Mind the data gap(s): Investigating power in speech and language datasets

Mind the data gap(s): Investigating power in speech and language datasets

Nina Markl
Institute for Language, Cognition and Computation
University of Edinburgh
nina.markl@ed.ac.uk

Abstract

Algorithmic oppression is an urgent and persistent problem in speech and language technologies. Considering power relations embedded in datasets before compiling or using them to train or test speech and language technologies is essential to designing less harmful, more just technologies. This paper presents a reflective exercise to recognise and challenge gaps and the power relations they reveal in speech and language datasets by applying principles of Data Feminism and Design Justice, and building on work on dataset documentation and sociolinguistics.

1 Introduction

Algorithmic systems disproportionately harm marginalised communities by reproducing existing structures of oppression within a society in a process called algorithmic oppression (Hampton, 2021). These harms occur in all contexts where AI is applied to people, including speech and language technologies (SLTs) (Blodgett et al., 2020; Bender et al., 2021). Understanding power relations in the datasets used to train and test SLTs is essential to designing fundamentally more just and less harmful technologies. In this paper, I suggest reflecting on the gaps in the content and documentation of language datasets as a way to guide data compilation (Benjamin, 2021) and the re-use of existing datasets in appropriate contexts (Koch et al., 2021).

The aim of this paper is to contribute to a (long overdue) conversation about power, representation and bias in SLTs (see e.g., Blodgett et al., 2020; Field et al., 2021; Havens et al., 2020). It is grounded in the understanding that (language) technologies are political tools which cannot be "neutral". Unless they are explicitly designed to benefit marginalised communities, they will (re)produce existing structures of oppression and cause harm (Benjamin, 2019; Nee et al., 2021; Field et al., 2021). One way of approaching algorithmic oppression has been to carefully document the datasets used to train and test machine learning systems. Gebru et al. (2021) provide a highly influential documentation framework which can be applied to all AI datasets and Bender and Friedman (2018) introduce an approach to documentation specific to datasets for natural language processing, which I draw on here. This transparency can help to anticipate "predictive bias", a systematic difference in error rates for different groups (Shah et al., 2020), which is one (but not the only) outcome of algorithmic oppression. Detailed documentation is absolutely crucial to not just equitable, but fundamentally *useful* SLTs because it allows practitioners to choose appropriate datasets for a particular task. By definition, documentation is interested in what is *included* in a dataset. To highlight power inequities, it's also useful to think about what is *missing* from a dataset. In SLTs, the exclusion of particular ways of using language (accents, dialects, etc.) can lead to the exclusion of communities. This paper is an invitation to reflect on why these "data gaps" exist, who is harmed by them and how this harm could be prevented. The questions I propose here are not exhaustive or definitive, and addressing them may be difficult in many cases. The point is not to create the "perfect" dataset but to highlight that all (language) datasets involve power relations.

In the context of limiting harm and challenging power, thinking carefully about the appropriateness of any (language) technology in a particular context is fundamental[1]. In some cases, the most effective way to challenge power is to refuse to build the technology or compile the dataset (Baumer and Silberman, 2011; Cifor et al., 2019). Just as technologies are not "neutral", they are also not inevitable. A technological "fix" to a structural social problem will often fall short (Greene, 2021; Broussard,

[1] I'd like to thank an anonymous reviewer for pointing out the omission of this "step" in the original framing of this paper.

Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, pages 1 - 12

2019). Moreover, entirely "unbiased" (in the narrow sense of predictive bias) and "inclusive" language technologies can be at least equally harmful to marginalised communities, as "inclusion" can expose communities to further marginalisation and violence (Hoffmann, 2021). For example, automatic speech recognition systems are used in US prisons to monitor phone calls between incarcerated people and their friends, families and legal support (Asher-Schapiro and Sherfinski, 2021). In this context, "better" or "more accurate" speech recognition based on "more diverse" or "inclusive" speech datasets may make it easier for authorities to harm incarcerated people and their communities. Inclusion in datasets owned by technology corporations or public or governmental institutions can further mean that the "data", i.e. voices of these communities, is no longer owned by or even accessible to them. As a first step in any SLT data compilation process it is therefore crucial to consider and ideally directly involve the affected language communities to understand their own needs and desires with respect to language technology, and to avoid perpetuating a long history of colonial approaches to data and language in which communities, especially in the Global South, are exploited by academic institutions, (neo)colonial states and multinational corporations (Heller and McElhinny; Bird, 2020; Birhane, 2020; Coffey, 2021).

In contexts where we do choose to use or compile a dataset, we need to be aware of how power operates within it. The goal is not just to identify or mitigate biases once a system is ready for deployment, to for example, "retrofit against racism" (Costanza-Chock, 2020, 60). Instead, similarly to Bender and Friedman (2018), I argue that these questions should guide the (dataset) design process. Although it may be too late to change the way the data was compiled when reusing a dataset (Koch et al., 2021), it is still useful to critically reflect on the contents and context of the dataset, to ensure it is appropriate. Since it's impossible to evaluate potential or actual harms of data gaps in isolation, this should be done with a particular deployment context in mind. I consider two examples, not to prove that datasets contain imbalances, but to illustrate the framework: Mozilla's Common Voice English (release 7.0) (Ardila et al., 2020) and the Linguistic Data Consortium's Switchboard-2 (Graff et al., 1998, 1999) used to train and test automatic speech recognition (ASR) systems. I chose these datasets because they were compiled in quite different ways, by different types of institutions, for different purposes and contain different data gaps as a result: CommonVoice is a crowd-sourced speech dataset compiled by Mozilla with the explicit aim to create "diverse" speech datasets for ASR development, while Switchboard-2 is a collection of telephone conversations collected by the Linguistic Data Consortium, an academic institution, to develop speaker recognition systems.

2 Background

2.1 Data, power, feminism and justice

"Data" is always socially constructed and situated within a specific cultural, social and historical context (Havens et al., 2020; Benjamin, 2021; Taffel, 2021; Guyan, 2022). The "compilation" or "curation" of datasets involves complex social processes in which practitioners decide what (and who) to include or exclude and how to label or annotate the "data" (Benjamin, 2021; Paullada et al., 2021). These decisions are both shaped by and in turn reproduce existing power relations within a society.

I use the term "power" to refer to the structural position a particular social group occupies in relations to others. Because these social hierarchies as well as relevant categories or groups within them are socially constructed, they vary depending on the cultural and historical context (see e.g., Saini, 2019, on race). Over the past century, constructs of race, gender and sexuality, (dis)ability, class, age and nationality have been used in a global and many local contexts to secure and uphold the dominant position of white people, in particular those who are cisgender, heterosexual, able-bodied, wealthy, men, and/or from the Global North. Hill Collins (2000 [1990], 227) introduces the concept of the *matrix of domination* to describe "the overall social organization within which intersecting oppressions originate, develop, and are contained". It encompasses social, cultural and legal institutions which uphold the dominant position of some groups, while marginalising others, for example through laws and policies (or their enforcement and application), as well as cultural discourses and ideologies and everyday social interaction (Hill Collins, 2000 [1990], pp 282). By "intersecting oppressions", Hill Collins (2000 [1990]) refers to fact that these categories are not separate or separable, but rather produced by interlocking systems of oppression such as white supremacy and

patriarchy (see also "intersectionality" as coined by Crenshaw, 1989).[2] This complex understanding of power also accounts for the fact that groups who are marginalised by one of those systems, can be privileged by another system and hold power, for example white women (see Lorde, 2017 [1984]).

This paper draws on a feminist perspective on data and power, in particular as articulated by D'Ignazio and Klein (2020). Feminism is not an unproblematic framing. Many feminists and feminisms (past and present) exclude, ignore and/or harm marginalised people of all genders, in particular people of colour, Black people and trans* and non-binary people (Vergès, 2021; Olufemi, 2020; Faye, 2021). In academia and other (neoliberal) institutions the concept of intersectionality is further frequently co-opted and misrepresented in a, ahistorical, "depoliticised" and often explicitly deracialised fashion (Bilge, 2013; Tomlinson, 2013). The invocation of and commitment to "ornamental intersectionality", and notions of "equality", "diversity" and "inclusion" can further serve to symbolically address structural inequalities without in any way redressing them (Bilge, 2013; Hoffmann, 2021). Mindful of both this misuse of radical frameworks to which praxis is central, and the genuine harm that has been perpetrated under the guise of "feminism", I understand "feminist work [as] justice work" (Olufemi, 2020, 5) which seeks to challenge all systems of oppression. It is a way of making sense of the world(s) we live in and of organising (for) world(s) we can and want to flourish in. As such, it is for everyone and (potentially) by everyone who wants to understand and challenge existing power structures.

I build directly on D'Ignazio and Klein's seven principles of "Data Feminism": "examine power", "challenge power", "elevate emotion and embodiment", "rethink binaries and hierarchies", "embrace pluralism", "consider context" and "make labor visibile" (D'Ignazio and Klein, 2020, 17-18). I am also drawing on "Design Justice" as a way of understanding how (technology) design reproduces structural oppression and an approach to reimagining those design processes (see Costanza-Chock, 2020, 23)[3]. The principles of Design Justice focus on using design to empower communities, centering the voices of those who are impacted by (technology) design and working towards sustainable and community-controlled designs.

2.2 Language and power

In the context of SLTs, the "data" is language data, such as text and speech recordings where power relations are extremely salient. (Dominant) discourses about marginalised groups (including harmful stereotypes and hateful rhetoric) are reflected and propagated through language. We therefore need to pay close attention to the way marginalised groups are talked about in language datasets.

Language users harness the variation inherent to language to construct social identities and social meaning (Bucholtz and Hall, 2005). Particular ways of speaking (e.g., accents, dialects) can express specific social meanings and become closely associated with a particular way of being in the world (e.g., a specific subculture or social group) (Eckert, 2008). The accents or dialects spoken by elites become associated with (markers of) prestige, while those used by marginalised groups become associated with (markers of) marginalisation (Rosa and Burdick, 2016; Irvine and Gal, 2000). As a result, *whose language* is included matters not just because of *what* is said, but also, *how* it is said.

3 Power in language datasets

"Challenge power. Data feminism commits to challenging unequal power structures and working toward justice."(D'Ignazio and Klein, 2020, 17)

I use the term "algorithmic oppression" as introduced by Noble (2018) and discussed in depth by Hampton (2021) very deliberately to draw attention to the fact that the "biased" system behaviours we observe, rather than being "bugs" which only require a technical fix, are the (mostly predictable) reproduction of existing structural oppression in machine learning systems. The gaps in data and documentation we identify in datasets are also caused by structural factors. To *challenge power*, therefore specifically means pushing for structural, societal change. Technical fixes, such as "debiasing" word embeddings capturing sexism and racism, don't address the underlying societal context (and sometimes merely hide "bias" (Gonen and Goldberg, 2019)).

What does it mean to "challenge power" when compiling or using datasets then? D'Ignazio and Klein (2020) showcase projects which compile "counterdata" filling (deliberate) gaps. For example

[2] While the term "intersectionality" was coined by Crenshaw, the concept has a longer genealogy in Black feminist thought (Hill Collins, 2000 [1990]; Cooper, 2016).

[3] Design Justice Network: https://designjustice.org/

a 1971 map compiled by the Detroit Geographic Expedition and Institute to highlight the disproportionate rate at which Black children were killed by white drivers (D'Ignazio and Klein, 2020, 49). Another way of challenging power using data is to analyse the way oppression is manifested in data, but importantly (data) feminism also encourages us to go beyond critiques of the world as it currently is to imagining the world as it ought to be. As noted above, sometimes the way to challenge power is refusal: refusal to compile data, refusal to share data or refusal to (re)use data (Cifor et al., 2019). However, when we choose to engage with data(sets), we can challenge power by investigating and highlighting power relations. While this is unlikely to prevent all harm, it allows use to act more carefully and hopefully reduce harm.

I outline three steps in reflecting on power relations reproduced in SLT datasets to guide the compilation or selection of a dataset. The first is to identify gaps in data and documentation and their consequences to analyse power relations. The second involves asking *why* those gaps exist (and persist) given the broader context. The final step is about imagining alternative ways of compiling and using the dataset to create more just, less harmful technologies.

3.1 Who and what is missing?

"Examine power. Data feminism begins by analyzing how power operates in the world." (D'Ignazio and Klein, 2020, 17)

As outlined above, the way broader power structures in society are maintained can be understood through the matrix of domination (Hill Collins, 2000 [1990]). In the context of language technologies, we can ask how these structures are reflected in language datasets. Because linguistic variation (in word choice, in pronunciation, etc) is deeply intertwined with social identity, *who* is included is not just important because of *what* they say, but also *how* they say it. Bender and Friedman (2018) lay out an extensive (and excellent) questionnaire to produce a "data statement". They are particularly interested in *who* the *speakers*, *annotators*, *curators* and *stakeholders* are (for definitions of these terms see Bender and Friedman, 2018).

We can also start by minding the gap(s): both who's not included in the dataset (compilation) and what's not specified in the documentation can be revealing. These gaps provide insights in who or what "doesn't matter" (to the curators, and often, society writ large) (Guyan, 2022), as illustrated by Mimi Onuoha's *Library of missing datasets* (Onuoha, 2016)[4]. Key questions to ask at this juncture concern the language variety and speech situation: Whose voices and whose language varieties are missing? Are included topics centering dominant perspectives and/or harmful discourses to the exclusion of alternatives? Are included genres likely to under- or misrepresent marginalised voices? We also need to question who the stakeholders are and what the curation rationale is: Who benefits from the data collection and who is harmed? Who plans the data collection and who owns the data? Lastly, we need to focus on the annotators and their work: Who categorises and annotates the data and how?

3.2 Who is harmed in what ways?

"Elevate emotion and embodiment. Data feminism teaches us to value multiple forms of knowledge, including the knowledge that comes from people as living, feeling bodies in the world."(D'Ignazio and Klein, 2020, 18)

The power inequities identified in the previous step directly relate to reported or potential harms of a SLTs. Where marginalised speech communities (e.g. speakers of a particular accent or dialect) are under-represented in training data, they might be adversely affected by algorithmic oppression. For example, US English commercial ASR works worse for speakers of African American English (Koenecke et al., 2020; Martin and Tang, 2020) and hate speech detection tools disproportionately flag "obscene" language used in neutral or positive ways by, for example, queer communities (Dias Oliva et al., 2021). In addition to under-representation, there is also potential for misrepresentation: Bender et al. (2021) note that marginalised groups are often misrepresented in text data drawn from the internet (see also Tripodi, 2021; Sun and Peng, 2021), which can lead to the reproduction of harmful stereotypes and dominant ideologies (such as islamophobia), further entrenching their marginalised position (Abid et al., 2021). Who annotates (linguistic) data also matters, as annotators' familiarity with particular accents and dialects as well as their own positionality affects how and how accurately they classify data (Sap et al., 2019). In other words, as Waseem et al. (2021) point out, despite the "disembodied" fram-

[4] https://github.com/MimiOnuoha/missing-datasets

ing of machine learning systems, the embodiment of speakers, annotators and curators involved in dataset compilation (and deployment) matters.

Listening to the concerns and experiences of marginalised communities in the understanding that knowledge is embodied and that emotions are a central way we experience and "know" the world (Hill Collins, 2000 [1990]; Haraway, 1988), can also help us understand the harms of algorithmic oppression. A deployed system could cause representational harms (e.g. reproduction of harmful stereotypes in natural language generation) or allocative harms (e.g. exclusion from social media service based on erroneous "hate speech detection") (Barocas et al., 2019) both of which impact what speakers can do and how they feel. Costanza-Chock (2020, 45) describes some harms of algorithmic oppression as "microagressions", which may be comparatively low-stakes inconveniences but are nevertheless (potentially painful) reminders who something is designed for. Of course, what counts as an "inconvenience" is also highly dependent on positionality: people who find keyboards or touchscreens difficult to use or find writing difficult may rely on ASR tools for many tasks.

3.3 Why are there gaps?

"Consider context. Data feminism asserts that data are not neutral or objective. They are the products of unequal social relations, and this context is essential for conducting accurate, ethical analysis."(D'Ignazio and Klein, 2020, 18)

Once we have identified who and what is excluded from a dataset and what the potential or actual harms of this of those exclusions are, we need to interrogate *why* those decisions were made. Recognising the broader social, historical, and technical context in which a dataset was compiled helps us in exploring potential reasons. We can consider for what purpose the dataset was compiled and whether it meets that purpose, what current use cases are and how it differs from other datasets. Specifically, we can ask *why* particular language varieties, genres, topics, speakers and stakeholders were prioritised, based on how, by whom, where and when the dataset was compiled. We can also question the labels and annotations applied to the dataset. Importantly, even if we find that designers were well-intentioned, or that broader social contexts can "explain" why a dataset contains gaps, that's not an excuse, especially if there are harms.

3.4 Who does the work?

"Make labor visible. The work of data science, like all work in the world, is the work of many hands. Data feminism makes this labor visible so that it can be recognized and valued."(D'Ignazio and Klein, 2020, 18)

This is about the annotators, speakers, curators identified in the previous step. We need to ask how were they: trained, paid, rewarded, acknowledged. Considering how the people involved in compiling a dataset were trained, and who paid for their labour helps us understand the decisions they made (Birhane et al., 2021). Reflecting on much they were paid or how they were acknowledged for their work is not just useful to understand their motivation though, but also a reminder that dataset compilation is (crucial) skilled labour which should be fairly renumerated (Gray and Suri, 2019).

3.5 How could this be different?

The final step of the reflection is one of *imagination*. While this may appear unusual or "untechnical", considering how something could have been built differently or how we would like something to be, is useful because it: a) reminds us that technologies are built by people and that, b) technologies can be built differently.

We can reflect on what an ideal dataset for the given purpose would look like. If we've identified many "data gaps" or "documentation gaps", how would we go about filling them? In the current context, it's helpful to reflect on how the data compilation (including sampling and annotation) could be or could have been done differently. We can broadly draw on two principles of Data Feminism to fill data gaps: rethinking binaries and hierarchies, and embracing pluralism.

3.5.1 Rethink binaries and hierarchies

"Rethink binaries and hierarchies. Data feminism requires us to challenge the gender binary, along with other systems of counting and classification that perpetuate oppression."(D'Ignazio and Klein, 2020, 18)

One way of challenging power in datasets is to question the way both the speakers and their language data is documented and categorised. Categorisation is never "neutral", as both relevant areas of classification and the categories within them are socially constructed (Bowker and Star, 2000). In the context of speakers we need to ask: which broad axes are used to classify them (e.g. "gender")

and what are the subcategories within them (e.g. "non-binary", "female", "male")? These systems of classification are central to the way oppression works because they establish hierarchies, often consisting of binaries, which shape our lives in a million ways. As a result of the way power and identity is (re)produced through language, in many contexts gender, race, ethnicity, social class and education are particularly relevant. How these social categories are operationalised within data documentation matters, and is itself an ideological choice that risks reifying or naturalising a particular frame of a fundamentally harmful way of categorising people. "Boundaries" between socially constructed categories such as "race" or "gender" are furthermore contingent on the historical, social and cultural context (Hanna et al., 2020; Guyan, 2022). Here, documentation gaps may also be intentional: contributors may choose not to disclose certain aspects of their identity or experience and in some contexts legal and/or institutional restrictions may prevent them from being included (Andrus et al., 2021; Bennett and Keyes, 2020; Guyan, 2022; Hoffmann, 2021). However, if this information is missing, it's often impossible to disaggregate the performance of an SLT system for different (sub)populations and account for differences *caused* by oppressive structures we seek to challenge. This leaves us in a complicated (and perhaps uncomfortable) position: missing documentation about contributors and annotations makes it harder to examine and challenge power, *and* existing documentation can reify existing hierarchies and binaries unless we work to contextualise and destabilise them. Similarly, both exclusion *and* inclusion of marginalised communities can expose them to harms depending on the context.

3.5.2 Embrace pluralism

"Embrace pluralism. Data feminism insists that the most complete knowledge comes from synthesizing multiple perspectives, with priority given to local, Indigenous, and experiential ways of knowing."(D'Ignazio and Klein, 2020, 18)

One way of addressing data gaps is to change the way we collect and annotate data. Design Justice principles urge us to centre the voices and needs of marginalised communities in design. Directly and meaningfully involving marginalised communities as co-designers is therefore central to designing equitable technologies. For example, while recruiting students is often convenient and cheap, they have (by definition) a particular educational background, and in the United Kingdom for example, the resulting sample is likely to over-represent young, white, non-disabled middle class English native speakers. Similarly, crowdsourcing via the internet has the potential to be more inclusive, in practise there are still many potential barriers in terms of interface design, access to necessary hardware and software, availability of free time and relevant skills as well as feeling welcome and included within the project. Some of the exclusions are also the result of explicit, established practises. Speakers who report any speech or hearing impairments are commonly excluded from datasets used for speech and language research and technology development (Henner and Robinson, 2021). Second language speakers and multilingual speakers are also routinely excluded.[5]

Embracing pluralism also means thinking about the complications that come with "pluralism". (Language) communities are not monoliths and might well on whether and how their language is represented and used in technology. Incorporating and working with (linguistic) variation in language datasets is important but not trivial.

4 Examples

4.1 Common Voice English

Common Voice English is part of a project to collect open-source crowd-sourced speech corpora for a wide range of languages and as a fairly large dataset is suitable for training current (end-to-end) ASR systems (Ardila et al., 2020). The release of Common Voice English considered here is 7.0, and all documentation analysed here is drawn from the Common Voice website[6] and (where indicated) Ardila et al. (2020), which introduced the corpus.

4.1.1 Who and what is missing?

Q: Whose voices & language varieties are missing?
A: The 2021 release of Common Voice English (7.0) contains 2,015 hours of (validated) speech submitted by over 75,000 speakers some of whom opted to provide some information about their gender and accent (see Figure 1 for full breakdown). There are important gaps in documentation: 51% of recordings are not assigned an accent label. Although Mozilla allows users to choose the label

[5] It is telling that these gaps in speech science and technology research have hardly received comment or critique.

[6] https://commonvoice.mozilla.org/, accessed 17/02/2022

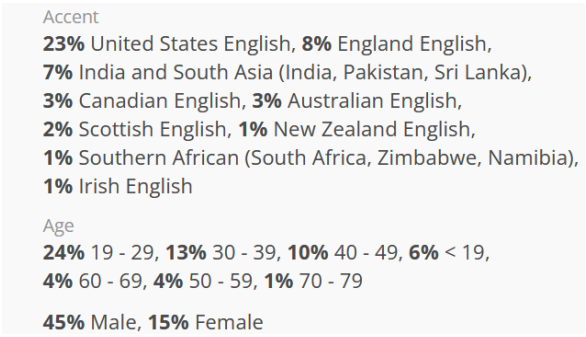

Figure 1: Screenshot of Common Voice English release 7.0 documentation (Accessed 17/02/2022).

"other" as a gender label, the documentation on the website only includes "male" and "female" speakers, and 40% of speakers are unaccounted for. There are also gaps in the data: only 15% of speakers identify as female (45% male), and only 15% are aged under 19 or over 50. While there is a range of varieties of English, only few speakers are from the Global South, with many global Englishes from Africa and Asia missing.

Q: Who plans data compilation & owns the data?

A: The corpus compilation is managed and designed by Mozilla with input from volunteers. Datasets are licensed under CC-0[7], meaning that they can be freely (re)used for any purpose.

Q: Which topics/genres/styles are included? What are likely risks of under- or misrepresentation?

A: Contributors are prompted to read sentences from public domain texts, including from film scripts and Wikipedia[8]. These are likely to reflect Standard English. There is some risk they misrepresent marginalised communities or contain stereotypes which perhaps mitigated by the fact language models used in ASR systems are very constrained because they are only used to decode already recognised phones (or strings of phones) (Bender et al., 2021).

Q: Who benefits from data compilation & who is harmed?

A: The validated datasets are open-source, so they could, in theory at least, benefit anyone who would like to use them for speech technology development. In practise the groups of people who can use open-source datasets, especially to train computationally expensive speech recognition tools is more limited and includes researchers in academia and industry (including at Mozilla). It is unclear that anyone is harmed in the data compilation process as contributors consent to making their recordings and associated information publicly available.

Q: Who annotates the data and how?

A: Speakers are encouraged (but not obligated) to provide their age, gender and choose an accent label from a drop-down list.[9] Recordings are validated by other volunteers via an interface[10]: after listening to the recording they are asked to confirm whether the utterance matches the prompt. Mozilla encourages volunteers to be mindful of accent variation when completing this task[11] but does not take annotator demographics into account.

Q: What are (potential) downstream harms of data gaps and documentation gaps?

A: DeepSpeech trained on an earlier iteration of Common Voice performed worse for African American English speakers, an outcome that could not have been anticipated from the documentation (Martin and Tang, 2020). Speakers of underrepresented varieties have a harder time using the resulting SLTs and report dissatisfaction. Mengesha et al. (2021) document that African American users of a (different) American English ASR tool felt "frustrated", "disappointed" and "angry" at errors which some of them attributed to their own way of speaking.

4.1.2 Consider Context

Q: What is the stated purpose of this dataset? Does it fulfil this purpose?

A: Common Voice is explicitly designed to capture a diverse range of voices, to enable speech and language technology development for minoritised and "low-resource" varieties and languages. In the context of English, this goal is not quite met. Only 49% of the recordings are labelled for accent, which makes it difficult to meaningfully assess the diversity of the corpus. Most of the labelled data represents US English or English English, the two most prestigious and best-resourced varieties.

Q: Why are some varieties and speakers excluded or underrepresented?

A: Mozilla notes on the website that contributions from a wide range of speakers are welcome, including groups usually under-represented in speech

[7] https://creativecommons.org/publicdomain/zero/1.0/

[8] https://github.com/common-voice/common-voice/tree/main/server/data/en

[9] Since 2022 speakers can self-describe their accent (Mozilla Common Voice, 2022; Mozilla Common Voice: Community Playbook)

[10] https://commonvoice.mozilla.org/en/listen

[11] https://commonvoice.mozilla.org/en/criteria

datasets such as second language speakers. However, like other crowdsourced projects, contributors are most likely to be young men[12], and more broadly, speakers from the United States and the United Kingdom. Likely factors shaping these skews include unequal access to technologies and skills privileging (younger) speakers from more affluent backgrounds. Attitudes and ideologies about what "counts" as ("good") "English" may further discourage speakers of minoritised varieties. Members of marginalised communities might also choose not to participate in crowd-sourced projects because they don't *want* (their voices or language) to be included in these datasets and the technologies they power. The problem of documentation gaps such as the fact that 51% of recordings are not associated with an accent label may be the result of the interface design as contributors are not obligated (or particularly strongly encouraged) to provide any information about themselves.

Q: Why are some genres/topics styles excluded or underrepresented?

A: Short snippets of read speech were probably chosen over conversational speech because they do not require expensive and laborious transcription. The use of sentences drawn from Wikipedia favours formal speech styles in standard(ised) English.

Q: How are speakers and annotators trained, paid, rewarded and acknowledged?

A: Speakers and annotators are (anonymous) volunteers. Aside from appearing on a leader board of top contributors, and setting custom goals there are no rewards. There is no required training for annotation or speaking, though volunteers are encouraged to read a short manual.

Q: Who funds the dataset compilation?

A: Work on Common Voice is supported by the Mozilla Foundation, investment from other organisations and grants (Mozilla, 2021b,a).

4.1.3 Re-imagine

Q: How could documentation gaps be filled?

A: Requiring speakers and annotators to provide some basic information about their linguistic background, gender and age could go a long way to fill documentation gaps. While this change could make the dataset more useful, it would also involve "taking" more private data from the contributors and lead some contributors to either not contribute or provide "incorrect" information. Actively encouraging contributors to provide basic information, informing them about the way this data will be used might alleviate some concerns.

Q: How could data gaps be filled?

A: Increasing participation from under-represented groups is likely difficult but could perhaps be achieved with targeted, local campaigns, similar to Wikipedia Edit-a-thons[13] with very clear downstream applications and use-cases designed by or with the relevant language communities.

Q: Do documentation and data gaps constrain appropriate use cases?

A: The documentation gaps mean that it's very difficult to anticipate or evaluate predictive bias using this dataset, as only small portions of it are fully labelled. ASR systems trained on datasets underrepresenting women have been shown to perform worse for female speakers (Garnerin et al., 2021). The data gaps suggest that we should be careful when training ASR systems on Common Voice.

4.2 Switchboard

Subsets of Switchboard-2 are well-established benchmarks for conversational ASR (e.g., Hannun et al., 2014; Tüske et al., 2020)[14]. All information here is drawn from the (more detailed) documentation of Switchboard-2 (Graff et al., 1998, 1999).

4.2.1 Who and what is missing?

Q: Whose voices & language varieties are missing?

A: .The Switchboard-2 (SWB-2) corpus contains (US) English telephone conversations between strangers recorded in the late 1990s. SWB-2 was compiled in two phases, with 657 and 679 speakers respectively (though some appear in both), and a total of a about 8,000 minutes of audio. Most of the SWB-2 speakers were students at US universities, the average age was around 24 years (underrepresenting older people), slightly more than half were female, and most were born and raised in the United States (mostly on the East Coast and the Midwest). Speakers' race or ethnicity is not recorded, the city and state they were raised in serves as a proxy for accent.

[12] Wikipedia has a long-standing an persistent gender gap among contributors: https://en.wikipedia.org/wiki/Gender_bias_on_Wikipedia

[13] https://en.wikipedia.org/wiki/Edit-a-thon

[14] The most popular benchmarks using Switchboard are the Hub5 English evaluation sets (LDC2002S23, LDC2002S09) which include a subset of Switchboard and a subset of CallHome, another LDC corpus, featuring telephone conversations between friends and family members: https://paperswithcode.com/sota/speech-recognition-on-switchboard-hub500

Q: Who plans data compilation & owns the data?
A: The Linguistic Data Consortium (LDC) planned the data compilation, owns and licenses the data.
Q: Which topics/genres/styles are included? What are likely risks of under- or misrepresentation?
A: The speech style is conversational. Topics and specific prompts suggested by LDC include uncontroversial topics (e.g., preferences for food, travel, pop culture, sports) and controversial topics (e.g., gun control, capital punishment, immigration, health care, changing gender roles) apparently designed to spark discussion. The latter could elicit dominant and/or harmful discourses about marginalised groups (e.g. migrants).
Q: Who benefits from data compilation & who is harmed?
A: The LDC and broader academic research community benefited from the compilation of the dataset. It is unclear that anyone was harmed directly by the way the recordings were collected, although some of the topics may have been uncomfortable for some speakers.
Q: Who annotates the data and how?
A: Demographic information about the speakers was collected by members of the research team during recruitment. Only subsets of SWB-1 and SWB-2 were orthographically transcribed (https://catalog.ldc.upenn.edu/LDC2003T02).
Q: What are (potential) downstream harms of data gaps and documentation gaps?
A: Speaker ethnicity or race is not recorded in SWB, but Martin (2021) shows that written African American English (AAE) is under-represented in the transcripts. Similarly, most speakers are young adults and have high levels of education, and almost all of them appear to be native speakers of a variety of US English. In the use of the corpus as a benchmark set this under-representation could cause evaluation bias (Suresh and Guttag, 2021): it's not possible to draw conclusions about the performance of a given system for a diverse range of users (including AAE speakers, second language speakers, older speakers) if they are not represented in the test set.

4.2.2 Consider context

Q: What is the stated purpose of this dataset? Does it fulfil this purpose?
A: SWB-2 (full dataset) was collected to research and develop speaker recognition techniques. Today subsets are used to evaluate conversational ASR systems.
Q: Why are some varieties and speakers excluded or underrepresented?
A: The skew towards young, highly educated, first language speakers of English is probably the result of the sampling method: speakers were primarily recruited via universities and personal networks of researchers.
Q: Why are some genres/topics/styles excluded or underrepresented?
A: Even though the speech style is more conversational and naturalistic than in other corpora (e.g. read speech in TIMIT), it might still be quite formal because the interlocutors don't know each other.
Q: How are speakers and annotators trained, paid, rewarded and acknowledged?
A: Speakers were paid after participation (the documentation does not mention the sum). Recordings were checked for audio quality, transcribed and annotated by members of the research team.
Q: Who funds the dataset compilation?
A: The compilation of Switchboard was funded by the US Department of Defense.

4.2.3 Re-imagine

Q: How could documentation gaps be filled?
A: Including information about speakers' race or ethnicity would have been quite simple (and was done for other LDC corpora, like TIMIT) but could have raised ethical challenges.
Q: How could data gaps be filled?
A: Specifically sampling participants from under-represented groups might have been achieved with a different sampling strategy, for example by advertising more widely or reaching out to particular communities via institutions like schools.
Q: Do documentation and data gaps constrain appropriate use cases?
A: The documentation gaps mean that it's very difficult to anticipate or evaluate predictive bias using this dataset, especially with respect to race.

5 Acknowledgments

This work was supported by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh, School of Informatics and School of Philosophy, Psychology & Language Sciences. I'd like to thank Catherine Lai, Lauren Hall-Lew, Gilly Marchini, Stephen McNulty and three anonymous reviewers for their comments.

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. *CoRR*, abs/2101.05783.

McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. 2021. What we cant measure, we cant understand. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. ACM.

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.

Avi Asher-Schapiro and David Sherfinski. 2021. U.S. prisons are installing AI-powered surveillance to fight crime, documents seen by the Thomson Reuters Foundation show, but critics say privacy rights are being trampled. *Thomson Reuters Foundation News*.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org. `http://www.fairmlbook.org`.

Eric P.S. Baumer and M. Six Silberman. 2011. When the implication is not to design (technology). In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 2271–2274. Association for Computing Machinery.

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

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

Garfield Benjamin. 2021. What we do with data: A performative critique of data 'collection'. *Internet Policy Review*, 10(4).

Ruha Benjamin. 2019. *Race after technology : abolitionist tools for the New Jim Code*. Polity Press, Newark.

Cynthia L. Bennett and Os Keyes. 2020. What is the point of fairness? Disability, AI and the complexity of justice. *SIGACCESS Access. Comput.*, (125).

Sirma Bilge. 2013. INTERSECTIONALITY UNDONE: Saving Intersectionality from Feminist Intersectionality Studies. *Du Bois Review: Social Science Research on Race*, 10(2):405–424.

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th international conference on computational linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2021. The values encoded in machine learning research.

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

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

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

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

M. Cifor, P. Garcia, T.L. Cowan, J. Rault, T. Sutherland, A. Chan, J. Rode, A.L. Hoffmann, N. Salehi, and L. Nakamura. 2019. Feminist Data Manifest-No.

Donavyn Coffey. 2021. Māori are trying to save their language from Big Tech. *Wired*.

Brittney Cooper. 2016. Intersectionality. In Lisa Disch and Mary Hawkesworth, editors, *The Oxford Handbook of Feminist Theory*, volume 1. Oxford University Press.

Sasha Costanza-Chock. 2020. *Design Justice*. MIT Press.

Kimberle Crenshaw. 1989. Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. *University of Chicago Legal Forum*, 1989(1).

Thiago Dias Oliva, Dennys Marcelo Antonialli, and Alessandra Gomes. 2021. Fighting Hate Speech, Silencing Drag Queens? artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online. *Sexuality & Culture*, 25(2):700–732.

Catherine D'Ignazio and Lauren F. Klein. 2020. *Data Feminism*. The MIT Press.

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

Shon Faye. 2021. *The Transgender Issue: An Argument for Justice*. Allen Lane, an imprint of Penguin Books.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A survey of race, racism, and anti-racism in NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1905–1925, Online. Association for Computational Linguistics.

Mahault Garnerin, Solange Rossato, and Laurent Besacier. 2021. Investigating the Impact of Gender Representation in ASR Training Data: A Case Study on Librispeech. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 86–92. Association for Computational Linguistics.

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

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them.

David Graff, Alexandra Canavan, and George Zipperlen. 1998. *Switchboard-2 Phase I*. Linguistic Data Consortium.

David Graff, Kevin Walker, and Alexandra Canavan. 1999. *Switchboard-2 Phase II*. Linguistic Data Consortium.

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

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

Kevin Guyan. 2022. *QUEER DATA: Using Gender, Sex and Sexuality Data for Action.* BLOOMSBURY ACADEMIC.

Lelia Marie Hampton. 2021. Black Feminist Musings on Algorithmic Oppression. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 1–11. ACM.

Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 501–512, New York, NY, USA. Association for Computing Machinery.

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng. 2014. Deep speech: Scaling up end-to-end speech recognition.

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

Lucy Havens, Melissa Terras, Benjamin Bach, and Beatrice Alex. 2020. Situated data, situated systems: A methodology to engage with power relations in natural language processing research. In *Proceedings of the second workshop on gender bias in natural language processing*, pages 107–124, Barcelona, Spain (Online). Association for Computational Linguistics.

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

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

Patricia Hill Collins. 2000 [1990]. *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment*, second edition. Routledge.

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

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

Bernard Koch, Emily Denton, Alex Hanna, and Jacob Foster. 2021. Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 1.

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.

Audre Lorde. 2017 [1984]. Age, Race, Class and Sex. In *Your Silence Will Not Protect You*. Silver Press.

Joshua L. Martin. 2021. Spoken corpora data, automatic speech recognition, and bias against African American Language: The case of habitual 'be'. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, FAccT '21, page 284, New York, NY, USA. Association for Computing Machinery. Number of pages: 1 Place: Virtual Event, Canada.

Joshua L. Martin and Kevin Tang. 2020. Understanding racial disparities in automatic speech recognition: The case of habitual "be". pages 626–630.

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*, 4:725911.

Mozilla. 2021a. Mozilla common voice receives $3.4 million investment to democratize and diversify voice tech in East Africa. Accessed: 24/02/2022.

Mozilla. 2021b. Mozilla partners with NVIDIA to democratize and diversify voice technology. Accessed: 24/02/2022.

Mozilla Common Voice. 2022. How we're making common voice even more linguistically inclusive. Accessed: 24/02/2022.

Mozilla Common Voice: Community Playbook. Community guidance for languages and variants. Accessed: 24/02/2022.

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, New York, NY, USA. Association for Computing Machinery.

Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press.

Lola Olufemi. 2020. *Feminism, Interrupted: Disrupting Power*. Outspoken. Pluto Press.

Mimi Onuoha. 2016. The point of collection. *Data & Society*. Accessed: 24/02/2022.

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

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

Angela Saini. 2019. *Superior: The Return of Race Science*. HarperCollins Publishers.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5248–5264, Online. Association for Computational Linguistics.

Jiao Sun and Nanyun Peng. 2021. Men Are Elected, Women Are Married: Events Gender Bias on Wikipedia. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 350–360. Association for Computational Linguistics.

Harini Suresh and John V. Guttag. 2021. A framework for understanding unintended consequences of machine learning. *CoRR*, abs/1901.10002v4.

Sy Taffel. 2021. Data and oil: Metaphor, materiality and metabolic rifts. *New Media & Society*, 0(0):0.

Barbara Tomlinson. 2013. Colonizing intersectionality: Replicating racial hierarchy in feminist academic arguments. *Social Identities*, 19(2):254–272.

Francesca Tripodi. 2021. Ms. Categorized: Gender, notability, and inequality on Wikipedia. *New Media & Society*, page 14614448211023772.

Zoltán Tüske, George Saon, Kartik Audhkhasi, and Brian Kingsbury. 2020. Single headed attention based sequence-to-sequence model for state-of-the-art results on switchboard.

Françoise Vergès. 2021. *A Decolonial Feminism*. Pluto Press.

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