Language variation and algorithmic bias: understanding algorithmic bias in British English automatic speech recognition

Nina Markl
The University of Edinburgh
Edinburgh, Scotland
nina.markl@ed.ac.uk

ABSTRACT
All language is characterised by variation which language users employ to construct complex social identities and express social meaning. Like other machine learning technologies, speech and language technologies (re)produce structural oppression when they perform worse for marginalised language communities. Using knowledge and theories from sociolinguistics, I explore why commercial automatic speech recognition systems and other language technologies perform significantly worse for already marginalised populations, such as second-language speakers and speakers of stigmatised varieties of English in the British Isles. Situating language technologies within the broader scholarship around algorithmic bias, consider the allocative and representational harms they can cause even (and perhaps especially) in systems which do not exhibit predictive bias, narrowly defined as differential performance between groups. This raises the question whether addressing or “fixing” this “bias” is actually always equivalent to mitigating the harms algorithmic systems can cause, in particular to marginalised communities.

CCS CONCEPTS
- Computing methodologies → Natural language processing; Speech recognition.

KEYWORDS
algorithmic bias, speech and language technologies, language variation, speech recognition

ACM Reference Format:

1 INTRODUCTION
As has been pointed out in recent years in particular by Black, queer and feminist scholars (e.g. [19, 20, 34, 40, 56, 97]), “algorithmic bias”, or as Hampton [56] and Noble [97] put it, “algorithmic oppression” (re)produces existing structures of oppression in a society. Tools frequently discussed in the context of algorithmic oppression often uphold oppressive systems in very direct ways: technologies used in carceral and border systems (“predictive policing and sentencing”, facial recognition) or the (uneven) distribution of housing, capital and services (credit allocation, hiring, education, healthcare) [19, 48, 97, 98]. In this paper, I argue that speech and language technologies (SLTs) are an increasingly important site of algorithmic oppression. They are embedded in high-stakes contexts such as hiring [101] and healthcare [83] and ubiquitous in daily technology use (e.g., voice assistants, language models embedded in web search). Their harms are in some ways more pernicious than those of other machine learning tools, especially where they reinforce existing cultural discourses and ideologies about, in particular, marginalised groups and their ways of using language.

Drawing on knowledge from sociolinguistics, I show that marginalised populations are disproportionately affected because language variation, power and social identity are deeply intertwined. Against this background (and exemplifying this dynamic), I evaluate two British English automatic speech recognition (ASR) systems developed by Google and Amazon. Both systems perform substantially worse on second language speakers of English and speakers of some (stigmatised) regional varieties of British English. I explore potential reasons and consequences of this and other types of SLT bias, as well as ways to detect and mitigate it. Then I turn to the limits of discourses of fairness and bias, and the harms even “unbiased” systems can cause, ending on an open question, pertinent to all discussions around algorithmic oppression – when should we attempt to “fix” biased systems and when should we avoid their use altogether?

2 ALGORITHMIC BIAS AND SPEECH AND LANGUAGE TECHNOLOGIES
2.1 Understanding algorithmic bias
"Algorithmic bias" and “algorithmic oppression” are valuable concepts because they highlight that it is the same systems of oppression and socio-technical contexts, and specifically, the same dominant groups within those contexts, which create and facilitate a wide range of technologies harming the same marginalised communities (albeit in different ways). The overarching frame allows us (or forces us) to recognise that origins and consequences of “biased” sociotechnical systems are systemic [56]. To disentangle the different ways in which these underlying structures “show up”

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

FAccT ’22, June 21–24, 2022, Seoul, Republic of Korea
© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9352-2/22/06...
https://doi.org/10.1145/3531146.3533117

In this paper, I often use the language of “(algorithmic) bias” rather than “(algorithmic) oppression” as this is the term used by most researchers whose work I draw on. Hampton’s critique [56] is, however, a crucial intervention in this field and, as will be evident, informs this paper.
in a sociotechnical system, more fine-grained terminology does, however, help. In recent years, several taxonomies to account for the origins [116], types [109] and consequences [13] of algorithmic bias have been proposed.

Speech and language technologies (SLTs), in particular machine-learning based systems designed to process or analyse text or speech, can harm language communities in different ways. As shown in the case study of British English ASR below, SLTs can exhibit “predictive bias” [109], producing systematically higher error rates for some, usually marginalised, groups (e.g., [39, 80]). Harmful outcomes of machine learning systems can be traced back to a variety of points, including sampling and measurement bias during (training) data generation and curation [116], aggregation and learning biases during model building [61, 66], evaluation biases which miss biased behaviours [116], and inappropriate deployment contexts [116]. The consequences for the people affected depends on the application context and the degree to which individual they rely on it, and include “degraded service”, a higher risk of adverse decisions in high-stake contexts, or representational harms [13]. It is through these harms, that these technologies (intentionally or not) (re)produce structures of oppression as described by [19, 40, 97]: in addition to mirroring the racist, (cis)sexist, ableist and queer-phobic context², they also further entrench and strengthen these structures.

2.2 (Social) meaning and context: inherent limitations to speech and language technologies

Like other machine learning systems, SLTs are often (unhelpfully, see [14]) framed, especially by technology corporations, as “solving” a wide range of (social or communicative) “problems”. Many of those “language problems”, from more abstract tasks like “automatic speech recognition” to concrete applications like “hate speech detection”, are extremely challenging to “solve”. Without detracting from impressive advances in SLTs in recent years, it is imperative that we not lose sight of the limitations inherent to these tools [26]. Sociolinguistics, the study of language in society, is a useful starting point to understand why “solving language” is so difficult.

Language, both in production and perception, is fundamentally social. All parties to any linguistic interaction are situated in a particular social context which they draw on when expressing and interpreting ideas. Indeed, we use language to convey and construct social meaning in addition to those ideas, both as speakers and as listeners [46]. The social context, social meaning (and, arguably all meaning [17]) are generally not available to SLTs [96]. Some tasks, such as hate speech detection, are very difficult for both algorithmic systems and humans because the specific social context (what is said, by whom, to whom) is crucial [96]. Harmful system behaviours in those cases are not (just) the result of insufficient or biased training data, but of the exceptionally difficult, and perhaps inappropriate, task.

Language is also fundamentally characterised by variation. This variation isn’t uniformly distributed across members of a language community, but strongly tied to language users’ identities. Since machine learning models generally improve performance with a higher number of training examples, they tend to perform worse for small (sub)populations in a training data set [116]. Even a system trained on a “perfectly representative” language dataset would be prone to make wrongful predictions for numeric minorities.³ Minoritised and marginalised communities are further often mis- and underrepresented even if they aren’t a numeric minority [16].⁴

2.3 Prior work on algorithmic bias in SLTs

[24] show in their survey of 146 papers on “bias in NLP” that discussions of “bias” are often divorced from social, historical and sociolinguistic context. They also often fail to critically engage with how existing power structures shape who does and does not have access to reliable SLTs, and who gets harmed in what ways as a result [24]. Here, I highlight some of the work on algorithmic bias in SLTs, and, following this critique I return to their origins and harms in 5.

2.3.1 Unequal access to SLTs. SLTs are extremely unevenly distributed both across and within languages. There are over 7000 languages in the world [45], only a small subset of which has been integrated in SLTs. [76] find that 88.38% of the 2679 languages whose typology is described in WALS ([44]) are essentially “no resource languages” (see also [21]). They argue that it is “probably impossible” to create SLTs for these languages which are spoken by more than one billion people globally [76, p. 6284]. The seven languages with the most “resources”, on the other hand, make up only 0.28% of all languages [76]. The framing of this inequality deployed in [76] as a “race” for language resources with “winners”, “left-behinds” and “hopefuls” obscures the (obvious) legacy of colonialism and the effect it has had both on which languages, and more importantly, which ethnic and national groups dominate the world to this day.⁵ It is no accident (and certainly not the result of a fair “win”) that English, Spanish, German, Japanese and French are five of the seven most “well-resourced” languages spoken by 2.5 billion people globally. The upshot of this distribution is that there are many language communities around the world who have no access to SLTs. Furthermore, because many “high-resource” languages for which SLT architectures are initially developed are typologically similar, they might generalise poorly to those which are currently “under-resourced” [76].

2.3.2 Unequal performance of SLTs. But even for the “high-resource” languages of the world, power shapes which language communities can use SLTs successfully, and which ones may even be harmed by them. Here I focus on English, in part because of my own research background, and in part because it is the most well-researched

---

²To name just a few prevalent structures of oppression. Many people are marginalised in multiple ways which are impossible to disentangle.

³As [66] points out, the pervasiveness of “skewed” distributions in the real world, is one of the reasons why careful model development is crucial.

⁴I use the terms ‘minoritised’ and ‘marginalised’ to highlight that these positions are the result of a socio-historical and political process. For example, women and non-binary people make up slightly more than half of the population (globally and in many nations) but are nevertheless marginalised.

⁵The framing of some languages as “left-behind” and fundamentally in need of language resources is particularly problematic. As I discuss below, it may well be that some communities do not want or need these technologies and in any case would like to be actively involved in their creation [22, 23, 65].
context of algorithmic bias in SLTs. However, many of the observations on English also apply to other languages, where SLTs perform much better on a dominant (standard) variety than other varieties and variants used by marginalised communities. This predictive bias can, for example, be seen in sentiment analysis and hate speech detection. [39] show that Google’s sentiment analysis tool Perspective API classifies tweets by popular drag queens as “more toxic” than those by white nationalists. Perspective flags tweets containing reclaimed slurs like gay and queer and “obscene” language in neutral, positive or non-offensive contexts as “toxic” (see also [42]), but does not account for the fact that “innocuous” words can be used in ways that are deeply hateful. In other words, it doesn’t capture the social context of the “obscene” language.  

Of course, which (or whose) language is considered “obscene” is itself an ideological choice imposed by the hearer [112]. Large language models are prone to reproduce structural oppression in a very direct way by “parroting” biases in the training data [16], for example islamophobic content [1, 43, 86]. Similar problems also exist in machine translation [107] and word embeddings [27] where gender bias proves particularly persistent. Gender neutral nouns or pronouns are often translated reflecting stereotypes or are simply ungrammatical [38, 107]. Machine translation also introduces stylistic bias, where translated text “sound[s] older and more male” than the original [68]. Recent work on US English automatic speech recognition systems show substantial performance differences between Mainstream US English and varieties used by marginalised communities such as African American English (AAE) [80, 90, 126]. [80] find that commercial ASR systems by Apple, IBM, Google, Microsoft and Amazon produce a much higher rate of errors for Black speakers of AAE than comparable White speakers of Californian English. Notably, they also find that error rates were influenced by both gender and race, with particularly high rates for Black men, who tend to use a very high rate of “non-standard” linguistic features in the recordings used in the study (sourced from CORAAL [77]) [80]. This highlights the need for not just disaggregated approaches to SLT evaluation, but specifically intersectional ones, which recognise that interlocking systems of oppression (such as race and gender) cannot be considered separately (as conceptualised in Black feminist thought [36, 62] and applied to other domains of machine learning evaluation [28, 59, 75]). Other work has found predictive bias for regional varieties of English such as Indian English [91], Scottish English and Southern US English [119, 120] (as compared to Mainstream US English). I add to this literature by considering predictive bias in British English commercial ASR systems as it affects first and second language speakers of English.

3 PATTERNS OF LANGUAGE VARIATION AS PROXIES FOR SOCIAL IDENTITIES

As noted above, language is both inherently social and inherently characterised by variation. This variation may appear random or free when we first encounter it (for example when we enter a new language community). However, as [129] put in a very influential formulation, language variation (and language change) is characterised by “orderly heterogeneity”. That is, patterns of language variation are not random, but are highly structured both in individuals and in communities and they can further be used to construct social identities in interaction [47]. As a result, particular linguistic features (or particular combinations of them) can be proxies for social identities. Worse SLT performance for particular language varieties and linguistic features thus often translates to worse performance for particular (usually marginalised) people. In the following section I explain the relationship between language variation and identity before outlining some work on variation in British English, in particular how it relates to power and discrimination.

3.1 Beliefs about language are beliefs about speakers

Language and language variation are always situated in a larger social context. All speaking, writing, signing and listening originates from somewhere. Sociolinguists have long been interested in how particular ways of using language can become associated with specific social identities and positionalities until they become indexical of them (i.e. until they point to a particular identity) [47, 73]. In short, as people using language we construct beliefs about language to make sense of the (arbitrary) correlations between particular linguistic forms and the people who use them. Put more precisely, we “locate linguistic phenomena as part of, and evidence for, what [we] believe to be systematic behavioral, aesthetic, affective and moral contrasts among the social groups indexed” [72, p 37]. These ideologies are used to justify and re-entrench particular power structures and construct notions of normativity, markedsness, difference and similarity between social groups [35]. Like other ideologies, they can become deeply embedded in our understanding of the world and shape how we produce and interpret language (variation).

Language ideologies can surface in “attitudes about language”, often framed as “apolitical”, aesthetic preferences for one form over another. But these attitudes about language are almost always reflective of attitudes about the speakers who use them. This is evident in the fact that the same linguistic feature is often interpreted differently depending on who produced it. For example, a creaky voice, a phonation type commonly also known as “vocal fry”, is among English speakers, much more stigmatised and pathologised in young women’s speech than men’s [5, 31]. The terms used to evaluate the feature are also evaluations of the women who use them: “annoying”, “grating”, “too much to bear” [31, 53]. Similarly, linguistic features common in some varieties of British English, such as “glottal replacement” of /t/ in words like butter or Scotland are stigmatised when used by working class speakers in formal contexts, but interpreted as signalling authenticity and solidarity.
when used by upper-class speakers (e.g. politicians) in those very same contexts [79, 111]. Listeners’ judgements of speakers (e.g. attractiveness, trustworthiness, friendliness) are also influenced by their perceived race, gender and social class background [9, 41]. These attitudes also have structural implications (see also [35]). [5] ask listeners to rate speakers with and without vocal fry according to their “hireability”, and find that those without creaky voice are preferred. This is just one example among many culturally-specific language ideologies around “professional”, “educated” and “articulate” speech [11, 87]. In anglophone settings, hiring committees disprefer second language speakers [67, 121] who have also been found to be perceived as “less credible” [84] than first language speakers. Just like algorithmic oppression, language ideologies are not just underpinned by or reflective of structural oppression, but also serve to secure it [104]. It is the language used by powerful social groups in a given societal context (e.g. White, upper and middle class, men) that becomes the “prestigious” or “right” way to speak. [5] conclude that women should avoid creaky voice to avoid discrimination and similar advice is often given to anyone who doesn’t speak the “standard variety” [35]. I strongly reject this conclusion - it is listeners (and hiring committees) and language technologies should resist sexist (language) attitudes [31].

3.2 Linguistic variation in the British Isles

The British Isles encompass a lot of linguistic diversity. In addition to English, there are many minoritised languages, including Scottish Gaelic, Scots, Irish, Welsh, Manx, Polish, Punjabi and Urdu, which have different levels of legal recognition within the United Kingdom and Ireland [45]. English in the British Isles is also characterised by significant variation, conditioned both by region and social class (see e.g. [51, 70, 130]). This variation is apparent both in dialectal variation (broadly: variation in syntax, morphology and lexicon) and accent variation (variation in pronunciation). Linguists tend to define regional accent or dialect regions along “linguistic borders” (so-called “isoglosses”) where two (or more) different ways of expressing the same concept or structure meet. These different pronunciations, words or syntactic structures are often rooted in the distinct historical developments of English in different regions. Especially in the context of accents, these differences are not isolated to individual words, but tend to affect the entire “inventory” of sounds in a particular Accent (the “phonology”). For example, accents in the South of Britain distinguish between the vowel in words like can and the vowel in words like can’t, while those in the North generally do not [70]. In addition to these geographical differences in the presence and distribution of particular sounds, speakers also vary in their language use depending on style, context and social class.

In the British Isles (in particular in the UK), the classic example of a highly prestigious accent is Received Pronunciation (RP) [12]. RP, also colloquially referred to as “the Queen’s English”, is “supra-local”: rather than being interpreted as an index of the speaker’s geographical origin or identity, it is interpreted as indicative of their social (class) and educational background [3, 49]. It is spoken by a small group of people and was historically particularly widely used in British media and in elite spaces (private schools, politics, aristocracy) [3]. As [3] highlights, the association between RP and upper class status is very strong, and has been reinforced over centuries through prescriptive teaching (inside and outside classrooms) and popular media. Crucially, other “native” accents especially those associated with working class speakers both in urban and rural areas of the North and northeast of England, the Scottish central belt, Wales, and London continue to be stigmatised in many elite spaces and rated as “less prestigious” and “less pleasant” [110]. Recent research shows that attitudes towards (some) second language accents have improved, or, at the very least, that increased awareness of the negative effects and arbitrary nature of linguistic discrimination lead study respondents to suppress negative judgments [103]. Nevertheless, accent discrimination and open prejudice against speakers of particular accents (especially second-language speakers) or people who use particular linguistic features appears more socially acceptable than other forms of discrimination. Recent work shows that accent discrimination still plays a role high-prestige hiring contexts such as corporate law, although not all regional accents are equally stigmatised [29, 85]. Accent bias has also been documented in teacher training and schools in the UK, affecting both first and second language speakers of English [11, 37].

4 INVESTIGATING ALGORITHMIC BIAS IN BRITISH ENGLISH ASR SYSTEMS

To add to the understanding of predictive bias in commercial automatic speech recognition systems, I tested the off-the-shelf systems by Google (Google Speech-to-Text) and Amazon (Amazon Transcribe) using two corpora of read speech: the Speech Accent Archive (SAA) [128] and Intonational Variation in English (IViE) [54]. The subset of SAA contains a wide range of first and second language speakers of English, and is useful to illustrate that second language accents have improved, or, at the very least, that second language speakers are disadvantaged in the context of ASR systems as compared to first language speakers. IViE allows for a detailed analysis of how speakers of different varieties of British English are impacted by algorithmic bias in ASR.

4.1 Experimental setup

The Speech Accent Archive is a database of short English language speech samples. Each entry consists of a recording of a read elicitation passage which contains most sounds of English, some demographic information about the speaker (binary gender, age, native language and other languages, birthplace, current place of residence, age and mode of acquisition of English), a detailed phonetic transcript and some linguistic analysis. For this experiment, I initially chose a subset of 495 recordings provided by first and second

---

I’d like to thank an anonymous reviewer for their generous comments regarding the role of the “listening subject” in the context of language technology.  
[12] Polish, Punjabi and Urdu are the most common “non-indigenous” languages in the UK, though the list of languages spoken by residents of the UK is, of course, very long and ever-changing.  
[14] Data, code and analysis available at https://github.com/ninamarkl/FAccT22_ASRBias  
[16] See also the Accentism Project which collects personal stories about this experience: https://accentism.org/
language speakers of English (as self-defined by each speaker). Both
groups are internally heterogeneous: the first language speakers
are from a range of regions in the UK, and all 430 second language
speakers speak one of ten randomly chosen languages as a first
language (Arabic, French, German, Hindi, Italian, Mandarin, Por-
tuguese, Spanish, Thai or Urdu) and vary in how, and for how long,
they have been learning English. To ensure that any differences
weren’t due to systematic differences in recording quality, signal
to noise ratio was measured using Praat and recordings with a
measure higher than 50dB were excluded from the subsequent
statistical analysis. 445 recordings were retained.

The IViE corpus was collected around 2000 to study intona-
tional variation in the British Isles. It contains audio recordings
from 102 adolescent speakers from 9 cities in the British Isles: Lon-
don, Dublin, Cambridge, Liverpool, Leeds, Bradford, Newcastle,
Cardiff and Belfast. The IViE corpus does not contain any informa-
tion about the speakers, aside from their age (16), binary gender,
city and the fact that the speakers from Cardiff and Bradford are
bilingual (Welsh and Punjabi, respectively) and the speakers from
London are of Caribbean descent. I chose recordings of the speak-
ers reading the first two paragraphs of a longer retelling of the
fairy tale Cinderella for analysis (about one minute per speaker).
All reference transcripts were validated and, where speakers made
speech or reading errors, adjusted by me.

All audio recordings were converted to 16kHz FLAC files, up-
loaded to Google Cloud and Amazon S3 Storage and processed
using their Python APIs. Both corpora were processed with the de-
fault models for British English (‘en-GB’). The generated transcripts
were evaluated against the reference transcripts using sclite from
the SCTK toolkit. Further analysis of the evaluation outputs was
conducted in R and Matlab to compare performance on speaker
subgroups within each corpus.

4.2 Results
I employ a mixed methods approach to analysing the experi-
tmental results. To quantify the extent of predictive bias experienced
by second language speakers of English and speakers of different
regional varieties of English, I report word error rate (WER), a stan-
dard metric in ASR evaluation for both experiments. I then apply
a qualitative error analysis to the results of the IViE experiment to
explore the effect of phonetic variation on word error rates.

4.3 Quantitative results
WER is an edit-distance metric presenting the number of errors
(deletions, insertions, substitutions) in an automatic transcript relat-
tive to the number of correct words in a reference transcript. It is
usually computed over an entire test set, often a well-established
benchmark. This aggregated approach risks obscuring systematic
differences between subsets of the test set. To avoid this, I analyse
the WER using multiple linear regression models which include
factors such as variety, gender, speech rate and age.

4.3.1 SAA: L1 vs L2 speakers. While SAA also contains information
about the speakers’ first language and when they started learning
English, I decided to focus specifically on sex, age, speech rate
and L1/L2 status. Error rates varied greatly by L1 (and individual)
for both systems, but they were lowest for L1 speakers of English.
Age of second language acquisition, as well as phonological charac-
teristics of a first language can influence speakers’ accents in many
ways which may also impact ASR error rates. However, for the
purposes of this paper, I am more interested in highlighting that a
wide range speakers whose (potentially only) common characteristic
is that they’re not “native speakers” of English, likely encounter
problems when using these ASR systems. While this category is
itself problematic, it is also how many speakers are perceived (and
judged by, in particular, “native speakers”.

For both systems, linear regression models show that word error
rates are significantly higher for L2 speakers than L1 speakers.
Categorical predictors (variety: L1/L2 English, sex: male/female)
deviance coded and numeric predictors (age, speech rate in
syllables per second) are scaled and centered. There is a significant
main effect for variety at p<0.05 for both systems (Table 1a and
Table 1b). Sex is not a significant factor for either model, and adding
an interaction term of sex and variety does not improve model fit.
Age is a significant factor for Google, with higher error rates for
older speakers. Speech rate is a significant factor for Amazon, with
higher speech rates corresponding to lower WER.

For speakers from Newcastle, Liverpool, Belfast, and Bradford,
Amazon produces error rates which are significantly higher than
those for speakers from Cambridge (p<0.05). There is also a signifi-
cant main effect of gender, whereby recordings by female speakers
show significantly lower error rates (p<0.05). Adding an interaction
term for gender and variety did not improve model fit.

4.3.2 IViE: Variation with British L1 varieties. To investigate the
impact of accent variation, I chose the variety with the lowest error
rate for each system as the reference levels in linear regression mod-
els. Amazon performs best on recordings from Cambridge, while
Google performs best on those from London. Categorical predic-
tors (gender: male/female) deviance coded and numeric predictors
(speech rate in syllables per second) scaled and centered.

For speakers from Newcastle, Liverpool, Belfast, and Bradford,
Amazon produces error rates which are significantly higher than
those for speakers from Cambridge (p<0.05). There is also a signifi-
cant main effect of gender, whereby recordings by female speakers
show significantly lower error rates (p<0.05). Adding an interaction
term for gender and variety did not improve model fit. Compared
to speakers from London, Google only performs significantly worse
for speakers from Belfast (p<0.05) (see Fig 1b). There is an interac-
tion effect between variety and gender (which improves model fit):

39For each “native language subgroup”, I selected up to 70 recordings: recordings
are numbered consecutively, so the dataset contains, e.g., files “arabic1” to “arabic70”
and “urdu1” to “urdu16” as there are a total of 16 urdu speakers in the sample. See
https://github.com/ninamarkl/FAccT22_ASRBias for further details.
40https://github.com/unmistgov/SCTK
41While this type of analysis would also be appropriate and interesting for the SAA
dataset, I focus on IViE here as the speakers in SAA show a lot of individual variation,
the analysis of which is outwith the scope of this paper.
42Note that despite the name, word error rate can be larger than 1 and is conventionally
represented in percentages (i.e. WER * 100)
error rates are significantly higher for women from Belfast, Cardiff and Newcastle (see also Fig 1b).

4.4 Qualitative results: applying context-sensitive evaluation

Quantitative evaluation fails to capture the context of errors. WER (as computed above) does not distinguish between different error types (insertion, deletion or substitutions), linguistic contexts (word class, phrase position) or different “triggers” for errors (phonetic variation, speech errors, unusual phrases). WER therefore obscures both the origins and consequences of an error. While architectures vary [93, 131], most speech recognition systems make use of an acoustic model, which contains representations of speech sounds, a dictionary mapping sequences of sounds to words, and a language model, which is used to decode words into longer sequences. Because errors can be the result of a mismatch between the training and test data, the errors we can observe here can be the consequence of under-representation (of a particular pronunciation, turn of phrase or word) in any of those components. To understand origins and impacts of ASR errors, we can qualitatively analyse these errors [88]. The ability to pinpoint which linguistic features “trigger” errors with the help of sociolinguistic expertise could be very useful in developing more robust technologies [126].

4.4.1 Error types. WER considers three types of errors: “substitution errors” where a word is substituted with a wrong transcription, “insertion errors” where the ASR system inserts a word not present in the speech signal, and “deletion errors” where the ASR system fails to transcribe a word. The two systems differ in the distribution of those errors: substitutions are the most common type for both systems while insertion errors are rare, but Google has a much higher deletion rate than Amazon. These patterns are consistent across all speaker groups, which perhaps suggests different model settings. Systematic differences in error type could be problematic as they have distinct impacts on the transcripts. A very high deletion rate can render a transcript useless, in particular as they sometimes appear to cause knock-on effects (see also [89]). Substitution errors can vary in impact: substitutions tend to be phonetically similar (but semantically unrelated) or morphologically related (but not necessarily phonetically similar).

4.4.2 Errors related to phonetic variation. Analysing substitution errors more closely is useful to understand origins and impacts of the errors. We would expect the system to be most accurate for the variety the acoustic model was trained on (or the variety best represented in the training dataset). In addition to simply looking at WER by variety (recall: lowest WER for Cambridge & London), comparing what a speaker actually said to what the system transcribed can also provide clues to varieties the system was trained on. For example, for several of the Belfast speakers’ the word hair, pronounced by most of them as /həːr/ is mis-transcribed as her.25 In RP, the sequence /həːr(r)/ is indeed most likely her, while the actual target hair is produced as something like /həːj/. Transcribing /həːr(r)/ as her is therefore entirely expected if the system was trained on RP. This is just one small example of a systematic difference in the phonology of different varieties, which can lead to predictive bias. This kind of approach could be applied in evaluation on a larger scale by systematically tracing correlations between error rates and sociolinguistic variation [126].

4.4.3 Morphological and syntactic errors. For both systems, substitutions are often morphologically related forms, differing from the target only in number or tense. Sometimes these substitutions are phonetically similar. For example, in more than half of the Google transcripts lived in the phrase Cinders lived with her mother is substituted with live. These errors may reflect differences between connected speech (and in particular, faster speech) and more careful speech a system was trained on. However, some substitutions are quite phonetically distinct. In 47 (Amazon)/41 (Google) of the 102 IViE recordings, the word would in The ball would be held is replaced with will. This error might be introduced by the language model used in the ASR system (for example, because it was trained on text containing mostly present tense verb forms).

5 DISCUSSION

5.1 British English commercial ASR & linguistic hierarchies

The quantitative analysis shows that the performance of Amazon Transcribe and Google Speech-to-Text differs broadly by speaker group, with higher error rates for second language speakers of English, male speakers (Amazon), and speakers of some varieties spoken in the North and Northeast of England (Newcastle, Liverpool, Bradford) and Northern Ireland (Belfast) as compared to L1 speakers, women, and those from the South of England. These differences are particularly notable considering the nature of the

25Belfast English (like some other varieties of British English) collapses the distinction between the vowels in hair and her [130].
Table 2: Word error rates differ by variety. For each system, the variety with the lowest error rate was chosen as the reference level (Amazon: Cambridge, Google: London). Amazon: sig. (p<0.05) worse: Bradford, Liverpool, Newcastle, Belfast; better for women. Google: sig. worse: Belfast.

<table>
<thead>
<tr>
<th></th>
<th>Variable</th>
<th>Estimate</th>
<th>Std Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td></td>
<td>11.21</td>
<td>1.56</td>
<td>7.19</td>
</tr>
<tr>
<td></td>
<td>Cardiff Welsh</td>
<td>4.45</td>
<td>2.50</td>
<td>1.78</td>
</tr>
<tr>
<td></td>
<td>Bradford Punjabi</td>
<td>8.84</td>
<td>2.18</td>
<td>4.05</td>
</tr>
<tr>
<td></td>
<td>Leeds</td>
<td>1.43</td>
<td>2.23</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>Liverpool</td>
<td>6.00</td>
<td>2.20</td>
<td>2.72</td>
</tr>
<tr>
<td></td>
<td>London West Indian</td>
<td>4.80</td>
<td>2.39</td>
<td>2.01</td>
</tr>
<tr>
<td></td>
<td>Newcastle</td>
<td>5.25</td>
<td>2.19</td>
<td>2.40</td>
</tr>
<tr>
<td></td>
<td>Belfast</td>
<td>9.03</td>
<td>2.19</td>
<td>4.13</td>
</tr>
<tr>
<td></td>
<td>Dublin</td>
<td>1.18</td>
<td>2.23</td>
<td>0.53</td>
</tr>
<tr>
<td></td>
<td>Gender: female</td>
<td>-2.46</td>
<td>1.11</td>
<td>-2.42</td>
</tr>
<tr>
<td></td>
<td>Speech rate</td>
<td>0.22</td>
<td>0.63</td>
<td>0.35</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Variable</th>
<th>Estimate</th>
<th>Std Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(Intercept)</td>
<td>11.21</td>
<td>1.56</td>
<td>7.19</td>
</tr>
<tr>
<td></td>
<td>Cambridge</td>
<td>-0.07</td>
<td>3.38</td>
<td>-0.02</td>
</tr>
<tr>
<td></td>
<td>Cardiff Welsh</td>
<td>-0.12</td>
<td>3.45</td>
<td>-0.03</td>
</tr>
<tr>
<td></td>
<td>Bradford Punjabi</td>
<td>0.48</td>
<td>3.41</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>Leeds</td>
<td>-5.06</td>
<td>3.41</td>
<td>-1.49</td>
</tr>
<tr>
<td></td>
<td>Liverpool</td>
<td>2.82</td>
<td>3.22</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>Newcastle</td>
<td>-0.34</td>
<td>3.24</td>
<td>-0.11</td>
</tr>
<tr>
<td></td>
<td>Belfast</td>
<td>9.11</td>
<td>3.25</td>
<td>2.80</td>
</tr>
<tr>
<td></td>
<td>Dublin</td>
<td>-1.68</td>
<td>3.37</td>
<td>-0.50</td>
</tr>
<tr>
<td></td>
<td>Gender: female</td>
<td>-8.72</td>
<td>4.38</td>
<td>-1.99</td>
</tr>
<tr>
<td></td>
<td>Speech rate</td>
<td>6.41</td>
<td>0.96</td>
<td>6.67</td>
</tr>
<tr>
<td></td>
<td>Cambridge*female</td>
<td>8.31</td>
<td>6.08</td>
<td>1.37</td>
</tr>
<tr>
<td></td>
<td>Cardiff*female</td>
<td>15.55</td>
<td>6.83</td>
<td>2.28</td>
</tr>
<tr>
<td></td>
<td>Bradford*female</td>
<td>-2.21</td>
<td>6.24</td>
<td>-0.35</td>
</tr>
<tr>
<td></td>
<td>Leeds*female</td>
<td>9.68</td>
<td>6.31</td>
<td>1.53</td>
</tr>
<tr>
<td></td>
<td>Liverpool*female</td>
<td>10.48</td>
<td>6.12</td>
<td>1.71</td>
</tr>
<tr>
<td></td>
<td>Newcastle*female</td>
<td>12.73</td>
<td>6.09</td>
<td>2.09</td>
</tr>
<tr>
<td></td>
<td>Belfast*female</td>
<td>28.34</td>
<td>6.09</td>
<td>4.65</td>
</tr>
<tr>
<td></td>
<td>Dublin*female</td>
<td>1.35</td>
<td>6.28</td>
<td>0.22</td>
</tr>
</tbody>
</table>

5.2 Specific origins of “bias” in SLTs

My finding that Amazon Transcribe and Google Text-to-Speech perform best for Southern British varieties of English (and L1 speakers), suggests that some regional varieties of British English, especially Northern Ireland and the North of England are under-represented in the training datasets for these systems. Because the lexical content of the recordings was tightly constrained, these biases most likely originate in the training data for the acoustic models. Similarly, in the US context, [80] conclude that the higher error rates for African American English speakers are due to under-representation of AAEE in the acoustic models [90] further suggest that some AAEE constructions are also under-represented in the language models used by commercial and open-source ASR. They find that Google Speech-to-Text and Mozilla’s DeepSpeech produce significantly higher error rates in the vicinity of “habitual be” [90], a feature absent in Mainstream US English, than other uses of “be” [90]. The way training datasets are sourced thus warrants particular attention. Like many commercial machine learning systems, commercial ASR systems are trained on proprietary datasets. Documentation 28 of the voice user interfaces of both Amazon (Alexa) and Google (Google Assistant) suggest that voice data collected from users is part of this training data. Even setting aside any privacy concerns, this reliance on customer data is problematic because customers of large technology corporations who already use an ASR tool are unlikely to be representative of any given language community. According to the British Office for National Statistics, 35% of adults in Great Britain used voice user interfaces in 2020 [50]. The survey only considered variation in age and sex, with younger age groups

26Google shows the opposite pattern for some varieties (Belfast, Cardiff, Newcastle) - perhaps as a result of different error types: one recording by a Belfast woman has a WEB of 9% for Amazon but 60% for Google with 86% deletion errors.

27E.g. AAEE “I be at my office at 7.30” is equivalent to Mainstream US English “I am usually at my office at 7.30”, see also [55].

American English was not being developed because “Apple products (and no difference by sex), but similar studies on smartphone and home broadband use from the United States suggest that income is also an important factor [30]. While 96% of households in the UK had broadband access in 2020, speeds vary by region [10], which, among many other impacts, means that users in some geographical regions (e.g. Wales and the Scottish highlands) are probably less likely to make use of devices reliant on cloud-computing (such as ASR systems). This points to a more fundamental problem with commercial SLTs and predictive bias: their market-oriented design. Smaller or marginalised language communities are less likely to be considered valuable markets for technology companies [see also 82]. [19] presents a striking quote by a former engineer working on Apple’s ASR systems who was told by a manager that African American English was not being developed because “Apple products are for the premium market”. This especially egregious example of a (racist) language ideology linking all speakers of a particular language variety to a particular social and economic position, implying that AAE speakers are not part of the “premium market” emphasises that corporations design technologies with particular users in mind. Communities who aren’t perceived as “desirable” markets are less likely to be catered to. In the British context, this means that regional varieties, especially relatively stigmatised ones, are not developed for. This is probably also in part due to the “standard language ideology” [92], the belief that there’s an “ideal” or “standardised” variety of the language which is often the most prestigious, or “canonical” form of the language. Importantly, as this ideology becomes part of a “common sense” understanding of the world, so does the belief that all speakers of the language should (strive to) be able to speak the “standard variety”, as failing to do so is simply speaking “incorrectly”. In the context of SLTs, a consequence of this standard language ideology is that catering for different accents or dialects of the same language is considered less important as speakers are expected to be able to switch to the “standard”. While open-source crowdsourced datasets like Mozilla’s CommonVoice [7] could in theory be more representative, in practice, they are also unbalanced. This reflects broader issues with crowd- or community-sourced language data (e.g., text) used to train machine learning systems. A well-researched example to illustrate this point is Wikipedia, the text of which is frequently used to train SLTs. Wikipedia is a microcosm of the larger problem [16] identify where marginalised groups are both under- and potentially misrepresented in the text used to train large language models and other SLTs. For example, text by and about women and non-binary
people is both under-represented on English language Wikipedia [122], and qualitatively different from entries about men [115, 125]. This leads on the one hand, to a very particular skewed perception of the world by readers, and, much less obviously to people editing or reading Wikipedia, to, for example, “co-occurrence bias” (e.g. the word *family* co-occurs more often with *woman* than *man*) [61], which is reproduced in machine translation systems [107], word embeddings [108], and other tools relying on large language models [97]. This “gender gap” on Wikipedia is an interesting example because it highlights how large structural factors and platform-specific policies and contexts work together to, to some extent inadvertently, create deeply skewed collections of knowledge and text which are then used to build SLTs. The majority of Wikipedia editors are men [6, 58] due to (perceived) skills gaps [58], lack of leisure time and perceptions of the existing (misogynistic or antagonistic) culture on Wikipedia. A seemingly common sense but in practice pernicious Wikipedia policy further favours biographies linking to existing Wikipedia articles [2, 122], making it difficult to add articles about notable people from marginalised communities who have been excluded from mainstream histories. [122] shows that, in addition to only making up 19% of all 1.5 million biographies about inventors, scientists and writers on English language Wikipedia, biographies about notable women are much more likely to be flagged for deletion or disputed on Wikipedia than those about men. The way women are written about is also different with more focus on personal relationships and a high rate of gendered terms (e.g. *feminine*, *woman* etc) [115, 125]. In short, women and non-binary people are under- and misrepresented on Wikipedia, and by extension in the “training data” for many SLTs.

Data bias can also be introduced to training data sets through (usually crowdsourced) annotation. Predictive bias in hate speech detection [39, 42, 106] has been shown to be affected by both data bias and annotation bias. [106] report a higher error rate (erroneous “toxic” label) for African American English phrases than Mainstream US English. They also show that White annotators are more likely than Black annotators to label non-toxic AAE tweets as a “toxic” [106]. In the US context, White speakers of “Standard American English” often associate African American English with racist stereotypes [87], while “raciolinguistic” ideologies position African American English as inferior to (white) Mainstream US English [105]. [39] and [42] conclude that the disproportionate associations between “obscene” words and reclaimed slurs and “toxic” labels in training data give rise to the predictive bias they observe.

### 5.3 Concrete harms of biased SLT

Worse performance on particular language varieties is often equivalent to worse performance for particular communities. Because SLTs tend to be trained on “prestigious” varieties, such as Standard English and RP, these communities are likely to be already marginalised. In this way SLTs further cement the power of “high-status” varieties (and speakers) – and contribute to the devaluation of all other varieties (and speakers). Availability of SLTs in powerful varieties could even accelerate language shift in some domains, such as computing and digital media [63].

Allocative harms [116], or “adverse decisions” [13] of biased SLTs are understudied. But SLTs are now commonly embedded in high stakes contexts where bias could prevent a language community from being allocated resources. Automatic speech recognition systems are part of complex algorithmic systems used to automatically rank job application videos [101]. Predictive bias of the kind shown in this paper, with worse performance for second language speakers or speakers of stigmatised varieties, could have very concrete negative consequences for marginalised applicant groups. As voice user interfaces become increasingly important tools to access services (e.g. banking and customer service) and devices, predictive bias in ASR can both “degrade service[s]” [13] and actively deny services. Predictive bias in hate speech detection can also represent an allocative harm, if users are unfairly excluded from, for example, social media platforms because a post was erroneously flagged as “toxic”. While the prescriptive linguistic rules of social media platforms which are enforced by algorithmic tools, have been shown to inspire (fascinating) new language practices to avoid penalties [114, 123], creative resistance to algorithmic systems is not enough to avoid harm.

Stereotypical, discriminatory and hateful predictions in the context of language about communities in natural language generation tasks are representational harms [13, 86, 116]. For example, the islamophobic predictions by GPT-3 documented in [1] are both harmful to any Muslim users who see the offensive output and feed into (and, of course, result from) much larger islamophobic discourses harming all Muslims. Similarly, machine translation systems which reproduce gender and racial bias, can harm the immediate users and further perpetuate sexist and racist discourses [58, 107]. Hate speech detection tools also create representational harm to marginalised groups when they fail to detect actually hateful content, as highlighted in [39] by not “protecting” marginalised users from this content, while simultaneously reinforcing existing ideologies about, for example, what is and is not “obscene” language [112].

### 5.4 Mitigating bias in SLTs

Mitigation of data bias is an active field of research, but there are fundamental limitations to the extent to which data can be “de-biased”. Data is often likened to a “natural resource” [113] and specifically often (positively) compared to oil, due to its central role in many societies and economies. As [118] points out, the metaphor is also apt because machine learning is fundamentally extractive, and, I would add, just like the fossil fuel industry it does a lot of harm, in particular to marginalised populations. Unlike (unrefined) oil, however, data is not a “naturally occurring resource” that can simply be “collected” [18] – it is socially constructed [40, 99]. As such, it can never be fully free of “bias” (see also [60]). It is thus crucial that we account for all (structural) factors that have shaped a dataset, and highlight the labour that went into it [18, 40]. This includes where, how, when, why, about whom or what, by whom and for whom the dataset was compiled and, where relevant, who annotated it [52]. Understanding datasets as infrastructure and prioritising good documentation is a first step to mitigating harms of biased systems as it allows us to anticipate them [15, 71, 99].

Despite inherent limitations, “debiasing” approaches are increasingly popular in the field and have seen some success. In the context
of hate speech detection [42] show that relatively simple interventions such as adding non-toxic examples of identity terms which are associated with disproportionate levels of toxicity such as “gay”. While far from perfect, the Mozilla’s CommonVoice corpora do cover a much broader set of varieties, including some minoritised languages which do not have broad SLT support. Recent work in ASR furthermore shows that different model setups [74, 124] can mitigate accent bias. [108] show that recent efforts in increasing the representation of women on Wikipedia has had some limited positive effect on bias in word embeddings. [106] highlight that annotator bias can be limited when annotators are informed about sociolinguistic variation before labelling data. Other research on crowd-sourcing also highlights the importance of balancing annotators from diverse backgrounds and accounting for variation in values and opinions [12, 69]. [43] highlight that filtering large text corpora to mitigate bias is very challenging, especially because using a simple list of “bad words” risks removing all kinds of “non-toxic” uses of those terms by, in particular, marginalised groups (e.g. reclaimed slurs, other “obscene language”). As suggested in 4.2, better evaluation strategies of SLTs are also central to mitigating harms. An intersectional approach to quantitative evaluation can identify predictive bias which may be missed by more “aggregated” techniques. Furthermore adding qualitative methods allows us to pinpoint the exact implications, and sometimes, causes of undesirable, “biased” system behaviours.

Finally, it should be noted, as raised by [66], “not everything is a data problem”. While imbalances in training and test datasets are one important source of predictive bias in machine learning, particular model structures can amplify or even introduce biases [66]. But focusing on the data we use is crucial, especially as concerns about “data bias” are often dismissed (including by senior figures in the field [116]) as “non-technical” issues which fall (by extension perhaps) outwith the remit of machine learning engineers. Within this perspective, predictions of machine learning models which reflect, reproduce or actively worsen structural oppression are not considered an injustice (or even erroneous), but the “neutral” consequences of “accurately” reflecting the (racist, sexist, queerphobic etc.) world as it exists. But of course, we can (and, I would argue, should) instead choose to build technologies that work towards more liberated futures [34, 40]. Creating more representative datasets and carefully selecting models can limit harms of machine learning technologies. Another important consideration is whether to deploy (machine learning) technology at all [14].

5.5 Limits of bias and fairness

While the recent interest algorithmic oppression (or “bias”) in machine learning, including SLT applications, is a step in the right direction, there are fundamental limits to these discourses of bias and fairness [20, 56, 64]. As noted by [56], the framing of “bias” elides the structural nature of the issue and decouples it from power and oppression. Shifting our attention from “bias” to both oppression and concrete harms of algorithmic systems, allows us to account for power, forces us to reflect on our normative position towards harms, and focus on people’s lived experiences [24]. It also allows us to expand the discussion to include harms of systems which aren’t “biased” in a narrow sense. For example, an algorithmic system ranking job applicants based on “voice data” may not produce a higher rate of transcription errors for one group of speakers, but still applies language ideologies about “professional speech” when ranking applicants. This is particularly pernicious as, contrary to the standard language ideology, not everyone has equal access to these “right” ways of speaking (e.g. through education). Other seemingly “trivial” or “harmless” SLT applications like grammar checkers also encode deeply harmful language ideologies about “good” or “correct” language use without showing any predictive bias in a narrow sense. As [4] notes, those who can make themselves “legible” to the algorithm (e.g. by using the right words), according to the model of the world (or language) it has constructed will succeed (and continue to do so). Arguably, even hate speech detection tools flagging positive uses of (reclaimed) slurs as “toxic” are not “wrong” – they just lack access to the social context that licences only some people to use particular words depending on their positionality.

That does not mean that excluding users or removing content based on those decisions is not harmful. Aside from reinforcing (linguistic) ideologies, entirely “unbiased” systems can also be applied in deeply harmful ways. Right now, Amazon’s ASR, machine translation and other natural language processing tools are being used to facilitate surveillance of incarcerated people (and their contacts) in the United States [8]. Arguably, the harms resulting from this use of SLTs are particularly large if these systems exhibit predictive bias as they could result in, for example, criminal investigation based on incorrect transcripts.

On the other hand, “fixing” or mitigating predictive bias in this context, risks rendering marginalised populations even more legible [4] (to the state and to corporations) against their wishes (or at least, without their consent). As [100] put it: “[b]ias is real, but it’s also a captivating diversion”. Even when focusing on the harms of predictive bias of an algorithmic system, we risk overlooking the harms of the algorithmic system, full stop. Sometimes, conversations about the technical challenges or even the social contexts of “bias” (such as this paper), distract us from perhaps more urgent political conversations about the kinds of technologies we want to build and the kinds of futures we want these technologies to exist in. One way of integrating these conversations in the development of not just “fairer” but fundamentally (more) just SLTs [94] is to actively involve language communities. Participatory, community-led approaches to the development of speech and language infrastructure (such as Masakhane [95]) could be particularly beneficial for smaller or marginalised language communities, which are often overlooked or purposefully excluded by large technology corporations and academic institutions. Participatory and community-oriented approaches to data creation, curation and compilation are a way to ensure that ownership of the data stays with the community [18, 33]. Crucially, it would also give them a say in how their language(s) are represented and space for refusal [32].

Note that this is distinct from the high toxicity scores for entirely “neutral” terms like gay, trans or lesbian, which does constitute predictive bias.

Leo Technologies frames their services as safeguarding inmates, but [8] report that they are also being used to monitor “conversations involving mention of the Spanish word for lawyer or accusations that detention facilities were covering up COVID-19 outbreaks”.

Inaccurate or incorrect transcriptions of speakers of non-standard varieties are also a dangerous problem of human transcribers [162].
6 CONCLUSION
Like other machine learning technologies, SLTs reproduce the structures of oppression which shape the contexts in which they are designed and deployed, as a form of algorithmic oppression. SLTs tend to be designed for and by (language) communities which hold more power (within and across societies). As a result they are not only less useful for marginalised communities, but because of the complex interaction of (social) meaning, social context and identity, SLTs can inflict allocative and representational harms on marginalised communities. For example, as this paper shows, British English commercial ASR performs significantly worse for second language speakers of English and speakers of regional non-standard accents. Beyond this predictive bias, SLTs can also entrench existing ideologies about communities and the linguistic varieties they speak. Shifting our focus to the experiences of, in particular marginalised, people who use SLTs (or to whom they are applied) forces us to think carefully about what to do with “biased” systems and invites us to actively involve them in or let them lead technology design.

ACKNOWLEDGMENTS
This work was supported in part 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 and three anonymous reviewers for their generous feedback. Special thanks to Lukas Eigentler for help with Matlab.

REFERENCES


