelf language identifiers, for example, researchers found that Tweets by Black Americans were more often mistakenly labeled as non-English than those by White Americans [18]. Automated speech recognition (ASR) has also shown higher error rates for Black speakers than White speakers [19], likely due to insufficient audio data from Black speakers in the training data [51].[6] Higher error rates can have negative impacts on affected speakers. For example, if YouTube auto-captioning has higher error rates for one set of people, viewers who rely on captioning may not understand those speakers as well [69]. Since indexing text enables better searching, information from some speakers may also be missing from search results [69].

As these studies suggest, relying on language data that does not equitably represent the population can result in inequitable performance of an NLP tool - and inequitable outcomes for individuals - based on language variety. However, adherence to a monolingual language "standard" is far from the norm globally [12]; code-switching and use of different registers is common across a wide range of contexts. By creating datasets that contain a wider set of language varieties beyond the "standard" and which feature multilingual, dynamic ways of using language, researchers will be better equipped to create tools for realistic linguistic encounters in a wider range of environments. Furthermore, they will contribute to greater social equity by pushing back against standard language ideologies that place the burden on marginalized speakers to learn and use the "standard". Instead, if tools are built to serve a diverse set of speakers equally, their language varieties may also come to have equal status and prestige.

At the same time, data collected from marginalized speakers and communities must be done in a way that respects privacy and choice. Not all communities have the same definitions of intellectual property, and some language data may contain sensitive information including culturally specific knowledge [27, 70]. For some, language is considered communally held property and its use by individuals from outside the speaker community may require permission or special practices [25, 45, 63, 70]. In 2005, for example, Mapuche leaders sued Microsoft for translating their software into Mapudungun without prior permission [68], seeing it as a violation of their rights. To better serve those whose data may be used by NLP, learn about community standards for language and data use before diving into a project.

There is immense opportunity to approach projects from a collaborative framework through which members of language groups are fully included in the process of determining whether or how their language data is collected, what is done with it, and who maintains ownership. Utilizing participatory research models [26, 35, 49, 74] can help to facilitate equitable participation between various stakeholders. Given that some varieties are likely to remain underrepresented, NLP developers should recognize what data any given system was trained on, the sources of the data, and how those texts may contain biases or prioritize certain voices. Developers can document this information using tools like data statements for NLP and be transparent about the limitations of the tools built with a given dataset [16].

A second challenge at this layer is related to our perceptions of language. Determining whether language is considered "professional" or "unprofessional," for example, requires making a subjective judgement that can be influenced by our linguistic biases. For example, research shows that Tweets written in AAE are often considered more hateful or offensive than Tweets in "S"AE [30]. In addition to being trained on data over-representing "standard"

[6] Disparities by gender have also been found [51, 69].

language varieties and therefore being subject to greater error rates, deciding what hate speech is and labeling it as such is subjective [18]. The biases about language that data labellers hold may be reflected in the labels they assign. Also, whether a term is categorized as hate speech changes over time and depends on context [57]. As discussed in layer 1, words that may be offensive in one context may be reclaimed for self-empowerment in another. Collaborating with members of marginalized groups in the process of NLP development is key to understanding contextual differences.

Finally, as mentioned above, data labellers may introduce their own linguistic biases into the dataset [18]. If annotators don't speak or sign the language variety they're labelling, for example, they might label it as "unintelligible" or mislabel it. This can cause NLP systems trained on the data to perform worse for speakers of marginalized languages. NLP developers can ensure data labellers speak or sign the specific language varieties they are labelling, and are trained to counteract biases. Recruiting and training individuals who use marginalized language varieties may be a more time intensive and costly undertaking. However, it is more effective to ensure language data is accurately labeled.

Beyond how labels are applied to language, it is important to consider how linguistic profiling may result from NLP tools. Speech analysis is used for things like selecting candidates to hire and promote [20]. NLP outputs can include predictions of personality traits like enthusiasm, organization, and empathy based on training data gathered by individuals who submitted speech recordings and associated personality tests. However, the tools may perform worse for protected classes or penalize non-native speakers and those with speech disorders [20]. Additionally, people can manipulate the system by altering their speech, while some people may refuse to give a speech sample and may get excluded. Because the analysis involves many variables whose meanings can be opaque, it is quite possible for the tool to rely on associations between particular aspects of speech and their correlation with social identities. As in the case of automated resumé screeners picking up on subtle connections between word choice and gender to discriminate against women, tools analyzing speech may pick up on specific aspects and associate them with race, gender, or other protected categories. Given that many stereotypes about what "professional" or "articulate" speech sounds like are based on a prototypical White male speaker, these tools may use language to discriminate on the basis of social categories by proxy. Without transparency around how outputs are determined, we can't determine whether or not these tools are biased towards speech that is more like that of White, middle-class men.

All NLP systems have the opportunity to support and advance linguistic justice by including and serving speakers of a wide variety of languages. Because language can provide access to power, ensuring that speakers of different language varieties are given equitable access is essential. Furthermore, NLP tools provide a unique possibility to achieve greater access to information cross-linguistically as these tools become increasingly accurate at tasks like translation and allow for a greater set of language varieties to circulate and gain the prestige and recognition that they deserve.

Questions to utilize in developing NLP systems that advance equity and inclusion:

- 4.1 What language varieties are represented in our training data and outputs? Do these varieties reflect the range of language used by the population of potential users? Is our target population maximally inclusive?
- 4.2 Have we ensured that consent for use of language data has been given following culturally appropriate practices for the particular language community? Have we collaboratively and fairly engaged with marginalized language communities so that members of those groups can provide input and/or lead throughout the process from deciding whether or not to participate, to informing data collection, labeling and processing, to tool development and implementation? Does the tool address the needs and goals of the particular language community/ies?
- 4.3 Have we ensured appropriate privacy and ownership of language data?
- 4.4 Are data labellers fluent in the language variety they are working with? Have data labellers been trained to counter their implicit biases?
- 4.5 How might we be more transparent about the data our NLP tool is trained on and associated limitations of the tool? Have we audited our NLP systems to make sure that they work well for different language varieties, particularly target and potential user populations?
- 4.6 Who is the target population for our tool? Why? Are our choices of target audience inclusive or do they reflect harmful stereotypes? Have we included members of the target audience in the development of the tool?
- 4.7 Is the tool picking up on associations between language and social categories that could be used for discriminatory purposes? Can the data be curated to ameliorate such discrimination? If not, is the tool worth pursuing?

4 TECH COMPANY CASE STUDY

This framework was partially informed from a collaboration between researchers at UC Berkeley and leaders and practitioners at a large Silicon Valley tech firm, as well as broader academic research and industry trends. This short case study illustrates how the framework can be put into practice hypothetically within industry.

Within Layer 1, some teams at the tech firm have been interested in identifying imprecise and inaccurate terms within their content (Q1.3) and code (Q1.2), and recommend that those terms be updated with more inclusive alternatives. In fact, many tech companies have undertaken similar initiatives (e.g., Adobe [4], Apple [3], Google [7], and Microsoft [1]). The firm is interested in exploring automated tools that can flag potentially harmful terms for human users to evaluate and replace with more inclusive suggestions (Q1.1). We have discussed how to ensure the tool allows for context-specific insight, as well as handle situations like using person-first versus identity-first language, for example, and identified that adopting a rigid policy could have potentially harmful impacts for individuals whose language choices are not set as the standard (Q1.4). Layer 2 requires examination of words and phrases within a broader context. Automated tools being developed to flag potentially harmful terms can be expanded and made available to flag issues related to the organization of words and phrases (Q2.1). Ensuring datasets include

accurate, abundance-based representations of marginalized groups (Q2.2) is difficult; yet, a step towards this is greater documentation and transparency of datasets used to build NLP systems. The firm has expressed commitment and developed tools to support dataset transparency. Within layer 3, the firm has collaborated with interdisciplinary experts, including our team and community leaders (Q3.2), to develop inclusive language practices that can be built into the tool for flagging harmful content (Q3.1). Implementing layer 4 of the framework remains a challenge, which is not unexpected given this layer seeks to address power inequities broadly. However, the questions have prompted fruitful discussion about what types of data can/should be used to build NLP models (Q4.1) and who is qualified to label that data (Q4.4). Addressing concerns about consent for language data use (Q4.2), privacy and ownership (Q4.3), and transparency (Q4.5) is also a priority for the firm. Finally, the firm seeks to design products for a diverse set of target audiences (Q4.6) and has encouraged data labeling partners to better ameliorate discriminatory outcomes and linguistic bias (Q4.7). With one such group, UC Berkeley researchers piloted a lesson plan for data labellers that incorporates ideas in the framework on responsible language in machine learning training datasets, and the group has expressed interest in further piloting it in their internal training offerings.

5 CONCLUSION

Advancing equity and inclusion through language is critical to enabling a more just society. Within NLP, it requires careful reflection on what data is being used in training and evaluation of NLP tools. We must examine to what extent datasets include harmful terms, organizations of words, and patterns of use, and explore ways to mitigate or remove such occurrences from the datasets. We should also ask ourselves whose language data is included in datasets and whose language data is prioritized. If privileged language varieties continue to be the first varieties included in NLP development, they will continue to hold their position as gatekeepers. Instead, developing tools that serve a wider range of language users may help to decenter the hegemonic power of "standard" language varieties and can lead to greater empowerment for marginalized language users.

At the same time, we recognize a tension between the desire to include a greater diversity of language varieties within NLP, and the need to respect the decisions of language users who may not want their language data to be used for NLP. While creating NLP systems that encompass a greater diversity of languages has the potential to shift power and prestige, we must not use this as an excuse to engage in harmful practices such as extracting data from language communities, violating community standards for intellectual property rights, building tools that work against community needs, or violating individual and community privacy. Instead, we should pursue linguistic equity and justice by building NLP tools collaboratively, involving members of language communities as partners throughout the process, and with a willingness to shift or abandon a project that does not fit community needs.

In addition, we must strive to create equitable workplaces where NLP development is equitably led by members of different backgrounds, and whose ideas and contributions are valued. Creating an environment where concerns can be brought up and discussed will support a culture conducive to acknowledging, working through, and tackling these issues. Furthermore, creating a more inclusive environment will require that those in power not only allow for marginalized voices to be heard, but also uplift and amplify those voices in positions of decision-making and power.

Finally, we must remember that language and reality are mutually reinforcing. NLP has an incredible and growing influence, and the recommended practices are important to implement, but we must also continue to understand how discrimination manifests around us. We must work to eliminate the factors that make it difficult for marginalized groups to equitably participate within communities, organizations, and society. We hope this work serves as a useful starting point for considering how NLP tools can center equity and inclusion, enable linguistic justice, and ultimately advance social justice.

ACKNOWLEDGMENTS

We thank Kellie McElhaney, Edwin Ko, members of our Working Group including practitioners and leaders in the tech industry, and the anonymous reviewers for providing feedback on this work. This research was funded by the Center for Equity, Gender, and Leadership at the UC Berkeley Haas School of Business along with sponsorship by a leading Silicon Valley tech firm. Any errors are our own.

REFERENCES

[1] [n.d.]. Bias-free communication. https://docs.microsoft.com/en-us/style-guide/bias-free-communication
[2] [n.d.]. *Evaluating neural toxic degeneration.* Technical Report. Allen Institute for AI. https://toxicdegeneration.allenai.org/
[3] [n.d.]. Overview - Apple Style Guide. https://help.apple.com/applestyleguide/#/apdcb2a65d68
[4] [n.d.]. Terminology changes in Premiere Pro, After Effects, and Audition. https://helpx.adobe.com/x-productkb/multi/terminology-changes-video-products.html
[5] [n.d.]. Textio. https://textio.com/products/tone/
[6] [n.d.]. *Toxic Twitter - The Psychological Harms of Violence and Abuse Against Women Online.* Technical Report. Amnesty International. https://www.amnesty.org/en/latest/research/2018/03/online-violence-against-women-chapter-6/#topanchor
[7] [n.d.]. Writing Inclusive Documentation. https://developers.google.com/style/inclusive-documentation
[8] 2020. A Scalable Approach to Reducing Gender Bias in Google Translate. ai.googleblog.com/2020/04/a-scalable-approach-to-reducing-gender.html
[9] 2021. *Responsible Language for AI and ML.* Technical Report. Center for Equity, Gender, and Leadership at the UC Berkeley Haas School of Business. https://haas.berkeley.edu/equity/industry/playbooks/responsible-language-for-ai/
[10] Yásnaya Elena Aguilar Gil. 2016. El nacionalismo y la diversidad lingüística. *Tema y Variaciones de Literatura* 47 (2016), 45–47.
[11] H. Samy Alim. 2007. Critical Hip-Hop Language Pedagogies: Combat, Consciousness, and the Cultural Politics of Communication. *Journal of Language, Identity & Education* 6, 2 (2007), 161–176. https://doi.org/10.1080/15348450701341378
[12] Peter Auer and Li Wei. 2007. Introduction: Multilingualism as a problem? Monolingualism as a problem? In *Handbook of Multilingualism and Multilingual Communication*, Peter Auer and Li Wei (Eds.). De Gruyter Mouton, Berlin. https://doi.org/10.1515/9783110198553.0.1
[13] April Baker-Bell. 2020. *Linguistic Justice: Black Language, Literacy, Identity, and Pedagogy.* Routledge, New York.
[14] John Baugh. 2000. Racial Identification by Speech. *American Speech* 75, 4 (2000), 362–364.
[15] Charlotte Baxter-Read. 2021. How Acrolinx is Helping Salesforce Create More Inclusive Technical Documentation. https://www.acrolinx.com/blog/how-acrolinx-is-helping-salesforce-create-more-inclusive-technical-documentation/
[16] Emily 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 (2018), 587–604. https://aclanthology.org/Q18-1041.pdf

[17] Emily Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Virtual Event, 610–623. https://doi.org/10.1145/3442188.3445922

[18] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. *Annual Meeting of the Association for Computational Linguistics* (July 2020), 5454–5476. https://aclanthology.org/2020.acl-main.485.pdf

[19] Su Lin Blodgett and Brendan O'Connor. 2017. Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English. https://arxiv.org/pdf/1707.00061.pdf

[20] Miranda Bogen and Aaron Rieke. 2018. *Help Wanted: An examination of Hiring Algorithms, Equity, and Bias.* Technical Report. Upturn. https://www.upturn.org/reports/2018/hiring-algorithms/

[21] Gerd Bohner. 2001. Writing about rape: Use of the passive voice and other distancing text features as an expression of perceived responsibility of the victim. *British Journal of Social Psychology* 40 (2001), 515–529.

[22] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *30th Conference on Neural Information Processing Systems*. Barcelona, Spain, 4356–4364. https://doi.org/doi/pdf/10.5555/3157382.3157584

[23] Lydia X. Z. Brown. 2011. The Significance of Semantics: Person-First Language: Why It Matters. https://www.autistichoya.com/2011/08/significance-of-semantics-person-first.html

[24] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186. https://doi.org/10.1126/science.aal4230

[25] S. R. Carroll, I. Garba, O. L. Figueroa-Rodríguez, J. Holbrook, R. Lovett, S. Materechera, M. Parsons, K. Raseroka, D. Rodriguez-Lonebear, R. Rowe, J. D. Walker, J. Anderson, and M. Hudson. 2020. The CARE Principles for Indigenous Data Governance. *Data Science Journal* 19, 1 (2020), 43. https://doi.org/10.5334/dsj-2020-043

[26] J. M. Chevalier and D. J. Buckles. 2019. *Participatory Action Research: Theory and Methods for Engaged Inquiry.* Routledge, New York.

[27] Jon Corbett and Tim Kulchyski. 2009. Anti social-computing: indigenous language, digital video and intellectual property. In *Change at Hand: Web 2.0 for Development*, Holly Ashley (Ed.). International Institute for Environment and Development, United Kingdom, 52–58.

[28] 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 (2020), 389–407. https://doi.org/10.1146/annurev-linguistics-011718-011659

[29] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters* (Oct. 2018). https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

[30] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In *Proceedings of the Third Workshop on Abusive Language Online*. Florence, Italy, 25–35. https://doi.org/10.18653/v1/W19-3504

[31] Bethany Davila. 2012. Indexicality and "Standard" Edited American English: Examining the Link Between Conceptions of Standardness and Perceived Authorial Identity. *Written Communication* 29, 1 (2012), 180–207. https://doi.org/10.1177/0741088312438691

[32] Katie Drager. 2010. Sociophonetic Variation in Speech Perception. *Language and Linguistics Compass* 4, 7 (2010), 473–480. https://doi.org/10.1111/j.1749-818X.2010.00210.x

[33] Dana S. Dunn and Erin E. Andrews. 2015. Person-first and identity-first language: Developing psychologists' cultural competence using disability language. *American Psychologist* 70, 3 (2015), 255–264. https://doi.org/10.1037/a0038636

[34] Nelson Flores and Jonathan Rosa. 2015. Undoing Appropriateness: Raciolinguistic Ideologies and Language Diversity in Education. *Harvard Educational Review* 85, 2 (2015), 149–171.

[35] Michael Gaffney. 2008. Participatory Action Research: An Overview, What makes it tick? *Kairaranga* 9 (2008), 9–15.

[36] Peter Glick, Korin Wilk, and Michele Perreault. 1995. Images of Occupations: Components of Gender and Status in Occupational Stereotypes. *Sex Roles* 32, 9-10 (1995), 565–582. https://link.springer.com/content/pdf/10.1007/BF01544212.pdf

[37] Kate Goddard, Abdul Roudsari, and Jeremy C. Wyatt. 2012. Autmation bias: a systematic review of frequency, effect, mediators, and mitigators. *Journal of the American Medical Informatics Association* 19, 1121-127 (2012). https://doi.org/10.1136/amiajnl-2011-000089

[38] Stephan Greene and Philip Resnik. 2009. More than Words: Syntactic Packaging and Implicit Sentiment. In *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL*. Boulder, Colorado, 503–511. https://www.aclweb.org/anthology/N09-1057.pdf

[39] R Harlow. 1998. Some languages are just not good enough. In *Language Myths*, Laurie Bauer and Peter Trudgill (Eds.). Penguin Books, London, 9–14.

[40] Andrew Hartman. 2003. Language as oppression: The English-only movement in the United States. *Socialism and Democracy* 17, 1 (2003), 187–208. https://doi.org/10.1080/08854300308428349

[41] Nancy M. Henley, Michelle Miller, and Jo Anne Beazley. 1995. Syntax, Semantics, and Sexual Violence: Agency and the Passive Voice. *Journal of Language and Social Psychology* 14, 1-2 (1995), 60–84. https://doi.org/10.1177/0261927X95141004

[42] Jane H. Hill. 2008. *The Everyday Language of White Racism.* John Wiley & Sons Ltd., West Sussex.

[43] Frank Houghton and Sharon Houghton. 2018. "Blacklists" and "whitelists": A salutary warning concerning the prevalence of racist languge in discussions of predatory publishing. *Journal of the Medical Library Association* 106, 4 (2018). https://doi.org/10.5195/jmla.2018.490

[44] Claretha Hughes and Ketevan Mamiseishvili. 2013. Linguistic Profiling in the Workforce. In *Diversity in the Workforce: Current Issues and Emerging Trends*, Marilyn Y. Byrd and Chaunda L. Scott (Eds.). Routledge, New York, 249–265.

[45] Christopher Hutton. 2010. Who owns language? Mother tongues as intellectual property and the conceptualization of human linguistic diversity. *Language Sciences* 32 (2010), 638–647. https://doi.org/10.1016/j.langsci.2010.06.001

[46] John Earl Joseph. 1987. *Eloquence and power: the rise of language standards and standard languages.* Blackwell, Oxford.

[47] David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. 2017. Incorporating dialectal variability for socially equitable identification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vol. 2. Vancouver, Canada, 51–57. https://doi.org/10.18653/v1/P17-2009

[48] Lorcan Kenny, Caroline Hattersley, Bonnie Molins, Carole Buckley, Carol Povey, and Elizabeth Pellicano. 2015. Which terms should be used to describe autism? Perspectives from the UK autism community. *Autism* (2015), 1–21. https://doi.org/10.1177/1362361315588200

[49] S. A. Kidd and M. J. Kral. 2005. Practicing participatory action research. *Journal of Counseling Psychology* 55, 2 (2005), 187–195.

[50] Sharese King. 2020. From African American Vernacular English to African American Language: Rethinking the Study of Race and Language in African Americans' Speech. *Annual Review of Linguistics* 6 (2020), 285–300. https://doi.org/10.1146/annurev-linguistics-011619-030556

[51] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connon Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences of the United States of America* 117, 14 (April 2020), 7684–7689. https://doi.org/10.1073/pnas.1915768117

[52] Claire Kramsch. 2004. Language, Thought, and Culture. In *The Handbook of Applied Linguistics*, Alan Davies and Catherine Elder (Eds.). Blackwell Publishing Ltd, Malden, MA, 235–261.

[53] Rosina Lippi-Green. 2012. *English with an Accent: Language Ideology, and Discrimination in the United States.* Routledge, New York.

[54] Robert W. Livingston, Ashleigh Shelby Rosette, and Ella F. Washington. 2012. Can an Agentic Black Woman Get Ahead? The Impact of Race and Interpersonal Dominance on Perceptions of Female Leaders. *Psychological Science* 23, 4 (2012), 354–359. https://doi.org/10.1177/0956797611428029

[55] Rachael Graham Lussos and Lourdes Fernandez. 2018. Assault and Accusation Without Agents: Verb Voice in Media Narratives of Campus Sexual Assault. *Journal of Mason Graduate Research* 5, 2 (2018), 108–127. https://doi.org/10.13021/G8jmgr.v5i2.1984

[56] Francesca A. López. 2017. Altering the Trajectory of the Self-Fulfilling Prophecy: Asset-Based Pedagogy and Classroom Dynamics. *Journal of Teacher Education* 68, 2 (2017), 193–212. https://doi.org/10.1177/0022487116685751

[57] Louise Matsakis. 2018. To Break a Hate-Speech Detection Algorithm, Try 'Love'. https://www.wired.com/story/break-hate-speech-algorithm-try-love/?utm_source=WIR_REG_GATE

[58] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *Comput. Surveys* 54 (2021), 1–35. Issue 6. https://doi.org/10.1145/3457607

[59] Robert B. Moore. 2006. Racism in the English Language. In *The Production of Reality* (fourth edition ed.), Jodi O'Brien (Ed.). Pine Forge Press, Thousand Oaks, CA, 119–126.

[60] Luis Noe-Bustamante, Lauren Mora, and Mark Hugo Lopez. 2020. *About One-in-Four U.S. Hispanics Have Heard of Latinx, but Just 3% Use It.* Technical Report. Pew Research Center. https://www.pewresearch.org/hispanic/2020/08/11/about-one-in-four-u-s-hispanics-have-heard-of-latinx-but-just-3-use-it/

[61] Matthew R. Pearson. 2010. How "undocumented workers" and "illegal aliens" affect prejudice toward Mexican immigrants. *Social Influence* 5, 2 (2010), 118–132. https://doi.org/10.1080/15534511003593679

[62] Martha Ramos Duffer. 2018. Language Matters: Competent Mental Health Treatment for Latina/Latino/Latinx Undocumented Immigrants - A Comment on Alfaro and Bui. *Ethics & Behavior* 28, 5 (2018), 389–392. https://doi.org/10.1080/10508422.2018.1463533

[63] Angela R. Riley. 2000. Recovering Collectivity: Group Rights to Intellectual Property in Indigenous Communities. *Cardozo Arts & Entertainment Law Journal* 18, 1 (2000), 175–226. https://ssrn.com/abstract=2449057
[64] Tracy L. Robinson. 1999. The Intersections of Dominant Discourses Across Race, Gender, and Other Identities. *Journal of Counseling & Development* 77, 1 (1999), 73–79. https://doi.org/10.1002/j.1556-6676.1999.tb02423.x
[65] Donald L. Rubin. 1992. Nonlanguage Factors Affecting Undergraduates' Judgements of Nonnative English-Speaking Teaching Assistants. *Research in Higher Education* 33 (1992), 511–532.
[66] Mike Seele. 2020. Striking Out Racist Terminology in Engineering. https://www.bu.edu/articles/2020/striking-out-racist-terminology-in-engineering/
[67] Michael Silverstein. 1979. Language Structure and Linguistic Ideology. In *The Elements*, P. Clyne, W. Hanks, and C. Hofbauer (Eds.). Chicago Linguistic Society, Chicago, 193–248.
[68] Margaret Speas. 2013. Language Ownership and Language Ideologies. In *Negotiating Culture: Heritage, Ownership, and Intellectual Property*, Laetitia La Follette (Ed.). University of Massachusetts Press, Amherst, 101–121.
[69] Rachael Tatman and Conner Kasten. 2017. Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions. In *INTERSPEECH 2017*. Stockholm, Sweden, 934–938. https://doi.org/10.21437/Interspeech.2017-1746
[70] Sheri Tatsch. 2004. Language Revitalization in Native North America - Issues of Intellectual Property Rights and Intellectual Sovereignty. *Collegium Antropologicum* 28, 1 (2004), 257–262. https://pubmed.ncbi.nlm.nih.gov/15156749/
[71] Eve Tuck. 2009. Suspending Damage: A Letter to Communities. *Harvard Educational Review* 79, 3 (2009), 409–427.
[72] James Vincent. 2016. Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day. *The Verge* (March 2016). https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist.
[73] J. Celeste Walley-Jean. 2009. Debunking the Myth of the "Angry Black Woman": An Exploration of Anger in Young African American Women. *Black Women, Gender + Families* 3, 2 (2009), 68–86. https://www.jstor.org/stable/10.5406/blacwomegendfami.3.2.0068
[74] W. F. Whyte, D. J. Greenwood, and Peter Lazes. 1991. Participatory Action Research: Through Practice to Science in Social Research. In *Participatory action research*, W. F. Whyte (Ed.). Sage Publications, Thousand Oaks, CA.
[75] Kelly Wright. 2021. Race and Language: Considering Black Linguistic Experiences. https://www.youtube.com/watch?v=np46elWCaKc
[76] Mihir Zaveri. 2020. This lawmaker wants to remove the words 'illegal alien' from the law. *New York Times* (Feb. 2020). https://www.nytimes.com/2020/02/13/us/politics/colorado-illegal-immigrants.html

A RESEARCH METHODS

This paper is informed by a systematic literature review carried out by a multidisciplinary research team spanning linguistics, sociology, economics, and computer science from August 2020 through June 2021. The research was undertaken as part of a collaboration between researchers at the Center for Equity, Gender, and Leadership at the UC Berkeley Haas School of Business and, to understand real-world needs for practical application, a leading tech firm in Silicon Valley. We sought to 1) understand the links between language and power,[7] 2) explore the construction of race and anti-Black racism in the United States, and the connections to language, 3) interrogate the links between machine learning (ML) and NLP in relation to power and inequality, and 4) identify equitable and inclusive language practices within artificial intelligence (AI) systems, particularly related to race[8] and ethnicity. We searched academic journals (language / linguistics, social psychology, sociology, anthropology, ethnic studies, computer science, engineering, data science) to inform these subjects. Beyond academic sources, we relied on other sources including articles and reports. The search for sources was confined to the past 30 years (1990-2020), with exceptions for particularly influential texts.

Our framework is consciously designed from the perspective of *abundance*[9] to counter dominant narratives that promote harmful *deficit* framings of marginalized groups [34, 71] that focus on how they contrast with dominant groups rather than recognizing their inherent, independent value. For example, instead of framing AAE as "non-standard" in comparison to "S"AE, we can instead affirm that AAE and "S"AE are two equally positioned varieties of English. By questioning how different language varieties are valued within NLP development, we can develop new standards that more equitably value a diversity of linguistic practices.

[7] We consider power to be the ability to implement one's own will.
[8] In this work, we conceptualize race as a socially constructed category that emerged during the Trans-Atlantic slave trade. Race is fundamentally about power differentials between racial groups that advantage some while marginalizing others.
[9] We use the term *abundance* (suggested by Beth Piatote, p.c.) to refer to what other scholars frame as assets [56].