Understanding the Impacts of Language Technologies' Performance Disparities on African American Language Speakers

Jay L. Cunningham*
University of Washington
jaylcham@uw.edu

Su Lin Blodgett
Microsoft Research
sulin.blodgett@microsoft.com

Hal Daumé III
University of Maryland
Microsoft Research
hal3@umd.edu

Christina Harrington
Carnegie Mellon University
Google Research
cnharrington@google.com

Hanna Wallach
Microsoft Research
wallach@microsoft.com

Michael Madaio
Google Research
madaiom@google.com

Abstract

This paper examines the experiences of African American Language (AAL) speakers when using language technologies. Previous work has used quantitative methods to uncover performance disparities between AAL speakers and White Mainstream English speakers when using language technologies, but has not sought to understand the impacts of these performance disparities on AAL speakers. Through interviews with 19 AAL speakers, we focus on understanding such impacts in a contextualized and human-centered manner. We find that AAL speakers often undertake invisible labor of adapting their speech patterns to successfully use language technologies, and they make connections between failures of language technologies for AAL speakers and a lack of inclusion of AAL speakers in language technology design processes and datasets. Our findings suggest that NLP researchers and practitioners should invest in developing contextualized and human-centered evaluations of language technologies that seek to understand the impacts of performance disparities on speakers of underrepresented languages and language varieties.

1 Introduction

Prompted by the widespread availability of language technologies, researchers have questioned whether these technologies are sufficiently inclusive of speakers of underrepresented languages and language varieties (e.g., Koenecke et al., 2020; Aji et al., 2022; Harrington et al., 2022; Kantharuban et al., 2023; Martin and Wright, 2023). Most notably, previous work has found that language technologies often exhibit performance disparities between African American Language (AAL) speakers and White Mainstream English (WME) speakers (Koenecke et al., 2020; Deas et al., 2023).[1]

Much of this work—as well as work on performance disparities between other social groups when using language technologies—has used quantitative methods to uncover the existence and extent of performance disparities (e.g., Dixon et al., 2018; Koenecke et al., 2020; Sarı et al., 2021). However, for the most part, this work has not sought to understand the impacts of these performance disparities on AAL speakers. For example, although Koenecke et al. (2020) wrote that *"the performance gaps we have documented suggest it is considerably harder for African Americans to benefit from the increasingly widespread use of speech recognition technology... These disparities may also actively harm African American communities...,"* they did not examine the experiences of AAL speakers when using language technologies. In contrast, our paper focuses on understanding these experiences in a contextualized and human-centered manner, illuminating the relationships between linguistic discrimination and failures of language technologies.

We take a qualitative research approach to demonstrate the need for contextualized and human-centered evaluations of language technologies that seek to understand the impacts of performance disparities on speakers of underrepresented languages and language varieties. Specifically, we conduct interviews with 19 AAL speakers to examine their experiences when using language technologies, asking, *What fairness-related harms might AAL speakers encounter when using language technologies?* We find that AAL speakers often undertake invisible labor of adapting their speech patterns to successfully use language technologies (section 4.1), and they make connections between failures of language technologies for AAL

*Jay Cunningham was a research intern with the Microsoft Research FATE group when this work was conducted.

[1]We use "White Mainstream English" to refer to the racialized dominant variety of American English (see section 2).

Findings of the Association for Computational Linguistics: ACL 2024, pages 12826–12833

speakers and a lack of inclusion of AAL speakers in language technology design processes and datasets (section 4.2). We emphasize that although quantitative methods can uncover the existence and extent of performance disparities, they do not help us understand the impacts of these performance disparities. To understand such impacts, qualitative methods are a better fit because they can be used to probe deeper into issues around social inequalities and illuminate the relationships between linguistic discrimination and failures of language technologies.

2 Related Work

2.1 African American Language

We use "African American Language" (AAL)[2] to refer to *"the grammatically patterned variety of English used by many, but not all and not exclusively, African Americans in the United States"* (Grieser, 2022). AAL is a community-based variety of American English. We use "White Mainstream English" (WME) to refer to the racialized dominant variety of American English.[3]

Many sociolinguistic studies have investigated the role of AAL in African American[4] communities, finding that its deeply rooted history encompasses cultural, social, and linguistic values (Rickford, 1999; Yosso, 2005). AAL has been sustained through generations of African American communities, shaping how some African Americans navigate society. However, because of language ideologies that treat AAL as incorrect, undesirable, inappropriate, or unprofessional, AAL speakers often experience linguistic discrimination in a range of social settings, including asylum, citizenship, criminal justice, education, employment, and housing (e.g., Lippi-Green, 2012; Rickford and King, 2016; Baugh, 2018; Baker-Bell, 2020b; Craft et al., 2020). As a result, some AAL speakers adapt their speech patterns or even code-switch to WME to avoid this linguistic discrimination or to adhere to societal norms for appropriateness and professionalism (e.g., Jones and Shorter-Gooden, 2009; Young and Barrett, 2018; Kinzler, 2020; Johnson et al., 2022).

2.2 Fairness-Related Harms Caused by Language Technologies

Previous work has focused on identifying, measuring, and mitigating a variety of fairness-related harms caused by language technologies (e.g., Crawford, 2017; Bird et al., 2020; Weidinger et al., 2021; Dev et al., 2022). In the context of automated speech recognition (ASR) technologies, researchers have found that such technologies often exhibit poor performance for members of historically marginalized groups, including AAL speakers (Koenecke et al., 2020; Ngueajio and Washington, 2022). For example, Martin and Wright (2023) explored how linguistic features of AAL likely cause failures of ASR technologies, Tatman and Kasten (2017) demonstrated that ASR technologies exhibit higher word error rates for African American users than for white users, and Wassink et al. (2022) found that ASR technologies exhibit poor performance for African American, Native American, and ChicanX users. Researchers have also used quantitative methods to uncover the existence and extent of performance disparities between AAL speakers and WME speakers across a variety of text-based language technologies (e.g., Blodgett et al., 2018; Davidson et al., 2019; Sap et al., 2019; Harris et al., 2022; Deas et al., 2023).

2.3 Impacts of Performance Disparities

In addition to quantitative work showing that language technologies often exhibit poor performance for AAL speakers, emerging qualitative work has also explored African Americans' experiences with language technologies. For example, Harrington et al. (2022) found that African American elders face communication challenges with voice assistants because they must code-switch to WME to ensure their speech is recognized, while Brewer et al. (2023) additionally found that African American elders do not expect voice assistants to understand their speech patterns or to possess cultural and regional knowledge relevant to African American communities. Other researchers have studied the impacts of these experiences, finding feelings of being othered and emotional responses of anger, frustration, and disappointment (Mengesha et al., 2021), as well as lowered individual and collective

[2]AAL has had many different names, including African American English (AAE) and African American Vernacular English (AAVE) (Green, 2002; Wolfram and Schilling, 2015; Rickford and King, 2016; King, 2020; Becker, 2013). We emphasize that people's perceptions of AAL, and the names they use for it, vary widely (Wassik and Curzan, 2004).

[3]WME as the dominant variety of American English has likewise had many different names, including Standard American English (SAE) (e.g., Wolfram and Schilling, 2015) and Mainstream U.S. English/Mainstream American English (MUSE/MAE) (e.g., Baker-Bell, 2020a; Harris et al., 2022).

[4]We use "African American" to refer to people who identify as Americans originating from Black racial groups in Africa. During the interviews, we encouraged each participant to use their preferred terminology, which we then also used.

self-esteem (Wenzel et al., 2023). Our paper builds upon this body of qualitative work by arguing that NLP researchers and practitioners should invest in developing contextualized and human-centered evaluations of language technologies (e.g., Heuer and Buschek, 2021) that seek to understand the impacts of performance disparities on speakers of underrepresented languages and language varieties, illuminating the relationships between linguistic discrimination and failures of language technologies.

3 Methods

3.1 Participants

To investigate our research question, we used semi-structured interviews—a qualitative method commonly used in HCI research to understand people's experiences, desires, and concerns (Olson and Kellogg, 2014; Brinkmann and Kvale, 2018). We recruited participants through posts on a large technology company's internal message boards for affinity groups, posts on social media, and snowball sampling through email and word of mouth. Participants were required to reside in the U.S. and to be at least 18 years old. In total, the first author interviewed 19 participants, each of whom identified as being an AAL speaker and as being African American. All participants were between the ages of 22 and 59, with a mean age of 40. Seven participants identified as being men, while 12 participants identified as being women. Participants resided in 11 different U.S. states, covering all five major U.S. geographical regions. We asked participants to list the names they use for AAL.[5] During the interviews, we encouraged each participant to use their preferred terminology, which we then also used.

3.2 Data Collection

All interviews were conducted between July and September, 2021. Each interview was 45–60 minutes long. Due to the COVID-19 pandemic, all interviews were conducted on a remote meeting platform. Participants were compensated with $30 USD gift cards, and our study was approved by our institution's IRB. Each interview focused on that participant's use of AAL, their use of language technologies, and their experiences as an AAL speaker when using language technologies. We presented participants with a broad definition of language technologies[6] and examples of different types of language technologies.[7] We asked questions including: *Are you familiar with what language technologies are and can you tell me about a recent experience you've had?* and *As an AAL speaker, have you had negative experiences with language technologies, and if so, can you describe them?*

3.3 Data Analysis

To understand the data from our interviews, we conducted an inductive thematic analysis, following Braun and Clarke (2006). Inductive thematic analysis is a method commonly used in psychology research, in which researchers create "codes" or labels to identify passages in the data that are relevant to their research questions, inductively grouping these codes into higher-level themes through iterative discussions and consensus-building exercises with multiple members of the research team (Braun and Clarke, 2006; Olson and Kellogg, 2014).

4 Findings

Our findings shed light on the experiences of AAL speakers when using language technologies, in turn helping us understand the impacts of performance disparities on speakers of underrepresented languages and language varieties. First, we find that AAL speakers often undertake invisible labor of adapting their speech patterns to successfully use language technologies. Second, we find that AAL speakers make connections between failures of language technologies for AAL speakers and a lack of inclusion of AAL speakers in language technology design processes and datasets. In section 5, we unpack the implications of these findings for NLP researchers and practitioners.

4.1 AAL Speakers Undertake Invisible Labor to Use Language Technologies

We find that AAL speakers often undertake considerable labor to successfully use language technologies,[8] including correcting their outputs, proac-

[5] Names included African American Language (AAL), African American Vernacular English (AAVE), Ebonics, Slang, Black Language, and other self-defined names.

[6] Our definition: "Language technologies can be defined as computer programs, applications, or devices that can analyze, produce, modify, or respond to human text and speech."

[7] Our examples: ASR technologies (e.g., Amazon's Alexa, Google's Assistant, Apple's Siri), automated transcription technologies (e.g., in Microsoft Teams and Zoom), and text augmentation technologies (e.g., auto-correct, text prediction).

[8] Although we presented participants with a broad definition of language technologies and examples of different types of language technologies, most participants described experiences with ASR technologies via smart home devices such as Amazon's Alexa or voice assistants like Google's Assistant

tively code-switching to WME, asking others to help, or manually entering inputs instead of using ASR technologies. However, this labor is not made visible by quantitative methods for uncovering the existence and extent of performance disparities.

Correcting language technologies' outputs. Participants described how language technologies often fail to recognize AAL, thus requiring them to correct their outputs. P8 explained, *"I do qualitative research, interviewing Black patients, and these transcription services are not interpreting my participants' speech data correctly... I often have to listen and manually correct these transcripts, causing my research to take more time, but it's not a surprise at all."* This labor means that participants often do not enjoy language technologies' supposed benefits. For example, P8 went on to share, *"if I have to change the sentence three or four times, I might as well just do it manually."*

Proactively code-switching to WME. Participants described proactively code-switching to WME when using language technologies, which they experienced as being effortful. For example, P4 said, *"you have to alter your voice to make it more white sounding."* For some participants, code-switching to WME was about more than just adapting their speech patterns. As P7 described, *"I feel like I got to change me to interface with this technology and that's just not worth the effort."*

Asking others to help. In other cases, participants described how they needed to find other people to communicate with their smart home devices or voice assistants because code-switching to WME was difficult or still didn't work for them. For example, P11 told us, *"I got a Google Home. Ah, man, I don't even talk to it. I let my daughter or somebody talk to it because it never understands."*

Manually entering inputs. Some participants described how they manually entered inputs instead of using ASR technologies. For example, P18 said, *"Look, I could type it faster than trying to speech-to-text.' It makes me not want to use the technology, because it doesn't work for me and how I communicate."* Another participant (P5) explained, *"I was using Siri to send out messages, but it was misinterpreting what I was saying and the context... It can be extremely frustrating to the point where I don't want to use it anymore—I don't care to."*

or Apple's Siri. That said, some participants described experiences with other language technologies, such as automated transcription technologies or text augmentation technologies.

4.2 AAL Speakers Make Connections Between Failures of Language Technologies and a Lack of Inclusion in Design Processes and Datasets

Failures of language technologies imply that AAL is illegitimate. Participants often described failures of language technologies for AAL speakers as failures to treat their culture and language use as legitimate. For example, P5 spoke about their experiences with Grammarly, a tool to support the writing process: *"Why is it assumed that when I type AAL, it's a typo? I meant what I said."* Additionally, participants pointed out that language technologies can fail to correctly recognize names that are culturally African American. For example, P5 stated, *"I have a friend named Jhamal. He spells it J-H-A-M-A-L... But, if it's a Black-sounding name, Siri is going to get it incorrect."*

Some participants found these kinds of corrections to be useful in professional contexts, but not personal ones, reflecting pressures to code-switch to WME to adhere to societal norms for appropriateness and professionalism. For example, P10 said, *"I don't have auto-correct activated for my text messages, but when writing e-mails or using Microsoft Word, as it corrects grammar and gives your alternatives for word use... I love it for that!"*

Participants described how language technologies reproduce ideologies about what constitutes "good" English in ways that reinforce linguistic discrimination experienced in other parts of their lives. For example, P10 explained, *"I think that hearing certain voices, forces us into a subconscious code switch, or a subconscious acceptance of what is proper, what is standard, what is good English."*

A lack of inclusion of AAL speakers in design processes. Additionally, some participants theorized about the causes of failures of language technologies for AAL speakers, attributing them to a lack of inclusion of AAL speakers in language technology design processes. For example, P2 told us, *"I don't want to call a technology racist. Right? I can't call it racist. The code may lack input to take into consideration Black language, which isn't fair."* Similarly, P1 explained, *"I don't ever blame the technology; I blame the people that made it... clearly there aren't Black coders in the room to be like, 'Hey, you know what, this isn't picking up on these key words for Black speakers.'"*

Participants felt that this perceived lack of inclusion of AAL speakers in language technology

design processes meant that there were gaps in language technology developers' knowledge about AAL, in ways that may have led to performance disparities. For example, P2 said, *"It's a difference between racism and ignorance... Things you [designers] just don't know to consider, even though it may be particular to a specific racial group, doesn't make something racist, but it can be biased and discriminatory."* Other participants emphasized that just being in the room is not enough, noting that diversity does not always equal empowerment—e.g., *"Even if you have the diversity in the room, do those people feel comfortable enough to speak up on the issue?"* (P7)—and inclusion may not lead to prioritization of AAL speakers' needs.

A lack of inclusion of AAL speakers in datasets. Participants often attributed failures of language technologies for AAL speakers to a lack of inclusion of AAL speakers in language technology datasets. For example, P1 said, *"[I]t's literally data-driven... you just need to have more diverse data collection sets of Black speakers."* As another participant (P4) put it, *"Apple's Siri, Amazon's Alexa and all her friends need to go and start getting [AAL] words in these systems."* Participants felt that even if language technologies appeared to learn from their users, they didn't appear to learn from them specifically. P2 explained, *"Alexa is supposed to learn, but she doesn't learn me, for sure. So no matter how much I say it, it's still going to assume something else, because it's learning from everybody, just not me."*

For one participant (P10), failures of language technologies for AAL speakers were inevitable without more diverse datasets: *"If you're using the same standard English database to teach different types of language technology, then the same inherent biases against AAL will be put into it."* This participant saw better data collection processes as being an essential part of inclusive design: *"How do we capture the vernacular? How do we gather the appropriate terminology? How do we initiate the collection of relevant data? How do we expand our data collection efforts to include diverse cultural perspectives from around the world?"*

5 Discussion

Although previous work has focused on identifying, measuring, and mitigating a variety of fairness-related harms caused by language technologies, there remains ample room for further examination of the experiences of AAL speakers when using language technologies. For example, much of the work to date on performance disparities between AAL speakers and WME speakers when using language technologies has used quantitative methods to uncover the existence and extent of such performance disparities (e.g., Koenecke et al., 2020), but has not sought to understand their impacts on AAL speakers. This narrow focus limits what NLP researchers and practitioners are able to understand about the experiences of speakers of underrepresented languages and language varieties. In contrast, qualitative methods can be used to probe deeper into issues around social inequalities and illuminate the relationships between linguistic discrimination and failures of language technologies.

Our findings offer insights into the impacts of performance disparities on AAL speakers, including invisible labor they undertake to adapt their speech patterns to successfully use language technologies, as well as connections they make between failures of language technologies for AAL speakers and a lack of inclusion of AAL speakers in language technology design processes and datasets. These impacts are not made visible by quantitative methods for uncovering the existence and extent of performance disparities. We therefore urge NLP researchers and practitioners to invest in developing contextualized and human-centered evaluations of language technologies that seek to understand the impacts of performance disparities on speakers of underrepresented languages and language varieties. These evaluations could use qualitative methods, but they could also also use quantitative methods, albeit in new ways. For example, an evaluation could measure the number of times users correct language technologies' outputs or manually enter inputs instead of using ASR technologies. Alternatively, an evaluation could measure how often users code-switch to WME.

We also urge NLP researchers and practitioners to prioritize improving language technologies for speakers of underrepresented languages and language varieties, so they do not feel excluded and need to undertake invisible labor. At a minimum, language technology developers should seek ways to involve such speakers in their design processes.

Limitations

Our findings may be limited in the extent to which they reflect the experiences of African Americans,

and even the experiences of AAL speakers. Although we recruited participants who identified as being AAL speakers and as being African American, we recognize that our participants likely do not reflect the full range of the African American diaspora. In addition, participants were required to reside in the U.S., and they may have needed to have access to smart devices (e.g., smartphones, voice assistants, smart appliances) on which to use language technologies and to have WiFi/cellular connectivity, although neither were a requirement for participating in our study. As a result, our findings may not reflect the experiences of people who are disproportionately impacted by the digital divide.

Ethics Statement

Although our study was IRB approved, we want to foreground one possible ethical consideration. By asking each participant to talk about their experiences as an AAL speaker when using language technologies, we might have reinforced the idea that AAL is not well supported by language technologies, causing participants to feel even more excluded. That said, we hope that by illuminating the relationships between linguistic discrimination and failures of language technologies, our paper motivates NLP researchers and practitioners to invest in improving language technologies for speakers of underrepresented languages and language varieties.

Acknowledgements

We thank the many colleagues who provided feedback on or reviewed this paper, including Johann Diedrick, Julie Kientz, and Daniela Rosner.

References

Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, Jey Han Lau, and Sebastian Ruder. 2022. One country, 700+ languages: Nlp challenges for underrepresented languages and dialects in indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7226–7249. Association for Computational Linguistics.

April Baker-Bell. 2020a. Dismantling anti-black linguistic racism in english language arts classrooms: Toward an anti-racist black language pedagogy. *Theory Into Practice*, 59(1):8–21.

April Baker-Bell. 2020b. *Linguistic Justice: Black Language, Literacy, Identity, and Pedagogy*. Routledge.

John Baugh. 2018. *Linguistics in Pursuit of Justice*. Cambridge University Press.

Kara Becker. 2013. Ethnolect, dialect, and linguistic repertoire in new york city. *Manuscript Submitted for Publication. In Malcah Yaeger-Dror and Lauren Hall-Lew (eds.), New Perspectives on the Concept of Ethnolect: Publication of the American Dialect Society (Pads) Durham: Duke University Press*.

Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. Fairlearn: A toolkit for assessing and improving fairness in ai. *Microsoft, Tech. Rep. MSR-TR-2020-32*.

Su Lin Blodgett, Johnny Wei, and Brendan O'Connor. 2018. Twitter Universal Dependency Parsing for African-American and Mainstream American English. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 1415–1425, Melbourne, Australia.

Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101.

Robin N. Brewer, Christina Harrington, and Courtney Heldreth. 2023. Envisioning equitable speech technologies for black older adults. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 379–388, New York, NY, USA. Association for Computing Machinery.

Svend Brinkmann and Steinar Kvale. 2018. *Doing interviews*, volume 2. Sage.

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).

Kate Crawford. 2017. The trouble with bias. In *Conference on Neural Information Processing Systems, invited speaker*.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Workshop on Abusive Language Online*, pages 25–35, Florence, Italy.

Nicholas Deas, Jessica Grieser, Shana Kleiner, Desmond Patton, Elsbeth Turcan, and Kathleen McKeown. 2023. Evaluation of African American language bias in natural language generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6805–6824, Singapore. Association for Computational Linguistics.

Sunipa Dev, Emily Sheng, Jieyu Zhao, Aubrie Amstutz, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. 2022.

On measures of biases and harms in NLP. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 246–267, Online only. Association for Computational Linguistics.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Lisa J. Green. 2002. *African American English: A Linguistic Introduction*. Cambridge University Press.

Jessica A Grieser. 2022. *The Black side of the river: Race, language, and belonging in Washington, DC*. Georgetown University Press.

Christina N. Harrington, Radhika Garg, Amanda Woodward, and Dimitri Williams. 2022. "It's Kind of Like Code-Switching": Black Older Adults' Experiences with a Voice Assistant for Health Information Seeking. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Camille Harris, Matan Halevy, Ayanna Howard, Amy Bruckman, and Diyi Yang. 2022. Exploring the Role of Grammar and Word Choice in Bias Toward African American English (AAE) in Hate Speech Classification. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 789–798, New York, NY, USA. Association for Computing Machinery.

Hendrik Heuer and Daniel Buschek. 2021. Methods for the design and evaluation of HCI+NLP systems. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 28–33, Online. Association for Computational Linguistics.

Darin G Johnson, Bradley D Mattan, Nelson Flores, Nina Lauharatanahirun, and Emily B Falk. 2022. Social-cognitive and affective antecedents of code switching and the consequences of linguistic racism for black people and people of color. *Affective science*, 3(1):5–13.

Ms Charisse Jones and Kumea Shorter-Gooden. 2009. *Shifting: The double lives of Black women in America*. Harper Collins.

Anjali Kantharuban, Ivan Vulić, and Anna Korhonen. 2023. Quantifying the dialect gap and its correlates across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7226–7245. Association for Computational Linguistics.

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*, 6:285–300.

Katherine D Kinzler. 2020. *How You Say It: Why We Judge Others by the Way They Talk—and the Costs of This Hidden Bias*. HarperCollins.

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.

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

Joshua L Martin and Kelly Elizabeth Wright. 2023. Bias in automatic speech recognition: The case of African American language. *Applied Linguistics*, 44(4):613–630.

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*, 4:169.

Mikel K Ngueajio and Gloria Washington. 2022. Hey ASR System! Why Aren't You More Inclusive? Automatic Speech Recognition Systems' Bias and Proposed Bias Mitigation Techniques. A Literature Review. In *International Conference on Human-Computer Interaction*, pages 421–440. Springer.

Judith S Olson and Wendy A Kellogg. 2014. *Ways of Knowing in HCI*, volume 2. Springer.

John R. Rickford. 1999. *African American Vernacular English: Features, Evolution, Educational Implications*. Wiley-Blackwell.

John R Rickford and Sharese King. 2016. Language and linguistics on trial: Hearing Rachel Jeantel (and other vernacular speakers) in the courtroom and beyond. *Language*, pages 948–988.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *ACL*.

Leda Sarı, Mark Hasegawa-Johnson, and Chang D Yoo. 2021. Counterfactually fair automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3515–3525.

Rachael Tatman and Conner Kasten. 2017. Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions. In *Interspeech*, pages 934–938.

Alicia Beckford Wassik and Anne Curzan. 2004. Addressing Ideologies around African American English. *Journal of English Linguistics*, 32(3):171–185.

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

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Kimi Wenzel, Nitya Devireddy, Cam Davidson, 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. Association for Computing Machinery.

Walt Wolfram and Natalie Schilling. 2015. *American English: Dialects and Variation*, 3 edition. Wiley Blackwell.

Tara J. Yosso. 2005. Whose culture has capital? a critical race theory discussion of community cultural wealth. *Race ethnicity and education*, 8(1):69–91.

Vershawn Ashanti Young and Rusty Barrett. 2018. *Other people's English: Code-meshing, code-switching, and African American literacy*. Parlor Press LLC.

A Appendix

A.1 Semi-Structured Interview Questions

Note that in semi-structured interviews, the actual questions asked by the interviewer may not correspond exactly to the interview protocol, as the interviewer may ask new questions to follow up on or respond to topics that emerge during the interviews.

- In the initial survey you identified as a speaker of a language variety used in some Black and African American Communities and indicated that you typically refer to this as: [variety] Are you ok with us using this term during our conversation today?

Significance of AAL in African American Communities

1. How (and in what contexts) do you currently use [variety] with your community (geographically or culturally)?

2. Can you tell me about any negative experiences or with societal biases you've encountered as an [variety] speaker?

Experiences with Language Technologies

1. Are you familiar with what language technologies are?

2. Can you tell me some types of language technologies you've used or interacted with?

3. Can you tell me about a recent experience you've had with one of these types of language technologies?

Negative Experiences with Language Technologies

1. Have you had negative experiences with language technologies as an [variety] speaker? If yes, can you tell me more about that?

2. What do you see as the connection between the societal biases you discussed in the context of being an [variety] speaker and those negative experiences with using language technologies?

Mitigating Performance Disparities

1. Technology makers have attempted to involve affected community members in mitigating these harms such as (e.g., consulting with users of these technologies, inviting users and people to co-design and develop technologies with engineers).

2. Would you want to be involved? If so, how? If not, why not?

3. Are there other approaches that you might suggest that we didn't discuss?

4. What could you or others in your community get out of collaborating with language technology makers on designing them to be more inclusive?

5. What would an ideal partnership with technologists look like for you? Are there concerns or trade-offs?