Language variation, automatic speech recognition and algorithmic bias

Nina Markl

OILTHIGH DHÙN ÈIDEANN

PhD in Natural Language Processing with Integrated Study
The University of Edinburgh
2023

Contents

Signed Declaration

I declare that the thesis has been composed by myself and that the work has not been submitted for any other degree or professional qualification. I confirm that the work submitted is my own, except where work which has formed part of jointly-authored publications has been included. My contribution and those of the other authors to this work have been explicitly indicated below. I confirm that appropriate credit has been given within this thesis where reference has been made to the work of others.

For convenience, references to published works and joint manuscripts and publications are reproduced at appropriate places in this thesis.

- The results presented in Chapter 3 (and some of the discussion) are published in:

  Nina Markl (2022b). "Language Variation and Algorithmic Bias: Understanding Algorithmic Bias in British English Automatic Speech Recognition". In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. Seoul, Republic of Korea: Association for Computing Machinery, pp. 521–534. DOI: 10.1145/3531146.3533117

  For the purposes of the chapter, I have expanded on the background and discussion.

- The case study of using ASR to transcribe sociolinguistic data presented in Section 4.3 and Section 4.5.1 is published in:

  Nina Markl (2022a). "(Commercial) Automatic Speech Recognition as a Tool in Sociolinguistic Research". In: *University of Pennsylvania Working Papers in Linguistics* 28.2. URL: https://repository.upenn.edu/pwpl/vol28/iss2/11

  For the purposes of the chapter, I have re-arranged the contents of the published paper.

- I am the lead author of the manuscript presented in Section 4.4. I designed, conducted and analysed the survey evaluation study with my co-author Electra Wallington, provided suggestions for the design of the transcription workshop and analysed recordings of that workshop conducted by my co-authors Thomas Reitmaier, Simon Robinson, Gavin Bailey, Jennifer Pearson and Matt Jones. Electra Wallington, Ondřej Klejch and Peter Bell designed the ASR systems discussed. All authors contributed to the write-up. Some of the contents of this manuscript have (since the submission of this thesis) been published in:

  Nina Markl, Electra Wallington, Ondrej Klejch, Thomas Reitmaier, Gavin Bailey, Jennifer Pearson, Matt Jones, Simon Robinson, and Peter Bell (2023). "Automatic transcription and (de)standardisation". English. In: *Proceedings - SIGUL 2023, 2nd Annual Meeting of the Special Interest Group on Under-resourced Languages*. URL: https://sigul-2023.ilc.cnr.it/wp-content/uploads/2023/08/9_Paper.pdf

- Section 5.4 is published (in full) in:

  Nina Markl (2022c). "Mind the data gap(s): Investigating power in speech and language datasets". In: *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Dublin, Ireland: Association for Computational Linguistics, pp. 1–12. URL: https://aclanthology.org/2022.ltedi-1.1

- Section 5.3 is published (in full) in:
  Nina Markl and Stephen Joseph McNulty (2022). "Language technology practitioners as language managers: arbitrating data bias and predictive bias in ASR". in: *Proceedings of the Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 6328–6339. URL: https://aclanthology.org/2022.lrec-1.680

  I am the lead author on the paper. My co-author Stephen Joseph McNulty and I both contributed to the framing, theoretical development and writing of the manuscript.

- Section 6.2 is published (in full) in:
  Nina Markl and Catherine Lai (2023). "Everyone has an accent". In: *Proc. INTERSPEECH 2023*, pp. 4424–4427. DOI: 10.21437/Interspeech.2023-1847

  I conducted the analysis and wrote the first draft of the paper. As co-author, Catherine Lai provided supervision, comments and guidance.

- Section 7.2 is published (in full) in:

  Nina Markl and Catherine Lai (2021). "Context-sensitive evaluation of automatic speech recognition: considering user experience & language variation". In: *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*. Online: Association for Computational Linguistics, pp. 34–40. URL: https://aclanthology.org/2021.hcinlp-1.6

  I am the lead author of the paper. I conducted the analysis and wrote the first draft of the paper. As co-author, Catherine Lai provided supervision, comments and guidance.

- The description of the dataset in Section 7.3, Section 7.4 and Section 7.5 relate to this publication:

  Ramon Sanabria, Nikolay Bogoychev, Nina Markl, Andrea Carmantini, Klejch Ondřej, and Peter Bell (2023). "The Edinburgh International Accents of English Corpus: Towards the Democratization of English ASR". in: *ICASSP 2023*, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10095057

  I am a co-author on the project alongside Ramon Sanabria, Nikolay Bogoychev, Andrea Carmantini, Ondřej Klejch and Peter Bell. My contribution focusses in particular on the design of the dataset compilation process, including establishing ethical protocols and

securing ethics board approval, designing the elicitation questions and recording framework, designing the participant questionnaire and standardising participant responses to the questionnaire. The evaluation and experiments discussed in Section 7.5 and Sanabria et al. (2023) were conducted by my co-authors. All authors contributed to the writing of Sanabria et al. (2023) and some of the contents of Section 7.3 and Section 7.5 are drawn from this paper in an abridged form.

- The work presented in Section 6.3 was conducted with Stephen Joseph McNulty. The methodology and framing of this paper was developed jointly through discussion. Data analysis was also conducted jointly.

Abstract

In this thesis, I situate the impacts of automatic speech recognition systems in relation to sociolinguistic theory (in particular drawing on concepts of language variation, language ideology and language policy) and contemporary debates in AI ethics (especially regarding algorithmic bias and fairness). In recent years, automatic speech recognition systems, alongside other language technologies, have been adopted by a growing number of users and have been embedded in an increasing number of algorithmic systems. This expansion into new application domains and language varieties can be understood as an expansion into new sociolinguistic contexts. In this thesis, I am interested in how automatic speech recognition tools interact with this sociolinguistic context, and how they affect speakers, speech communities and their language varieties.

Focussing on commercial automatic speech recognition systems for British Englishes, I first explore the extent and consequences of performance differences of these systems for different user groups depending on their linguistic background. When situating this predictive bias within the wider sociolinguistic context, it becomes apparent that these systems reproduce and potentially entrench existing linguistic discrimination and could therefore cause direct and indirect harms to already marginalised speaker groups. To understand the benefits and potentials of automatic transcription tools, I highlight two case studies: transcribing sociolinguistic data in English and transcribing personal voice messages in isiXhosa. The central role of the sociolinguistic context in developing these tools is emphasised in this comparison. Design choices, such as the choice of training data, are particularly consequential because they interact with existing processes of language standardisation. To understand the impacts of these choices, and the role of the developers making them better, I draw on theory from language policy research and critical data studies. These conceptual frameworks are intended to help practitioners and researchers in anticipating and mitigating predictive bias and other potential harms of speech technologies. Beyond looking at individual choices, I also investigate the discourses about language variation and linguistic diversity deployed in the context of language technologies. These discourses put forward by researchers, developers and commercial providers not only have a direct effect on the wider sociolinguistic context, but they also highlight how this context (e.g., existing beliefs about language(s)) affects technology development. Finally, I explore ways of building better automatic speech recognition tools, focussing in particular on well-documented, naturalistic and diverse benchmark datasets. However, inclusive datasets are not necessarily a panacea, as they still raise important questions about the nature of linguistic data and language variation (especially in relation to identity), and may not mitigate or prevent all potential harms of automatic speech recognition systems as embedded in larger algorithmic systems and sociolinguistic contexts.

Lay summary

Automatic speech recognition is a common tool for interacting with computing devices like smartphones, computers and smart speakers). It creates captions for audio and video content, for example on social media. These systems are designed to produce a text transcription from speech sounds. The text can then be presented to the user or cause another action, for example searching the internet.

In this thesis, I explore how automatic speech recognition tools fit into how people use language and what people think about language. How people speak differs depending on who they are and who they are speaking to. Identities and experiences such as age, ethnic background, gender, social class background, region and linguistic background are only some of the factors affecting how people speak. People have many beliefs about "correct", "good", "proper" ways to use language.

I first assess commercial automatic speech recognition systems in the United Kingdom. I find that they work better for speakers from some regions than others. These differences mirror the difference in prestige of the speakers' accents. As voice technologies become more important, these performance differences can prevent speakers from accessing services and opportunities. They also signal to speakers that their way of talking is not "correct". What we consider "correct" language is often the "standard language". This is the kind of language found in dictionaries, taught in schools and used in formal settings. Often this standard language appears to be the default for voice technology development. I study the language isiXhosa in Cape Town, where speakers might not speak, write or like the "standard language". I find that it is very important to think carefully about which language(s) the system should recognise and how it should transcribe it. I discuss the roles and responsibilities of technology developers, using existing theories on how we decide to use different languages in society. I also look at how developers talk and write about language(s) in their work. This discussion provides an insight into their beliefs about language(s). Users might also adopt these beliefs. Finally, I explore how we can build better speech technologies. Like in many other uses of machine learning, under-representation of marginalised groups in the datasets used to create automatic speech recognition systems, causes "biased" outputs. I focus on how we can create more representative speech datasets. I find that it is important that many different types of speakers are included in these datasets and that they use their natural speech. It is important that recorded speakers understand how their data will be used because datasets are re-used by many developers. It is also essential to keep in mind that some of the risks of voice technology, like privacy risks, are not necessarily "solved" through more inclusive design.

Funding

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.

The work of my collaborators presented in Chapter 4 was in part funded by EPSRC (grant EP/T024976/1). Stephen Joseph McNulty is supported by an ESRC studentship (grant ES/P000681/1). The dataset compilation presented in Chapter 7 was in part funded by the Institute for Language, Cognition, and Computation at the University of Edinburgh.

Acknowledgments

Catherine Lai and Lauren Hall-Lew have been wonderful supervisors for this project – I am grateful for their advice, feedback, insights, encouragement, empathy, humour, support and confidence. I probably would not have chosen to apply for a PhD programme in Informatics without Josef Fruehwald's encouragement (and his Language Variation and Change class). Linguistics and English language (LEL) has been a welcoming home for me since the first day of my undergraduate degree 8 years ago when I did not know what a (socio)linguist did or that I was going to become one – I am extremely grateful to have been able to learn and work with so many LEL students and staff over the years. Peter Bell has made me feel equally welcome in Informatics – I am grateful for his feedback and advice and introducing me and my work to valued collaborators. The students and staff of the CDT for Natural Language Processing have provided an interdisciplinary and collaborative work environment (even during a pandemic). I would also like to thank my examiners Bea Alex and Britta Schneider for their insightful and generous comments and inspiring discussion.

I have been extremely lucky to have worked with many collaborators over the last four years. In particular, I would like to thank Stephen Joseph McNulty for countless (endless) discussions about language technology and language management, and, much more importantly, his humour and friendship which brighten even the dreichest days since 2020. Thanks to the Lothian Diaries Project Team, including Stephen, Lauren and Catherine, I did not feel as alone and helpless during an extremely isolated and anxious period (in all of our lives). I would also like to thank the UnMute team for being so welcoming and for letting me add a sociolinguistics angle to every conversation and every paper. I appreciate that the EdAcc team trusted my suggestions even where they made things more complicated. Eddie Ungless, Laurie Burchell, Jie Chi, and Tom Hosking have been brilliant team mates (on projects not included in this thesis) who have all taught me something when I felt completely out of my depth.

I am also, of course, indebted to the researchers and scholars I cite, and the countless research participants who contributed to their findings and theories. Beyond those cited here, I would like to mention Hannah McGregor whose public feminist scholarship has deeply influenced how I understand academia, feminist movements and my role within them.

Annie Holtz, Emma Kouhi, Bran Papineau aka "the group chat", have been the world's best co-conspirators and friends. Edinburgh would not feel as much like home and a PhD would have felt a lot more daunting without my friends, in particular the linguists and mathematicians. In Austria, my family has been extremely supportive in everything I have pursued (or attempted to pursue): Alice, Bernd, Edith, Alex, Werner, Angelika and Karl – danke!

Finally, a very big "danke" to Lukas Eigentler who has provided lots of practical help and thoughtful feedback on my research over the years. It's maybe not true that I couldn't have done this without him – but his almost frustrating trust in my ability to do anything I set my mind to, his unshakable sense of humour, genuine curiosity and unwavering love and support makes every day more joyful, fun, and so much more interesting.

Chapter 1

Introduction

In this introduction, I set out the aims and scope of this thesis and outline the structure and contributions.

1.1 Terms, aims and scope

This thesis focuses on the intersection of language variation, automatic speech recognition and algorithmic bias. I am interested in the impacts of automatic speech recognition (ASR) on speakers and their language(s). At the same time, I am interested in the ways in which the wider sociolinguistic context can impact ASR design and mitigate harms of algorithmic bias. In exploring the impacts of ASR in a range of applications, I focus in particular on the unequal distribution of risks, harms and benefits. To explore these broad questions, I have conducted several research projects with a range of collaborators across multiple disciplines which I present in this thesis. As outlined below, most of this work has been published in a range of venues for different audiences.

1.1.1 Terms

Language variation

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. Language is also fundamentally characterised by variation. We use this variation to convey and construct social meaning, both as speakers and as listeners (Eckert 2008; Bucholtz and Hall 2005). As discussed in depth in Chapter 2, I approach the study of language variation from a variationist perspective (see e.g., Tagliamonte 2011; Eckert 2012). Throughout the thesis, I link variation in language use (e.g., in terms of accent) to frameworks of language ideology and language policy. The study of language ideologies, or beliefs about language(s), is concerned with how we (as speakers and listeners) make sense of language

variation (Irvine and Gal 2000). Following Spolsky (2003), we can understand attempts to modify language use and/or language beliefs as "language management", a type of language policy. Taken together, language variation, language ideologies and language policies provide a conceptual framework to understand how and why people use language in different ways.

I generally use the term "language variety" or "variety" to refer to named languages (e.g., English), (socially or geographically bounded) dialects (e.g., Scottish English, African American English) or accents (e.g., Scottish Standard English). The term is useful in part because it sidesteps the question of "what counts as a language or dialect" which is largely a political and social, rather than linguistic, question. This "neutral" descriptor favoured by (socio)linguists which expresses our shared assumption (and assertion) that all language varieties are "equal", obscures the fact that some varieties have much higher status than others. Nevertheless, I believe thinking in terms of "varieties" is a useful (if imperfect) shorthand. I occasionally also refer to languages as if they were unproblematic, bounded artefacts, but, as discussed in Chapter 6, named languages are themselves ideological concepts.

ASR, AI, algorithmic systems

I understand automatic speech recognition (ASR) as a speech technology or speech and language technology (SLT), which more broadly can be grouped within "artificial intelligence" (AI) and "algorithmic systems". As discussed in Chapter 2, the basic task of ASR is to map speech sounds to (usually graphic) representations and it can be embedded in a wide range of different applications and domains. In addition to this versatility, what ASR systems share with other AI tools is the central role of "training data", and increasingly similar machine learning algorithms which are used across domains. ASR is often a component of a larger algorithmic system, whose aim is not just mapping speech to graphic representation, but some kind of decision-making (e.g., voice user interfaces). Both "AI" and "algorithmic system" are somewhat nebulous, but extremely common, terms in popular and academic discussions around machine learning technologies. I generally avoid the term AI, both because it encompasses such a wide range of other tools (and architectures), and because of the way it obscures, as Crawford (2022, p. 7) puts it succinctly, that machine learning tools are "neither artificial, nor intelligent". Instead, "AI" is dependent on (and perhaps best understood as) the decisions and interaction of (thousands of) people involved (knowingly or not) in the creation and maintenance of "data", algorithms and material infrastructure.

Algorithmic bias

As I discuss in Chapter 2, "algorithmic bias" is a very broad term which, in practice, is often uncoupled from more useful notions of "impacts" and "harms". However, I do think there is utility in broad terms like algorithmic bias because they allow us to connect different types of system behaviours, in different kinds of systems which affect the same groups and are

ultimately rooted in the same social structures and inequalities. To disentangle the different ways in which these underlying structures "show up" in a sociotechnical system, more fine-grained terminology does, however, help. I therefore do not frequently use the term when discussing specific system behaviour or their consequences.

1.1.2 Aims

My main aim in this thesis is to draw together scholarship from different fields to make sense of the impacts ASR tools have on individuals, their communities and their language varieties.

I approach this work from the perspective that language technologies *should* support individuals, their communities and their language varieties in their interactions with one another. (Linguistic) Communities *should* have a degree of control over the development and deployment of these technologies. Technologies are not neutral, but, as science and technology studies scholars have long pointed out, shaped by and expressive of politics (Winner 1980).[1] This is particularly true for algorithmic systems which have been shown to reproduce existing structures of oppression which shape the data these systems are trained on, and the contexts they are employed in (e.g., Noble 2018; Benjamin 2019a; D'Ignazio and Klein 2020; Costanza-Chock 2020; O'Neil 2017). Hampton (2021, n.p.), drawing on Noble (2018), uses the term "algorithmic oppression" to discuss the "violent impact [of technology] on marginalized people's lives" as it draws (or forces) our attention to the systemic nature of these harms which are fundamentally rooted in (pre-existing) social structures and thus not easily addressed through technical solutions (alone). Tools frequently discussed in the context of this "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) (O'Neil 2017; Benjamin 2019a; Eubanks 2018).

Speech and language technologies (SLTs) are an increasingly important site of algorithmic oppression. They are embedded in high-stakes contexts such as hiring (Raghavan et al. 2020; Drage and Mackereth 2022) and healthcare (Lee 2021) and ubiquitous in daily digital technology use (e.g., voice assistants, language models embedded in web search). Recently, the domain of automatic speech recognition in particular has been expanding to both new applications (e.g., captioning informal speech, voice user interfaces) and new language varieties (e.g., increasingly "low-resource" varieties). The positive and negative impacts of these expansions into some of these deployment contexts, and how they come about are the focus of this thesis. I contend that while ASR tools have many real benefits, especially where they support people in their interactions with each other and the wider world around them, they also impact speakers and their language varieties in ways that could be harmful. These harms can occur

[1] Winner (1980)'s famous example of the New York City underpasses which are too low for public buses, has been critiqued (Joerges 1999; Woolgar and Cooper 1999), but the the broader point that technical artefacts are shaped by and can express political stances, stands (Costanza-Chock 2020).

when they don't work equally well for different speaker groups, arise from the way they transmit discourses and beliefs about language(s) and language communities, and relate to the wider political economy in which these tools are built, sold and deployed. Because technologies are not *inevitable* but *built and used by people*, designers and "users" (and regulators and the wider public) can shape how algorithmic systems are developed, deployed and interpreted.

1.1.3 Scope

Seaver (2019, p. 419) suggests that "algorithmic systems", which are "intricate, dynamic arrangements of people and code", not algorithms by themselves, are what has "sociocultural impacts". He argues that in the context of algorithmic systems "cultural details", related to their development and deployment – their wider context, "are technical details" and should therefore be the focus of critical inquiry (Seaver 2019, p. 419). This focus on the social and cultural, rather than "only" or predominately technical aspect has some key advantages for researchers, especially those interested in commercially developed algorithmic systems. As much of the literature on algorithmic bias acknowledges, technical details of algorithmic systems such as relevant datasets, models and code are often, if not usually, opaque to researchers. Where datasets are not proprietary they are often difficult to examine due to their sheer size[2], and where models (or detailed documentation) are open-source, significant expertise is required to audit them. Deployed algorithmic systems are also, in a sense, "moving targets". Outwith the control of the researchers who investigate them, they can be frequently updated (in terms of models and data), making it difficult to estimate the long-term validity of any descriptions.[3]

As a result, I focus on these wider contexts, rather than (only) technical details. Some of the wider dynamics in the way ASR is developed and its impacts on speakers and their language varieties are not dependent on most of these technical details but instead related to the (more durable) sociolinguistic context.[4]

1.2 The structure of this thesis

1.2.1 Chapter outline

- In Chapter 2, I present relevant research context and background. I begin with a brief discussion of ASR tools, before introducing theoretical frameworks of language variation, language ideologies and language management. I then turn to notions of "bias" and "fairness" in algorithmic systems and introduce the wider framing of the chapters which follow.

[2]Though Birhane et al. (2021) show that it is possible to audit these manually.

[3]In many contexts, though perhaps not ASR tools, systems will also "look differently" and "behave differently" for different users, as interfaces and output is personalised (Seaver 2019).

[4]As I discuss in Chapter 8, there are many aspects and approaches I have not considered in this thesis but which would be fascinating future research directions.

- In Chapter 3, I explore how language variation relates to performance differences of commercial ASR systems in the context of British Englishes. I discuss how error disparities between different groups reflect existing linguistic hierarchies, and highlight the harms of this "predictive bias" to speakers and their language varieties.

- In Chapter 4, I consider the task of automatic transcription in different sociolinguistic contexts. I present two very different case studies (transcription of English sociolinguistic data and transcription of isiXhosa voice messages) which expose underlying assumptions in the way ASR is conceptualised which favour (and entrench) standard varieties. I also discuss how engaging with speech communities and the historical context of their language varieties is essential to designing useful language technologies and avoiding harm.

- In Chapter 5, I build on the work presented in the previous chapters, and explore biases in speech datasets and ways in which they are shaped by and reflective of existing power structures. I am particularly interested in the role and agency of researchers and developers compiling these "language resources". I explore how framing data compilation as an act of language management, can allow us to understand the role of developers and imagine alternative ways of building language technologies. I then present a paper where I apply the framework of "Data Feminism" to the issue of "gaps" in commonly used speech datasets.

- In Chapter 6, I explore how language, linguistic diversity and language variation are conceptualised in language technology research. I first consider how the concept of "accented speech" is discussed and motivated in a leading speech technology conference to explore the language ideologies which shape academic research within the field. I then analyse how the discursive construction of "language" in promotional materials related to "off-the-shelf" ASR engines reflects language ideologies, especially regarding the "value" of language and its relation to the nation.

- In Chapter 7, I highlight a practical example of compiling a "more diverse" and transparent speech dataset. I discuss the limits of this approach to "inclusion" and explore the harms of ASR applications beyond "bias".

1.2.2 Publications

Much of the work in this thesis has been published in peer-reviewed venues. Two of these papers are included in-full (in Chapter 5). Two are included in slightly modified versions (Chapter 3 and Chapter 4). Two sections represent submitted manuscripts (Section 4.4 and Section 6.2).

I have been lucky to work with researchers from a range of backgrounds, and in turn, present sociolinguistic perspectives to researchers in other fields. As a result, each (co-authored)

paper presented in this thesis has not only a different set of collaborators, but also a different intended audience (which I introduce with the paper). To contextualise them further, each chapter includes an introduction, some background and an extended discussion.

- The results presented in Chapter 3 (and some of the discussion) is published in:

  Nina Markl (2022b). "Language Variation and Algorithmic Bias: Understanding Algorithmic Bias in British English Automatic Speech Recognition". In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. Seoul, Republic of Korea: Association for Computing Machinery, pp. 521–534. DOI: 10.1145/3531146.3533117

  For the purposes of the chapter, I have expanded on the background, data and discussion.

- The case study of using ASR to transcribe sociolinguistic data presented in Section 4.3 and Section 4.5.1 is published in:

  Nina Markl (2022a). "(Commercial) Automatic Speech Recognition as a Tool in Sociolinguistic Research". In: *University of Pennsylvania Working Papers in Linguistics* 28.2. URL: https://repository.upenn.edu/pwpl/vol28/iss2/11

  For the purposes of the chapter, I have rearranged some of the contents of the published paper.

- The case study of using ASR in the context of isiXhosa voice messages presented in Section 4.4 is an *unpublished* manuscript.

- Section 5.4 is published (in full) in:

  Nina Markl (2022c). "Mind the data gap(s): Investigating power in speech and language datasets". In: *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Dublin, Ireland: Association for Computational Linguistics, pp. 1–12. URL: https://aclanthology.org/2022.ltedi-1.1

- Section 5.3 is published (in full) in: Nina Markl and Stephen Joseph McNulty (2022). "Language technology practitioners as language managers: arbitrating data bias and predictive bias in ASR". in: *Proceedings of the Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 6328–6339. URL: https://aclanthology.org/2022.lrec-1.680

- Section 6.2 is published (in full) in: Nina Markl and Catherine Lai (2023). "Everyone has an accent". In: *Proc. INTERSPEECH 2023*, pp. 4424–4427. DOI: 10.21437/Interspeech.2023-1847

- Section 7.2 is published (in full) in:

  Nina Markl and Catherine Lai (2021). "Context-sensitive evaluation of automatic speech recognition: considering user experience & language variation". In: *Proceedings of the*

First Workshop on Bridging Human–Computer Interaction and Natural Language Processing. Online: Association for Computational Linguistics, pp. 34–40. URL: https://aclanthology.org/2021.hcinlp-1.6

- The description and discussion of the dataset in Section 7.3, Section 7.4 and Section 7.5 relate to the Edinburgh International Accents of English dataset, which we introduce in this publication:

  Ramon Sanabria, Nikolay Bogoychev, Nina Markl, Andrea Carmantini, Klejch Ondřej, and Peter Bell (2023). "The Edinburgh International Accents of English Corpus: Towards the Democratization of English ASR". in: *ICASSP 2023*, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10095057

1.2.3 Collaborators

Since January 2022, I have been involved in the UnMute Project, an interdisciplinary research project consisting of a team of speech technologists and human-computer interaction researchers and contributed to two collaborative research papers: Reitmaier et al. (2023) and a manuscript presented in Section 4.4 of which I am the lead author. I designed, conducted and analysed the survey evaluation study with my co-author Electra Wallington, provided suggestions for the design of the transcription workshop and analysed recordings of that workshop conducted by my co-authors Thomas Reitmaier, Simon Robinson, Gavin Bailey, Jennifer Pearson and Matt Jones. Electra Wallington, Ondřej Klejch and Peter Bell designed the ASR systems discussed. All authors contributed to the write-up. The manuscript also builds on our joint publication, Reitmaier et al. (2023), which is not included in this thesis.

The paper presented in Section 5.3 was co-authored with Stephen Joseph McNulty, who works on the sociology of language and language policy. I was the lead author of the paper but SJM's theoretical contributions were essential for the paper and we both contributed to the writing. The data presented in Section 6.3 is the foundation of our follow-up project which considers language management of commercial language technology providers. The data analysis and discussion presented in Section 6.3 was written up by me but is the product of collaborative work on a presentation and manuscript in preparation.

The manuscript submitted to Interspeech in Section 6.2 and the paper in Section 7.2 were co-authored by Catherine Lai who provided supervision, comments and guidance.

The dataset discussed in Section 7.3, Section 7.4 and Section 7.5 was compiled in collaboration with Ramon Sanabria, Nikolay Bogoychev, Andrea Carmantini, Ondřej Klejch and Peter Bell. As discussed in Section 7.3, my contributions focusses in particular on the design of the dataset compilation process, including establishing ethical protocols and securing ethics board approval, designing the elicitation questions and recording framework, designing the participant questionnaire and standardising participant responses to the questionnaire. The evaluation and experiments discussed in Section 7.5 and Sanabria et al. (2023) were con-

ducted by my co-authors. All authors contributed to the writing of Sanabria et al. (2023) and some of the contents of Section 7.3 and Section 7.5 are drawn from this paper in an abridged form.

1.3 Contributions of this thesis

In this thesis (and the published papers included it), I make contributions to different strands of research.

- I highlight how theoretical frameworks drawn from sociolinguistics and linguistic anthropology (variationism, language ideology and language management) can contribute to our understanding of the causes and effects of algorithmic bias in language technologies, as well as the development of language technologies more broadly.
- I contribute quantitative and qualitative evaluations of automatic speech recognition systems which show that existing linguistic discrimination is mirrored in the performance of these systems.
- I explore the role of automatic speech recognition in larger standardisation processes, in particular in the context of minoritised and/or "under-resourced" language varieties. This exploration consists of a critical discussion of existing literature on transcription and orthography, as well as specific case studies of the development and use of automatic transcription systems for two different language varieties.
- I discuss the relationship between social structures, power and speech datasets and explore ways of understanding language data compilation as a social process of language management. I also contribute a documentation guideline intended to highlight and interrogate gaps in datasets.
- I analyse how linguistic and social concepts like "language" and "accent" are conceptualised in speech technology development (in academic and corporate contexts). This investigation highlights differences in how these terms are understood across fields and surfaces language ideologies which may be reflected in the technology design.
- I contribute a discussion of the process of speech data compilation and speech data documentation, exemplified on the "Edinburgh International Accents of English" dataset, which I compiled with a group of collaborators. In this context, I also reflect on the limits of notions of "inclusion".

Chapter 2

ASR, language variation and algorithmic bias

2.1 Introduction

In this chapter I provide the wider research context within which the work in the following chapters is situated. I begin by explaining what automatic speech recognition (ASR) is, outline how it works, and how it is used today. The sociolinguistic perspective I draw on is introduced in Section 2.3: I first introduce variationist sociolinguistics and highlight how it relates to speech technologies. I then discuss language ideologies (Section 2.3.2) and language management (Section 2.3.3). I then provide a brief overview of the field of AI ethics, with a particular focus on how algorithmic bias and algorithmic harms have been conceptualised. Against this research context, I outline the overarching questions, which the following chapters aim to address.

2.2 Automatic Speech Recognition in society

The task of automatic speech recognition (ASR) is mapping input sound to output text. This deceptively simple task formulation belies the numerous challenges involved in every sub-task, from processing the input waveform, to correctly identifying sounds and words, and transcribing them according to a particular orthographic convention. Here I begin by tracing a brief history of ASR, outlining how modern ASR systems work (with a particular focus on the datasets used to train and test systems) and where they are commonly deployed.

2.2.1 A brief history of ASR

While people have been interested in creating "talking machines" since the 18th century[1], the first machines designed to recognise, rather than mimic, speech were developed in the 1950s and 1960s (Juang and Rabiner 2006; Furui 2005). Most of these early systems were designed to recognise isolated syllables or digits produced by specific speakers (e.g., Davis et al. 1952), while others were able to segment the speech stream (e.g., Sakai and Doshita 1961). The 1970s saw advances in time alignment between utterances using dynamic programming and the first commercial ASR tool VIP-100 (deployed for tasks like package sorting on conveyor belts) (Juang and Rabiner 2006). According to Juang and Rabiner (2006), this first commercial tool had a lasting impact as it sparked interest by the US Department of Defense whose (Defense) Advanced Research Projects Agency (DARPA)[2] would go on to fund a lot of ASR research over the following decades (and does so to this day). In addition to (D)ARPA funded projects[3], IBM and AT&T Bell Laboratories were working on speaker-dependent systems for transcription (which needed to be calibrated for a particular voice), and speaker-independent systems for simple voice commands (which could work for any speaker), developing crucial components of the ASR systems we see today, such as language models, signal processing techniques and clustering algorithms (Juang and Rabiner 2006).

Until the 2010s, most ASR systems modelled phones and sequences of phones using a stochastic process called Hidden Markov Models (HMMs) (Juang and Rabiner 2006; Huang et al. 2014). These "conventional" systems generally consist of several components including a speech processing unit, an acoustic model, a language model, a dictionary, and a decoding unit. Acoustic models contain probabilistic representations of speech sounds, while language models contain probabilistic representation of utterances (word sequences). Big advances in deep learning methods have since led to the adoption of novel ASR model architectures. In addition to better performance on many tasks and benchmarks, these new approaches also mark a change in the kind of data used to train ASR systems. Conventional systems rely on "supervised" data to separately train an acoustic model (AM) and a language model: speech data with paired "gold" transcripts (AM) and unpaired text data (LM). The most recent architectures, on the other hand, can be trained in a "semi-supervised" (Zhang et al. 2023; Radford et al. 2022) and "unsupervised" (Baevski et al. 2021) manner, requiring much smaller amounts of labelled speech data (i.e., paired transcribed speech) as they can instead harness larger amounts of unlabelled (i.e., untranscribed) speech data and unpaired text.

[1]Notable examples include the *Voder* designed by Bell Labs which was presented at the 1939 World's Fair in New York and was based on the human vocal tract: `https://archive.org/details/bellsystemtechni19amerrich/page/495/mode/1up?view=theater`.

[2]Since its founding in 1958 the agency has been known as DARPA and ARPA.

[3]For an overview of projects see Juang and Rabiner (2006) and Huang et al. (2014) – conducted at universities, they included engineering projects and data compilation efforts for some of the most common benchmark datasets as also discussed in Chapter 5.

2.2.2 ASR applications today

With the changes in architectures, performance, and data and hardware requirements we have also seen a proliferation of application contexts of ASR systems. They now include tasks like dictation, voice search and control, and transcription of voice messages, speeches, conversations and meetings and media like audiobooks, podcasts, movies and TV (Aksënova et al. 2021). As Aksënova et al. (2021) highlight, these tasks differ widely in terms of technical requirements, training and benchmarking datasets. For example, speeches and audiobooks represent planned and recited speech (Gabler et al. 2023), which is very different from unplanned conversation or scripted conversation with multiple speakers (and other sounds) (Aksënova et al. 2021).

There are many potential benefits of this proliferation of ASR technologies: automatic captioning of recorded and streamed video and audio content, including on social media platforms, has the potential to make this content more accessible to users who are Deaf or Hard of hearing[4]; voice user interfaces enable new ways of engaging with computing devices without typing or using a touchscreen, as well as controlling "smart home" features like lights and speakers which can be particularly useful for users with limited mobility (Pradhan et al. 2018); language learners have been shown to feel less anxious during conversation practice with an ASR tool than in the classroom (Bashori et al. 2020). Like other language technologies, ASR tools are also becoming available for more and more languages and language varieties, addressing a long-standing gap between so-called "high-resource" and "low-resource" or "under-resourced" varieties[5]. The former tend to be languages already associated with wealth and (geopolitical) power such as English, Spanish, Mandarin and French, while the majority of languages (and a plurality of first language speakers) do not have the (language) resources in the form of (transcribed and digitised) speech datasets, machine readable text data (Joshi et al. 2020), and, increasingly, computing infrastructure to build robust, modern language technologies. These developments have the potential to extend the benefits of ASR to more people and positively impact the status of minoritised varieties.

At the same time that ASR systems are expanding in terms of domains and language varieties, they are moving into "high-risk" applications where errors are costly. Though the use of algorithmic systems in employee recruitment is not new, recent years have seen the rise of "automatic video interview screening". Provided to employers by external providers like HireVue[6], these systems automatically analyse candidates' videos of answers to interview questions (Bogen and Rieke 2018). While some of these systems disregard much of the visual input [7], the

[4] If it works perfectly – where it does not, it is usually insufficient, as the Web Accessibility Initiative points out: `https://www.w3.org/WAI/media/av/captions/`

[5] These terms are problematic because of the way in which they define languages embedded in rich and complicated and messy social and cultural pasts, presents and futures by the absence of an extremely specific type of object, a point I return to in Chapter 5

[6] `https://www.hirevue.com/`

[7] HireVue discontinued its controversial analysis of facial expressions in 2021 (Maurer 2021) – other providers still analyse facial expressions (Drage and Mackereth 2022).

screening does involve automatic speech recognition and subsequent analysis of candidates' responses (Maurer 2021; Bogen and Rieke 2018). "Speech analysis" is also sold to companies to evaluate applicants for promotion, and analyse interactions between customers and employees (Morrison 2017).[8] In these contexts algorithmic bias in the form of higher ASR error rates or lower "scores" (in terms of hireability or other desired criteria) for some applicant groups could have serious consequences. In this way, the "risks" associated with ASR systems (and other language technologies) depend on both the user and application context.

Overall, steady advances in performance coupled with expansion of application domains of ASR opens up new avenues for research. Many of these new domains raise new or highlight old technical challenges such as variation in speech rate, differences in acoustic environment and encoding formats (Aksënova et al. 2021). In this thesis, I am interested in exploring the way in which different types of ASR systems affect speakers and languages – and ways in which speakers and languages might challenge ASR systems. As new and more diverse groups of users adopt and adapt ASR systems, pre-existing biases and assumptions embedded in datasets and task design are accentuated. New application domains also raise important ethical debates which benefit from a sociolinguistically-informed perspective. For example, understanding the wider historical contexts in which some languages have become "endangered", "minoritised" and "under-resourced" through violent oppression (Heller and McElhinny 2017) helps us avoid the "trap" of tech-solutionism (Broussard 2019) and colonial modes of technology development (Bird 2020; Bird 2022; Dourish and Mainwaring 2012; Srinivasan 2017) and encourages us instead to defer to language communities and speakers on decisions regarding their languages. Contextualising performance differences between varieties with their status in society allows us to understand the potential harms of these biases (Blodgett et al. 2020), and ways to mitigate them. Close examination of language technologies can also contribute to sociolinguistic theories, as they are a fascinating new "actor" in the sociolinguistic context. Recent work on human-computer interaction and communication has explored the role of ASR-based voice assistants as a new type of interlocutor especially in social interactions (Porcheron et al. 2018; Schneider 2021; Seymour and Kleek 2021). In this thesis, I am particularly interested in the, as-of-yet under-explored[9], role of other types of ASR tools (especially those used for transcription) and their developers in transmitting and shaping existing beliefs about language(s).

2.3 Language in society

Since I am interested in exploring how ASR interacts with language variation and existing sociolinguistic context, I introduce three core concepts in the following section: language variation,

[8]The company discussed by Morrison (2017) has since changed its name to VIER Precire GmbH and offers speech and emotion analysis tools but appears to have discontinued some of its initial products focussed on psychological analysis based on speech samples (Bös 2021).

[9]Though there is increasing interest, notably by large research groups like the EU-funded COST LITHME group (Sayers et al. 2021; Schneider et al. 2022)

language ideology and language management. Taken together, these concepts and the scholarship around them can help us understand both how and why language(s) vary and how and why different ways of using language have different status in society.[10] They also allow us to understand some of the effects language variation has on ASR design and performance. Conversely, we can also use them to theorise the ways in which ASR design, evaluation and deployment affects speakers and varieties.

2.3.1 Language variation

Language is both inherently social and inherently characterised by variation. In addition to communicating linguistic meaning, including information necessary to create, negotiate, maintain or even dismantle communities and social structures, we also use language to convey and construct social meaning. Variation occurs at all linguistic levels (syntax, morphology, phonology, phonetics, pragmatics, semantics), and, like change is a fundamental feature of language. The field of variationist sociolinguistics focusses on the origin and structure of this variation, especially how it relates to social context (Tagliamonte 2011). While it is only one analytical frame we can apply to linguistic variation, it has proven particularly useful through its combination of quantitative methods and careful analysis of social context. A central insight is that while linguistic variation may appear random or free when we first encounter it (for example when we enter a new language community), it is usually highly structured in both individuals and communities. As Weinreich et al. (1968) put it in a very influential formulation, language variation (and language change) is characterised by "orderly heterogeneity". Variation is furthermore often socially meaningful. How people express "the same" (kind of) meaning depends then on who their own position within a society, their interlocutors' position and identities as well as the specific interactional context.

Eckert (2012) describes the study of social meaning in sociolinguistic variation as consisting of three (not necessarily strictly historical or consecutive) waves. While the "first wave" focuses on the relationship between macro-social structures and patterns of language variation and change, others consider local social categories and interactional contexts, as well as the role of agency.

The first wave

As Eckert (2012) highlights, studies in the first wave consider language variation between individuals and groups, e.g., differences in phonology, to be shaped by macro-social categories. The central insight of these first wave studies is that language variation is structured in individuals and communities, and socially stratified. Much of the foundational work in this area was

[10] Here (and throughout the thesis) I focus on spoken languages – this is in part a reflection on biases in the field which mean that most (socio)linguistic theory takes spoken rather than signed or tactile languages as their object of study (Henner and Robinson 2023). Since ASR specifically relates to speech, examples relating to variation in speech may also be easier to follow.

conducted by William Labov. One of his key theoretical concepts here is that while we all have access to a range of speech styles, there is a particular way of speaking which comes most "automatically" to us called the *vernacular* (Labov 1972). Drawing on other social science fields, these studies explored the "effect" of categories like social class, race, ethnicity and gender[11] on the vernacular (Tagliamonte 2015). These social categories, which are of course, highly contingent on time, place and cultural context, are usually recorded "objectively" by applying empirical measures of social class tied to income, family history, occupation, and education. Many of the foundational studies in the field were conducted in the United States, in particular by William Labov in New York City in the 1960s, focussing on (phonological) variation in English (Labov 2009). Similar studies have since been conducted in other urban centres and other languages in North America (Sankoff and Cedergren 1972), the United Kingdom (Macaulay 1977; Trudgill 1972), Central America (Cedergren and Sankoff 1974). While the specific linguistic variables of interest differ, broad social patterns are remarkably similar in different sociolinguistic contexts. For example, lower socio-economic status of speakers is correlated with more frequent use of stigmatised variants than higher socio-economic status, and the degree of regional and ethnic variability is lower among high-status speakers. Within this framework, speakers are understood to shift their linguistic style as they pay more or less attention to their speech (Labov 1972).

In the context of language technologies, some insights from this "first wave" are crucial. The fact that language variation is socially stratified means that access to (reliable or useful) language technologies is too especially when these technologies are built for some (but not other) varieties. Language variation can be a proxy for macro-social categories and structures. If a language technology performs worse for a particular language variety or features, it therefore often performs worse for a particular group of people. This is particularly problematic because language technologies tend to be built only or primarily for the (high status) vernaculars of high-status speakers. The insight that all varieties and the variation they encompass are structured is therefore important for improving access to language technologies. Empirically grounded understanding of different varieties can be used to audit language technologies and discover and mitigate performance differences between varieties and build systems specifically for different varieties (see e.g., recent dissertations by Blodgett 2021; Martin 2022).

The second wave

While the first wave frames variation as indicative of macro-social structures or categories, the second wave considers the "local context" of variation through ethnographic methods. This perspective moves away from seeing style as (only) determined by speakers' (static) position within a social structure and the attention they pay to their speech based on the specific situ-

[11] Specifically, many studies refer to the effect of *sex*. While this was often inferred by researchers, rather than self-described by research participants, this usually refers to gender as a social category. In any case much of linguistic research has reproduced the gender binary in problematic ways (Tripp and Munson 2021).

ation. A new focus on speakers' agency to draw on social meanings associated with linguistic variation in interaction has proven extremely influential in the study of language variation since. The ethnographic approach has also uncovered the importance of "local" categories and "local" meaning which can account for differences *within* (macro-social) groups. Studies have, for example, considered differences between groups of adolescents who orient differently to authorities (Cheshire 1982) or school (Eckert 1989), and groups of otherwise similarly positioned adults in rural communities who relate differently to the local (Rickford 1986) or national economy (Holmquist 1985).

The second wave complicates the role of language variation in language technologies. While macro-social categories are historically and geographically contingent, they still generalise to some extent. Perhaps one insight to draw on is that *local context matters*. On the one hand, a more detailed understanding of the "users" helps to reveals differences in language use within a group, and on the other hand, it helps us understand how they interact with language technologies.

The third wave

"Third wave" variationist studies constitutes another shift in the theorising of the relationship between the social meaning of language variation and social identity. It is a shift from recognising that some linguistic variants are *correlated* with particular social identities, stances and social group membership, to theorising that the these identities and stances are *constructed through* these linguistic variants (Bucholtz and Hall 2005; Eckert 2008; Eckert 2012). Within variationist studies and (quantitative) linguistics more broadly, this framing has dramatically changed the understanding of speakers' agency in the context of language variation and change, and the notion of social identity. Socially meaningful linguistic variation is not the "incidental fallout" (Eckert 2012) of a broader social structure, but rather one of the ways in which we build, maintain and challenge social structure(s). Importantly the social meanings attached to any linguistic variable are not fixed. They only index macro-social categories indirectly, and their meaning depends on speaker, speech situation and hearer. Drawing on Silverstein (2003), Eckert (2008) introduces the idea of the "indexical field" encompassing ideologically related meanings of a linguistic variable. As we use language we can draw on these meanings to construct social identity, express stances by combining different linguistic variables into styles (e.g., Podesva 2007; Zimman 2017).

This decoupling of linguistic variation from static, predetermined social categories complicates our initial understanding of linguistic variation as proxies for particular social identities and raises important questions about language technology design and algorithmic bias. Firstly, it means that even though social identity is constructed through language variation we cannot straightforwardly "predict" social identity based on language use. Studying correlations between macro-social categories and language variation in online spaces has, for example, become increasingly popular at the intersection of sociolinguistics and natural language pro-

cessing (Nguyen et al. 2016). This kind of research can (sometimes inadvertently) amount to inferring user-level metadata such as gender, age, race and ethnicity without their consent. However, if people re-combine linguistic features associated with macro-level categories such as race, class and gender to create complex social meanings and identities in interactions, these approaches may not be reliable. Ilbury (2020) highlights a central limitation of this approach in his study of the use of features of African American Vernacular English (AAVE) by white British gay men on Twitter. Rather than using AAVE consistently in writing (or, presumably, speech) these men use specific features of AAVE to construct a particular style by evoking existing characterological figures associated in particular with Black women (Ilbury 2020).

Recent work has argued for the importance of incorporating social meaning in the design of language technologies which we want to be just or promote justice (Sutton et al. 2019; Nguyen et al. 2021; Nee et al. 2021; Blodgett 2021). The third wave perspective laid out above is crucial in this context. It helps us understand the complexity of language variation in interaction and raises important questions about how we could or should go about building language technologies which are better, more just and equitable. For example, as discussed above, linking language variation and macro-level social categories race, gender and class can help us audit language technologies for algorithmic bias. However, this very same linkage risks stereotyping (potential or actual) language (technology) users and glosses over a huge diversity in language use between and within social groups. The third wave also highlights that language is ideological: as discussed in more detail in the next section, the meanings we attach to variants and varieties are embedded within broader ideological frameworks and socio-cultural and historical contexts. Which varieties and variants we develop for is always a political and ideological choice. While researchers may be constrained by wider social structures (e.g., funding incentive structures, data availability etc.), these constrains too are the results by pre-existing social and linguistic hierarchies. In addition to potentially making language technologies more widely accessible, incorporating variation and social meaning can also make already marginalised groups subject to profiling, linguistic and cultural appropriation and automated classification (e.g., by race or gender).

2.3.2 Language ideologies

Language and language variation are always situated in a larger social context, and are interpreted differently depending on who is speaking, listening, signing, writing or reading. Haraway's notion of the "god trick of seeing everything from nowhere" (Haraway 1988, p 581) is relevant not just in the context of understanding science and technology but also in the context of language. The "god trick" is the illusion of complete, "transcendent" objectivity, when in reality our perception and interpretation, just like our knowledge is framed through a particular embodied lens (Haraway 1988). A multitude of studies looking at how linguistic variation is produced and perceived have contributed to understanding how language attitudes and language ideologies are developed and maintained, and how they in turn maintain broader social

structures.

As discussed in section 2.3.1, sociolinguists have long been interested in how particular ways of using language can become associated with and indexical of specific social identities and positionalities (Jaffe 2016). As people who use language, we notice correlations between particular language varieties or linguistic features and the people who use them. While these correlations are arbitrary, in the sense that there is no pre-social, essential link between form and social meaning, we nevertheless try to rationalise and justify them. In a highly influential paper, Irvine and Gal (2000) consider how "ideological representations of linguistic differentiation" are created through three semiotic processes (*iconisation*, *fractal recursivity* and *erasure*). As they put it, language users construct ideologies about language which "locate linguistic phenomena as part of, and evidence for, what [they] believe to be systematic behavioral, aesthetic, affective and moral contrasts among the social groups indexed" (Irvine and Gal 2000, p. 37). The relationship between a particular way of using language and a particular way of being in the world becomes indexical through iconisation (Irvine and Gal 2000, p. 37). Variation within different groups is erased and (binary) ideological oppositions and marco-social categories are reproduced and mapped onto distinctions between and within groups (fractal recursivity) (Irvine and Gal 2000, p. 38). The result of this process is that our knowledge about language variation becomes part of the broader ideologies we hold. Like other ideologies, language ideologies can become very deeply embedded in our sense of how the world is and should be.

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 particularly evident in the fact that the same linguistic feature is often interpreted differently depending on who produced it. For example, creaky voice, a phonation type commonly also known as "vocal fry" (Davidson 2020), is, among English speakers, much more stigmatised and pathologised in young women's speech than men's (Anderson et al. 2014; Chao and Bursten 2021). The terms used to evaluate the feature are also evaluations of the women who use them: "annoying", "grating", "too much to bear" (Chao and Bursten 2021; Glass 2015). 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 (Smith and Holmes-Elliott 2018; Kirkham and Moore 2016). Notably, these differences in the social evaluation of language variation depending on the (perceived) identity of the speaker can be reproduced experimentally. In "matched guise studies", listeners are asked to make inferences and evaluations of speakers based on short recorded utterances differing only in the realisation of a target variable (e.g., Campbell-Kibler 2009; Mack and Munson 2012). These studies indicate not only which kinds of social meanings are consistently associated with particular linguistic features,

but also show that these evaluations of speakers (e.g., education, attractiveness, sexuality, trustworthiness, friendliness) are also influenced by their perceived race, gender and social class background (Dixon et al. 2002; Baese-Berk et al. 2020; Campbell-Kibler 2009).[12]

Language "attitudes" and evaluations have (more and less obvious) structural implications (see also Craft et al. 2020). Anderson et al. (2014) asked listeners to rate speakers with and without vocal fry according to their "hireability", and found that those without creaky voice were preferred. This is just one example among many culturally-specific language ideologies around "professional", "educated" and "articulate" speech (Lippi-Green 2012; Baratta 2017). In anglophone settings, hiring committees disprefer second language speakers (Hosoda and Stone-Romero 2010; Timming 2016) who have also been found to be perceived as "less credible" (Lev-Ari and Keysar 2010) than first language speakers. 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 (Roessel et al. 2020). Nevertheless, in the UK, L1 accents associated with working class speakers 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" (Sharma et al. 2022). 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 to be unambiguously expressed than other forms of discrimination. This is evident, for example, from the omni-present metalinguistic public discourse about "good" and "bad" language on social media and in newspapers (Lukač 2018; Wright and Brookes 2018). First hand accounts of linguistic discrimination are also collected by the Accentism Project.[13] While some of this rhetoric, especially as expressed in right-wing media, is explicitly and transparently racist and/or xenophobic, the focus on the seemingly common-sense ideologies about language allows inherently racist and classist ideas to be expressed in ways that appear more broadly socially acceptable (Wright and Brookes 2018). Recent work shows that accent discrimination still plays a role in high-prestige hiring contexts such as corporate law, although not all regional accents are equally stigmatised (Cardoso et al. 2019; Levon et al. 2021). Accent bias has also been documented in teacher training and schools in the UK, affecting both first and second language speakers of English (Baratta 2017; Cushing and Snell 2022).

Given the pervasive and persistent nature of language ideologies, it stands to reason that speech technology development too is at least to some extent shaped by them. A central aim of the research in this thesis is to explore this role of language ideologies by paying close attention to how researchers and companies developing ASR tools conceptualise language variation and linguistic diversity, and ways in which some ideologies are already embedded in development

[12] Interestingly, it appears that speech perception can also be affected by visual stimuli which are not related to the speaker. Hay and Drager (2010) report that New Zealand England speakers' perceptions of whether a (synthesised) vowel token (and the word it was embedded in) was produced by a New Zealander or Australian were influenced by whether they saw stuffed kangaroo toys or stuffed kiwi toys in the room where the experiment took place.

[13] `https://accentism.org/`

pipelines and datasets. Understanding the already existing linguistic discrimination also helps us understand the harms of predictive bias in ASR.

2.3.3 Language management and language policy

Language ideologies are used to justify and re-entrench particular power structures and construct notions of normativity, markedness, difference and similarity between social groups (Craft et al. 2020; Rosa and Burdick 2016; Irvine and Gal 2000). The way this is achieved is, in part, through the way ideologies about language inform language *management* within a social context. Spolsky (2003) uses the term language policy to include the way a group of people use language (their "language practices"), their beliefs about language and the way they manage language. In the previous sections, I discussed how and why language variation arises and in turn forms the basis of language ideologies. In this section, I will show how the notion of language policy or management can help us understand how these ideologies are propagated through institutions and implemented as language policy. As I argue in later chapters of this thesis, speech technologies are also affected by, and in turn affect, language management.

Two common language ideologies particularly relevant to language technologies concern the notion of "standard language" and ideologies of the "boundedness" of language(s). The "standard language ideology" is, at its most basic, the belief that language can and should be "standardised" (Lippi-Green 2012; Milroy 2001; Spolsky 2003). Milroy (2001) draws attention to the use of "standardization" as a process of imposing uniformity. Like the standardisation of objects, measurements and tools (Bowker and Star 2000), language standardisation is also not a neutral, but a political process of language management or planning. Language planning as a field of research and practice developed in the 1960s, as linguists became involved as "experts" to "solve language problems" of post-colonial and/or newly independent nation states (Johnson 2013, p. 27). The tasks included developing standardised varieties to be used in education, though in many settings languages imposed by colonial powers retain(ed) official status after decolonisation (Heller and McElhinny 2017, 200 ff). Standardising languages involves selecting a variety (as there are always several different styles or varieties to choose from) and codifying (a written form of) this variety in dictionaries and grammars (Johnson 2013). Crucially, the choices involved in this process are guided implicitly and explicitly by language ideologies as well as pre-existing power structures (Spolsky 2003; Johnson 2013; Irvine and Gal 2000; Ricento 2000; Shohamy 2006). These ideologies can include beliefs about which varieties are already associated with powerful groups and, consequently, considered to be "good" or "appropriate" or "efficient" and which forms of that variety (e.g., in the context of orthography) are "good". This link between standardised language (as in uniform or codified) and Standard language (as in prestigious) is therefore no coincidence (Milroy 2001; Silverstein 1996). To further spread a standard and entrench its status, it is adopted in a variety of domains, including, education and government.

In the British context, Standard Southern British English (SSBE) and Received Pronuncia-

tion (RP) are highly prestigious. SSBE is a variety of English originating, as the name suggests, in the South of England. However, it is not only spoken by people in the South of England, and most people in the South of England also use other varieties. Fabricius calls SSBE the "generational successor" to Received Pronunciation (RP) (2018, p. 35) and many scholars use the terms interchangeably in more recent discussions (e.g., Hughes et al. 2013). Like RP it 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 (Agha 2003; Fabricius 2018). Only a small group of people use SSBE in all contexts, including informal ones, but it is widely used in formal contexts throughout the UK. In Scotland, Standard Scottish English (SSE), which differs from SSBE (almost) only in pronunciation, occupies similar status. RP has historically been particularly widely used in British media and in elite spaces (private schools, politics, aristocracy) (Agha 2003). The strong association between RP/SSBE and upper class status has been reinforced since the eighteenth century through prescriptive teaching (inside and outside classrooms) and popular media (Agha 2003; Mugglestone 2007; Cushing and Snell 2022). Cushing and Snell (2022) trace how ideologies around "Standard English" are implemented as language policing by government inspectors in British schools. They specifically consider Standard English as being produced by the standard language ideology and raciolinguistic ideologies (Cushing and Snell 2022). This raciolinguistic frame proposed by Rosa and Flores (2017) allows us to understand how "racialized speaking subjects who are constructed as linguistically deviant even when engaging in linguistic practices positioned as normative or innovative when produced by privileged white subjects" are produced (2015, p. 150). This perspective is particularly useful in the context of standard language ideology in the British context as it explicitly engages with the ways in which both Standard (British) English and many of the institutions tasked with transmitting it are tools of a colonial power (Cushing and Snell 2022). It also draws attention to the way listeners (and their ideologies) influence speakers (and their productions) (as discussed by, e.g., Inoue 2003; Pak 2021; Cushing and Snell 2022). This critique is particularly important because it reminds us that the cause of linguistic discrimination is *not* the way people use language but the structures of oppression which shape the way they can live. Changing these ways of speaking is unlikely to be enough to escape these structures because in an important way it is not *about* language (Rosa and Flores 2017). In the context of employment discrimination, Anderson et al. (2014), who, it should perhaps be noted, are not linguists, conclude that women should avoid creaky voice to avoid discrimination. Similar advice is often given to anyone who does not speak the "standard variety" (Craft et al. 2020), giving rise to, for example, an industry of "accent reduction" (Ramjattan 2022). This advice of "self-improvement" to become a "better speaker" (see also Silverstein 1996) is the wrong conclusion to draw. It is, in my opinion, normatively wrong as it does not change any of the structural barriers and constraints faced by marginalised speakers and, on the contrary, actively sustains the existing oppressive social structure. Rosa and Burdick also emphasise the limits of trying to change attitudes about language to achieve greater

justice and equality. Sociolinguistic work has often been motivated by an explicit or implicit desire to address linguistic discrimination. Understanding and explaining (to lay and academic audiences) the origins and uses of linguistic variation, emphasising in particular their "naturalness" and "order" has had some success (Charity 2008). However, this approach only goes so far, as it does not address any of the reasons of linguistic discrimination in the first place (Hudley and Flores 2022).

The concept of raciolinguistic ideologies was initially developed within the context of bilingual education in the US to capture the way racialised students' linguistic practices were always framed (and heard) as deficient (by the white listening subject) (Flores and Rosa 2015). One of these linguistic practices consistently heard as inferior when produced by Latino/a students was "translanguaging" (Flores 2019; Flores and Rosa 2015). Otheguy et al., who popularised the term in its current use, define translanguaging as "the deployment of a speaker's full linguistic repertoire without regard for watchful adherence to the socially and politically defined boundaries of named (and usually national and state) languages" 2015, p. 281. Raciolinguistic ideologies, and conceptions of standard language (at least in the context of English in the UK and the US), are *monoglossic* (Silverstein 1996; Flores and Rosa 2015). That is languages may not be "mixed". Implicit in this view is the assumption that languages are bounded and named linguistic objects (e.g., Spanish and English). As Otheguy et al. (2015) explain distinctions between named languages (and our ability to identify and name them in the first place) are entirely cultural and political, and historically contingent. Our observations of language as an object of study for linguists, are observations of speakers' idiolects rather than "a language" (Otheguy et al. 2015). Despite the long history and importance of this distinction in the development of modern theoretical and applied linguistics, its full implications, Otheguy et al. (2015) argue, are often set aside. In addition to being perceived as bounded, languages are often explicitly and implicitly framed as "belonging to" nations or nation states (Schneider 2019).

In this thesis, I make the case that language technologies are not only shaped by language ideologies and language management "upstream" but also have a "downstream" effect as they contribute to and entrench existing beliefs about speakers and language(s). For example, most ASR tools assume users who speak a standard variety. As a result, they often work poorly when speakers draw on multiple languages in the same interaction. In this way, these technologies suggest to users that translanguaging (or code-switching) is inappropriate. Where ASR tools produce a written transcript, they further (literally) reproduce an existing standard.

2.4 "Ethics" of ASR: understanding bias and harms

To understand the "harms" of these effects, and the wider ethical challenges posed by speech technology development, I draw on wider literature from AI ethics. In this way, I hope to be able to situate this work on ASR, within a broader research agenda and draw attention to the

complex social contexts of language within the AI ethics field. Here I provide an overview of recent work and debates in study of algorithmic fairness and algorithmic bias to motivate the approach I take in the studies presented in the following chapters.

I introduce the term "ethics" in quotation marks here, to draw attention to the fact that what we refer to as "AI ethics", or "ethics of language technologies" is rarely grounded in rigorous ethical analysis (drawing on ethics as a field of scholarship). Rather, it is frequently a "catch-all" term used to cover any and all scholarship which concerns itself with the social contexts and consequences of AI or language technology research. Similarly, notions of "bias", "harms" and "risks" are highly contested within this space (and elsewhere). I endeavour to provide a critical overview below and clarify how I approach these "contexts" of language technologies.

2.4.1 (Beyond) Fairness, Accountability and Transparency

Scholarship on the ethics of sociotechnical systems encompasses a wide range of perspectives, most notably from philosophy of technology (e.g., Anderson and Anderson 2011; Vallor 2016). In this thesis, I am instead drawing most heavily on notions of "fairness", "accountability" and "transparency" with a particular focus on algorithmic bias and harms. This focus has been popularised by the conference on Fairness, Accountability and Transparency (FAccT, formerly FAT-ML) which was first established as an independent conference in 2018 and has since become one of the largest venues for AI ethics research.

In an analysis of all papers published at FAccT between 2018 and 2021, Laufer et al. (2022) show that most "fairness" research was concerned with outcome disparities "across socially salient groups", while most "accountability" research focused on self-governance. This highlights two common critiques of this type of research: a tendency to focus on abstract notions of "bias" and "fairness", rather than harms as experienced by individuals which are rooted in broader social structures, and a tendency to present transparency and self-governance as a (sufficient) solution to this problem.

Through a focus on abstracted notions of unequal outcomes, Birhane et al. (2022b) argue, the understanding of "ethics" in AI ethics (too) often reproduces Western approaches to ethics in a way which abstracts away from people's lived experience with AI. Birhane (2021) argues instead for a "relational ethics" approach to what she terms "algorithmic injustice". This approach, rather than privileging the hegemonic Western rationalism, draws on "relationality" as theorised and practised by different schools of thought (including afro-feminism and complexity science) (Birhane 2021). At the heart of this perspective lies a focus on "interdependence, relationships and connectedness", and a rejection of the rationalist quest for "timeless and absolute knowledge" predicated on a "rational, static, self-contained, and self-sufficient subject" (Birhane 2021, p. 3). Instead, a relational ethics approach to algorithmic bias (and injustice), urges both breadth and depth of perspective. Away from abstracted metrics, it encourages us to consider both the broader deployment and development contexts of a system, and the specific ways it affects and interacts with people (Birhane 2021). Feminist science and tech-

nology studies have long pointed out the fundamental impossibility of the kind of disembodied objectivity (implicitly assumed or explicitly asserted) in rationalist science (Haraway 1988) and, more recently, machine learning (Talat et al. 2021). Building on this "situatedness" as established and popularised by Haraway (1988) in the context of "situated knowledges", Ehsan et al. (2022) propose "situated fairness". This concept of fairness centres the "people who live with [an algorithmic system]" (Ehsan et al. 2022). Rather than reducing the notion of an "unfair" or "fair" system to a simple mathematical score and narrow definitions of bias, this perspective can allow us to focus on harms and imagine a just (rather than "fair") system by scrutinising the wider origins, consequences and logics of a technology. As Hoffmann (2019, p. 901) puts it: "[I]n mirroring some of antidiscrimination discourse's most problematic tendencies, efforts to achieve fairness and combat algorithmic discrimination fail to address the very hierarchical logic that produces advantaged and disadvantaged subjects in the first place. Instead, these efforts have tended to admit, but place beyond the scope of analysis important structural and social concerns relevant to the realization of data justice."

In this thesis, I apply this "relational ethics" perspective laid out by Birhane (2021) by considering the impacts of ASR in a range of different (breadth) but specific (depth) contexts. I focus directly on language communities' experiences of language technologies – anchored in a particular time and place as algorithmic systems (and languages and communities) are dynamic. I also want to emphasise the embodied, material and situated nature of language (data) and language (data) practices.

In order to uncover and ideally mitigate or rectify "unfair" behaviour (however defined), governance structures and processes are required. One such process is the "audit", "a [tool] for interrogating complex processes [...] to determine whether they comply with company policy, industry standards or regulations" (Raji et al. 2020, p. 34). Drawing on audit structures in other domains (aerospace, medical devices, finance), Raji et al. (2020) propose audits conducted by internal audit teams. This approach is in contrast to external audits conducted after deployment by external parties such as independent researchers or journalists ("third party" audits), or contractors ("second party" audits) (Costanza-Chock et al. 2022), and "co-operative" audits where developer teams cooperate with an independent auditor team (Wilson et al. 2021, e.g.,). Especially in the context of commercial AI tools, external auditors generally do not have access to relevant information such as models and training data which are usually proprietary. While external auditors can scrutinise the output and behaviour of the system "from the outside", it can be very difficult to understand the origins of this behaviour without this "insider" information. As Raji and Buolamwini (2019) show, these "black box audits" can be effective, when commercial providers respond to public audits with changes to their systems. However, depending on the application domain and service, naming providers of commercial AI tools in an audit study might be a breach of the terms of service or be otherwise considered an unfair practise (Raji and Buolamwini 2019). While audits thus can be a pathway to accountability, additional factors such as robust internal processes, external regulation and legislation and

public pressure are necessary to ensure that audits are conducted in an appropriate manner taking into account a wide range of stakeholders and real-life impacts, and that action is taken as a consequence of the findings. Metcalf et al. (2021) discuss algorithmic impact assessment (AIAs) as another, related governance practice, where "impacts are proxies for sociotechnical harms" (2021, p. 735). However, as they caution, "impact assessments of an algorithmic system do not produce accountability unless the methods used to determine impacts are submitted to a forum that has the ability to mandate changes in the implementation of sociotechnical systems (or provide remedy for harms)" (Metcalf et al. 2021, p. 736). Recent work within the "FAccT" research community has critiqued the prevalent but very often counterfactual assumption that such governance structures exist (Gansky and McDonald 2022; Laufer et al. 2022).

In addition to Birhane (2021)'s critique, broader frameworks like "Data Feminism" (D'Ignazio and Klein 2020) and "Design Justice" (Costanza-Chock 2020) aim to foreground an analysis of power in understanding harms of technology design. D'Ignazio and Klein (2020) develop principles based on feminist theory which can be directly applied to data science, and by extension machine learning tools. These principles ("elevating embodied knowledge", "examining and challenging power", "considering context" and "embracing pluralism", "rethinking binaries and hierarchies" and "making labour visible") address some of the critiques raised in response to current discussions on "fairness", "bias" and "harms". Design Justice, as formulated by the Design Justice Network and presented by Costanza-Chock (2020), focusses on community-led practices for "just" rather than "fair" technology design[14].

2.4.2 (Beyond) Algorithmic bias

Friedman and Nissenbaum (1996) present a foundational discussion of different forms of "bias" in computing. Like much of the work that has followed, they present a taxonomy of algorithmic bias (see e.g., Mehrabi et al. 2021). Taking the origin of the bias as a starting point, Friedman and Nissenbaum (1996) distinguish between "preexisting", "technical" and "emergent" bias, and highlight the potential impacts of bias embedded in pervasive computing systems. "Preexisting" biases as they define them, can enter computing systems through individuals or social institutions which already hold them (Friedman and Nissenbaum 1996). These are different from "technical biases" which are the result of technical constraints (e.g., a pseudo-random number generator may be biased) (Friedman and Nissenbaum 1996). Lastly, "emergent" biases arise when computing systems are applied in new or different contexts from the ones they were developed for, for example because of changes in society (Friedman and Nissenbaum 1996). While much of the literature on "bias" has focussed on limiting the influence of "preexisting" bias and examining the impact of "emergent bias", as Dobbe et al.

[14]The distinction between "fair" and "just" systems crucially depends on the deployment context and is frequently discussed in the context of machine learning systems deployed in "high-stakes" contexts like loan distribution (e.g., Kasirzadeh 2022). In the context of Natural Language Processing, Blodgett (2021) and Nee et al. (2021) explore how mitigating algorithmic bias can advance social justice.

(2018) point out "technical bias" is also "an issue of epistemology". Decisions about how to measure and label training data points, what and how to optimise in model training, and how to evaluate systems, are all to some extent shaped by the "preexisting" biases. Focussing more specifically on natural language processing (NLP), Shah et al. (2020) set out to create a framework of "predictive biases". A "predictive bias" as they define it manifests as either systematic "outcome disparity" (outcomes differ with respect to an "ideal" distribution) or "error disparity" (error rates differ between groups) (Shah et al. 2020). Shah et al. (2020) also posit four "origins of bias": "label bias" (how the data is annotated), "selection bias" (which data is selected), "semantic bias" (how the data represents different groups) and "over-amplification" (how existing biases are amplified through the model training or structure).

To the extent that it is possible, separating out these different types of discriminatory system behaviours can be useful to identify exactly where "bias" entered the system. This knowledge could help people designing and interacting with algorithmic systems to anticipate, prevent or mitigate the harms of discriminatory system behaviour. However, there are real limits to the power of "bias" as a conceptual framework to understand the harms of algorithmic systems. One such limitation is the tendency to understand bias to have a "discrete source", like a biased individual (e.g., an engineer), or biased (or "bad") data and algorithms (Hoffmann 2019). The flip-side of the desire to pinpoint the "bad mechanisms" giving rise to bias in order to mitigate them, is the tendency to set aside bias where no such mechanisms can be easily identified (or fixed) (Hoffmann 2019, p. 905). The idea that unintentional or unconscious bias shapes decisions of individual engineers and historical bias shapes what kind of data we have is widely accepted at this point. But even so, as Hoffmann points out, what is lacking is an acknowledgement that, rather than being unfortunate "accidents" of history, people do actively maintain the underlying oppressive social structures – sometimes using algorithmic systems. Hampton (2021) argues that "algorithmic oppression" as coined by Noble (2018) is a useful concept to displace (algorithmic) "bias" because it emphasises that algorithmic systems are often deeply harmful to marginalised communities in ways that are systemic, structural, and intentional. A second limitation Hoffmann (2019, p. 905) identifies, is the tendency towards "single-axis thinking centred on disadvantage", which fails to (meaningfully) engage with the fact that structures of oppression interact (Hill Collins 2000 [1990]; Crenshaw 1989; Cooper 2016) and belies how advantages are created for already privileged groups.

In practice, the notion of bias in language technologies is often used in a highly decontextualised way. In their review of 146 papers on "bias" in natural language processing, Blodgett et al. (2020) find that many of them lack the normative reasoning and socio-cultural context necessary to meaningfully engage with the harms of "biased" systems. Similarly to Birhane et al. (2022b)'s analysis of the AI ethics literature, Blodgett et al. (2020) find that much of the existing literature fails to explicitly name who is harmed in what ways, focussing instead on broad categories (e.g., "racial bias", "gender bias") which might be defined in contradictory ways across different papers (if they are defined at all) (see also Field et al. 2021; Stanczak

and Augenstein 2021). By abstracting away from specific contexts, or failing to engage with the specific social or sociolinguistic context, much of this literature also fails to engage with power and broader social structures.

2.4.3 Algorithmic harms

Perhaps a more productive approach is not to focus (just) on "bias" but rather on the harms, the actual material or symbolic damage "biased" and "unbiased" systems do in conjunction with broader social structures. A slightly higher-level view helps us see potential negative impacts of ML tools beyond a narrow definition of "bias". While "bias" is important, so is the broader context of the way ML systems are built and deployed. Not all harms originate from "bias" and, consequently, not all harms can be mitigated by focussing on bias. As Powles and Nissenbaum (2018) put it: "bias is real but it's also a captivating diversion".

Suresh and Guttag (2021) present a framework which links different kinds of biases which can enter the machine learning system at various points, with harms. This framing, while still drawing on the "taxonomy of biases" approach discussed above, draws attention to harms to real people rather than undesirable system behaviours, and shifts our attention towards minimising and mitigating harm, rather than bias. Shelby et al. (2022) present a review of such "taxonomies" of such sociotechnical harms, emphasising the way harms are created through the interaction of technical and social systems. They identify 5 types of harms discussed in the academic scholarship on the topic: representational, allocational, quality-of-service, interpersonal and social (Shelby et al. 2022, p. 6).

Particularly popular in the field are notions of "representational" and "allocational" harms, often attributed to Kate Crawford and since further developed by many scholars (e.g., Barocas and Selbst 2016; Barocas et al. 2019; Weidinger et al. 2022). Allocational harms include what Shelby et al. (2022) term "opportunity loss" and "economic loss" – discriminatory distributions of material resources and opportunities. Examples are the outcomes of algorithmic bias in the distribution of loans, employment, housing, education and healthcare (as discussed by Eubanks 2018). Economic loss can also include discriminatory pricing algorithms and demonetisation of content and exclusion of individual content creators on social media based on algorithmic content moderation (Duffy and Meisner 2022). The application of ML in these high-stakes scenarios has received a lot of (warranted) attention in recent years. Not only is the risk of direct harm very high in these domains, but the underlying infrastructures and institutions where they are applied often have long histories of (non-algorithmic) discrimination. These legacies shape the training data and deployment contexts of ML systems, both of which can create harms.

"Representational harms" describe the fact that algorithmic systems can cause harm by reproducing and reinforcing existing structures of oppression and further the marginalisation of already marginalised groups by, for example, reinforcing harmful stereotypes (Barocas et al. 2019; Benjamin 2019b; Noble 2018; Wang et al. 2022). In her foundational work on this

topic, Noble (2018) describes the way search engines maintain, legitimise and re-entrench white supremacy and patriarchy by surfacing racist and sexist results (first)[15]. Shelby et al. (2022) identify six types of representational harms: stereotyping, demeaning social groups, erasing social groups, alienating social groups, denying people the opportunity to self-identify, and reifying essentialist social categories. These harms are, in some ways, more pernicious than allocational harms. Their effects can be long-lasting and not necessarily "fixable" as they feed into ideologies and discourses about marginalised groups. As Shelby et al. (2022) point out, the use of demeaning language and stereotypes about oppressed groups, can be understood through Patricia Hill Collins' concept of "controlling images" (2000 [1990]), a set of discourses which create and reinforce an oppressive social type or figure. In addition to wrong and bad "representation", the absence of representation also creates harm. Whole social groups can be "written" or "engineered out" of a system – in the same way that they can and have been written out of, for example, literary and scholarly canons – not only normalising and entrenching their "invisibility" to other social groups but also making it much harder for them to use and navigate algorithmic systems. Closely related is alienation of social groups who are "reminded", often quite violently, that they or their perspectives on the world are excluded from the mainstream culture.

Alienation is also what Shelby et al. (2022) call a "quality-of-service" harm. It includes the psychological harms being "misrecognised", as noted by Mengesha et al. (2021), Rincón et al. (2021) and Ahmed (2017) in the context of speech technologies. Other harms related to the reduced "quality-of-service", or "degraded service" (Barocas et al. 2019), include "increased labor", when having to compensate for said poor performance of an automatic system with manual corrections and "service or benefit loss" when poor performance completely negates any benefits of the system.

Interpersonal harm occurs when "algorithmic systems adversely shape relations between people and communities" (Shelby et al. 2022, p. 12). This includes loss of agency and social control, technology-facilitated violence, diminished health and well-being, and privacy violations. Loss of agency and social control includes instances where algorithmic systems are applied to people (rather than used by them) against their will, or are required to access basic services (e.g., housing) (Shelby et al. 2022). Technology-facilitated violence includes (but is not limited to) violence on digital platforms and violence enabled through digital devices (e.g., GPS trackers) (Shelby et al. 2022). Other types of sociotechnical harms might be accompanied by diminished (mental and physical) health and well-being. Privacy violations include both more conventional breaches of personal data, and, increasingly, (correct or incorrect) inferences about individuals based on existing data enabled by an algorithmic system (Shelby et al. 2022).

Finally, Shelby et al. (2022) refer to the way sociotechnical systems can destabilise societies and exacerbate social inequality as "social" or "societal" harms. As discussed in the context of

[15] As Shah and Bender (2022) also highlight, there are more fundamental problems with the way internet search is conceptualised today, which may create further harms.

other harms, algorithmic systems can and do reproduce existing power structures and inequalities, for example through allocational and representational harms. More specifically, Shelby et al. (2022) identify 5 types of social harms: informational harms, cultural harms, political and civic harms, macro-socioeconomic harms, and environmental harms. Informational harms include misinformation and disinformation, and the "subjugation" of particular discourses to the benefit of already hegemonic discourses, while cultural harms relate closely to the wider societal impact of representational harms and with the entrenchment of these same discourses (Shelby et al. 2022). The ways algorithmic systems are used to disempower or disenfranchise individuals and groups in a political and civic context, can be understood as "political harms". As pointed out by Crawford (2022) and Taffel (2021), machine learning is *extractive*, requiring both ever-large amounts of resources: energy, data, minerals, labour. The environmental impacts of ML technologies remain are slowly starting to receive the attention they deserve, with significant harms caused by mining and manufacturing and huge carbon footprints associated with training and deploying machine learning systems (Crawford 2022). "Macro-economic" harms include deteriorating working conditions for some workers[16], and already bad working conditions for others. Scholars like Lily Irani, Mary L. Gray and Siddarth Suri powerfully describe the conditions in which thousands of workers already perform the "data work" without which ML would be impossible (Gray and Suri 2019; Irani 2013). Annotating large datasets to train and evaluate ML models across domains is generally precarious, underpaid but ultimately crucial "cultural work" (Irani 2013) outsourced from the Global North where the technology is designed to workers in the Global South. The impacts of the way this work is done on workers should be a central ethical concern in the context of machine learning technology.

2.5 Understanding the impact of ASR

As laid out above, a central discussion in AI ethics research is *what* we should be focussing on. A focus on *bias* risks ignoring *harms* or the wider *context* (Blodgett et al. 2020; Siapka 2022; Powles and Nissenbaum 2018). Similarly, *fairness* as conceptualised as simple distributive justice might neglect to ask questions about how and why a system is deployed and whether making it "fair" is the same as "justice" (Kasirzadeh 2022; Birhane 2021; Birhane et al. 2022b; Ehsan et al. 2022; Hoffmann 2019). A single-minded focus on a narrow definition of bias and fairness outwith the broader context can lead to absurd outcomes: (in an only somewhat satirical paper) Keyes et al. (2019) show that a fictional algorithm selecting people to, essentially, be killed ("mulched"), can pass a range of "fairness" guidelines and checklists as long as no population group is more likely than any other to be selected.

Throughout this thesis, I try to avoid this narrow focus, and instead look at the wider impacts of algorithmic bias and ASR more broadly. In this section, I lay out this theoretical framing and

[16]While some tasks and jobs have been subject to automation and others could be, as Levy (2022) argues in the context of long-haul truckers in the US, the expansion of "AI" in the workplace often translates to a deterioration of working conditions and much reduced worker autonomy due to increased ML-facilitated surveillance.

briefly outline the chapters which follow.

2.5.1 Technical, contextual and infrastructural lenses

Ehsan et al. (2022) conceptualise these apparent tensions between different types of analysis as different "lenses": a *technical lens* and a *contextual lens*. "The technical lens finds the ethical consequences of algorithmic systems in the methods and technical details of how the systems are constructed, focusing interventions on best practices for constructing, modifying, or governing algorithmic systems" (Ehsan et al. 2022, p. 1307). The "contextual lens" on the other hand "finds the algorithmic ethical consequences in the social and political relations that the system interfaces with" (Ehsan et al. 2022). They then propose an *infrastructural lens* which "instead examines the social and political conditions that make the system possible in the first place"(Ehsan et al. 2022). This "infrastructural lens' might go some way to addressing some of the central critiques raised by Birhane (2021), Hampton (2021), Costanza-Chock (2020), and Hoffmann (2019) among others.

These three lenses represent three productive ways to think about the *contexts* (technical, social and infrastructural) of algorithmic systems. All three lenses are used in this thesis: I adopt a *technical lens* in my focus on data documentation (see Section 5.4 and Section 7.4), predictive bias in deployed systems (see Chapter 3) and data compilation methods (see Chapter 5 and Section 7.3). The *contextual lens* is central to the analysis of (explicit and implicit) language ideologies expressed in and through ASR development (see Chapter 6), and the close examination of application contexts and their impact on ASR design (see Chapter 4). Finally, like Ehsan et al. (2022), I am interested in the broader political, economic and social contexts which create the possibility (and perhaps necessity) for ASR systems in the first place. I apply an *infrastructural lens* when exploring different ways speech technology development can be organised and oriented (see Chapter 5 and Chapter 6) and drawing attention both wider structural constraints affecting developers *and* the agency they have in navigating those constraints to build more socially just technologies (see Chapter 5 and Chapter 7).

Ehsan et al. (2022) also propose the concept of the "algorithmic imprint" to describe the effect an algorithmic system both during and (crucially) after its deployment. These effects, they highlight, occur at several levels: infrastructural (e.g., data compilation practises required for the algorithm), social (e.g., how the algorithm mediates human relationships) and individuals (e.g., how people make sense of the algorithm and its output) (Ehsan et al. 2022, p. 1306).[17]

[17] They develop this framing using a specific example: in 2020, millions of students were unable to sit GCE A Levels (a UK-based internationally recognised high-school exam) due to the Covid-19 pandemic and initially received "algorithmic grades" (grades based on teachers' predictions, adjusted based on several factors including historical school-level results) which were eventually replaced with teacher assessments after protest from students, parents and teachers (Ehsan et al. 2022). In addition to unclear communication around the implementation of the algorithm, two assumptions of the system were particularly heavily criticised: that past performance would be indicative of exam performance, and that no two students could be equally good as teachers had to rank them without ties. Ehsan et al. (2022)'s fieldwork shows that the former is particularly problematic when applied across different education systems and learning cultures, as students in Bangladesh tend to prepare for exams during a short period leading up to them, rather than over the course of a year as is common in the UK. It is notable, (and perhaps not

We can use the notion of the imprint to draw attention to the way a (particular type of) ASR system intervenes in the pre-existing sociolinguistic context. At the infrastructural level, we see the framing of language data as a "language resource" and the "collection" of that data in highly specific forms which are required by the ASR system. In Chapter 4, we see that these "data requirements" can be at odds with the way language communities and speakers use their languages. For so-called "low-resource" language communities, i.e., communities where these data requirements are not easily fulfilled, new data practices might be required, as I also discuss in the context of crowd-sourced speech corpora for technology development in Chapter 6. Similarly, data practices may change as requirements change – for example because of the (often intertwined) emergence of new algorithms and new application contexts alongside wider changes in the political economy of technology development as discussed in Chapter 5. On a social level, ASR tools reconfigure social relationships whether they are embedded in "high-stakes" tools like automatic assessment, or in voice user interfaces, automatic transcription and dictation devices where they could improve accessibility and usability and enhance social relationships between people (see Chapter 4 in the context of isiXhosa) and potentially create new social relationships between people and machines (Schneider 2021; Porcheron et al. 2018). Finally, on an individual level, Ehsan et al. (2022, p. 1306) write that "algorithms can leave imprints on how people make sense of algorithmic operations and interpret their lives experiences with the algorithm, carrying deep psychological impact on their mental well-being." One of the arguments I develop throughout this thesis is that both the availability of speech technology in a particular language variety, and its performance for different (kinds of) speakers can impact both the status of the variety, and, more importantly in my view, speakers' self-perceptions. In this way, speech technologies can re-produce existing language ideologies and linguistic hierarchies – by performing better for high-status varieties (see Section 3.3), imposing aspects of spoken and written standards (see Chapter 4) and reifying broader ideologies around language variation and language (see Chapter 6).[18]

2.5.2 How does predictive bias in ASR affect speakers?

Chapter 3 explores the extent of *predictive bias* in commercial ASR systems for British English by comparing performance between different regional varieties, and between L1 and L2 speakers of English. I present these findings against the background of research on accent bias and linguistic discrimination in the UK and argue that predictive bias reflects and entrenches "pre-existing" biases to some extent. I then turn to the harms of predictive bias in ASR systems.

---

surprising) that the disproportionate harm to students in the Global South was in part due to the way officials in the UK expected that assumptions about learners and teachers which may be reasonable in the UK are "universal".

[18]While Ehsan et al. (2022) consider how these effects persist *after* the algorithm has been removed, I pay attention to the impact ASR has both *during* deployment and, to a lesser extent, how the imprint is "carried" to other contexts by speakers and language communities.

2.5.3 What is the role of automatic transcription in standardisation?

Chapter 4 explores how the task of automatic transcription interacts with language standardisation in two different application contexts. I first look at how ASR can be used to transcribe English language audio and video recordings collected in the Lothian Diaries Project (Hall-Lew et al. 2022) and in sociolinguistic research more broadly. I then turn to the very different context of developing an ASR tool to transcribe informal voice messages for and with the isiXhosa speaking community in Langa, South Africa. Both contexts raise important questions about the purpose of transcription, orthographic practises, standardisation and language data.

2.5.4 How are speakers and language varieties reflected in speech datasets?

In Chapter 5, I focus on the "language resources" used to build ASR systems. I first draw attention to *gaps* in commonly used datasets – and the need to interrogate their origins. I then consider how language policy can be applied to clarify the role of developers dealing with such gaps. Exploring how speakers and language varieties are reflected in language resources raises important issues about the values and goals underpinning language technology development.

2.5.5 What is the relationship between language ideologies and ASR development?

In Chapter 6, I take a closer look at language variation and linguistic diversity is conceptualised in academic research and discussed by prominent commercial and open-source language technology developers. This provides an insight into how the language ideologies held by developers may influence development. A close analysis of the way corporate developers discuss language variation in the context of their products also allows us to understand their language management.

2.5.6 How can speakers and developers build better ASR tools?

Finally, in Chapter 7 I turn to the question "How do we build better ASR systems?" Here I first present work on a "diverse" multimodal dataset of English language conversations, the design and documentation of which was inspired by best practices in sociolinguistics and machine learning as discussed in Chapter 5. I then discuss some limits of focussing on "bias" and "fairness".

Chapter 3

Predictive bias and harms in Automatic Speech Recognition

3.1 Introduction

In this chapter, I explore performance disparities of commercial automatic speech recognition systems for different speaker groups. Before presenting a quantitative and qualitative evaluation of British English commercial ASR systems, I contextualise this research in Section 3.2 by outlining prior work on this "predictive bias" in speech and language technologies and ASR specifically. I then provide an overview of the sociolinguistic context in Britain, with a particular focus on variation in English and linguistic discrimination.

In Section 3.3, I present a mixed-methods evaluation of two commercial British English ASR systems conducted using two corpora of read speech. I highlight performance disparities for different varieties in both commercial systems, with better performance for first language speakers than second language speakers, and speakers from the South of England than those from other regions of the British Isles.

In Section 3.4, I reflect on the results of the evaluation. After reflecting on ways of measuring bias and categorising varieties for the purposes of such audit studies, I draw attention to some of the potential and reported harms of predictive bias.

Much of this chapter (and this thesis) focusses on English. The reason for this focus is not that these issues are either more interesting or more important in the context of English than they are in other languages. The purpose of this chapter is not primarily to present those performance differences without context – as Blodgett et al. (2020) note, there is already a large volume of research doing just that. While highlighting these differences is one novel contribution of this chapter (as predictive bias in ASR is generally underexplored), the larger point I aim to make is that these performance differences interact with existing sociolinguistic hierarchies and linguistic discrimination. I use the example of English (both in the British and the global context) because it is the context I am personally most familiar with, and it is the context

which has been best explored in the prior literature on predictive bias in speech and language technologies. This disproportionate focus on English is, of course, itself the result of (sociolinguistic) hierarchies in technology development and academic research in sociolinguistics (see e.g., Heller and McElhinny 2017) and technology development (see e.g., Srinivasan 2017). I do believe that many of the key insights on the way linguistic discrimination interacts with algorithmic bias also apply in other contexts, though as I discuss in Chapter 4, in different contexts we also see other kinds of harms arising from both biased and unbiased speech technologies (see also Chapter 7).

3.2 Background

3.2.1 Predictive bias and language variation

As discussed in Section 2.4.2, defining "bias" or "algorithmic bias" is not straightforward. Here, I adopt Shah et al.'s notion of "predictive bias", which they define as: "[a] label distribution of a predictive model reflects a human attribute in a way that diverges from a theoretically defined 'ideal distribution' " (2020, p. 5248). They posit that we observe predictive bias whenever there is "outcome disparity" or "error disparity" between groups (Shah et al. 2020). The former refers to systematic differences between the distribution of predictions and an ideal distribution: for example, an image captioning system where "captions overpredict females in images with ovens and males in images with snowboards" (Shah et al. 2020, p. 5250) as identified by Hendricks et al. (2018). Error disparities are present when "model predictions have larger error for individuals with a given user attribute" (Shah et al. 2020, p. 5250), with a classic example being the "Wall Street Journal Effect" as discussed by Hovy and Søgaard (2015) where the error rate of part-of-speech taggers is related to how demographically similar an author is to the authors featured in the Wall Street Journal in the 1980s and 1990s on which the taggers had been trained.

While predictive bias looks differently depending on the specific task, the way language (variation) conveys and constructs social meaning (as discussed in Section 2.3.1) is crucial to understanding both the *origins* and *consequences* of predictive bias. In some contexts, such a sentiment analysis, predictive bias is in part due to the fact that the same surface form (word, phrase) can have different social meanings depending on (among other things) the identity of the speaker (and the listener). In others, such as ASR, different surface forms (different pronunciations) also carry social meaning as they act "as a proxy" for or marker of a range of macro- and micro-social categories identities. If these (aspects of) speaker identities, the "user attributes" (Shah et al. 2020), are under- or misrepresented in the training data (with respect to the application context), we can observe worse performance for the people who hold them. I will illustrate the different ways predictive bias surfaces below. I first focus on some applications from text-based language technologies (Natural Language Processing) where predictive bias is an increasingly popular topic of research. I then turn to the specific context of ASR, where the

topic has received less attention.

Predictive bias in Natural Language Processing

Predictive bias can, for example, be seen in sentiment analysis and hate speech detection. Dias Oliva et al. (2021) 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 Dixon et al. (2018)), 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.[1] Ungless et al. (2023) conduct a large-scale quantitative analysis of six (different) sentiment analysis tools commonly used in social science research. Instead of using existing social media posts, they construct a large data set comprising almost 30,000 sentences with the same template structure to probe how different queer identity terms (and by extension, people who use them) are affected by algorithmic bias (Ungless et al. 2023). Their findings suggest that some commercial sentiment analysis tools avoid producing obviously biased outcomes by effectively ignoring identity terms such as *straight*, *gay*, *queer*, *asexual* (Ungless et al. 2023, p. 9). The surface-level nature of this approach, does, however, mean that infrequent identity terms (like *demisexual*) are less likely to be ignored, and that there are still differences between different subgroups (e.g., *bisexual woman* receives a lower score than *bisexual man*) (Ungless et al. 2023). They also note that while systems provided by Google and Amazon appear to operate with such an "ignore-list", the IBM system tested does not and exhibits significant bias towards transgender identities as opposed to (explicit and assumed) cisgender ones (Ungless et al. 2023, p. 10). They also find that one of the tested systems gave more negative scores to longer sentences, thereby penalising people who mention marked identities inadvertently. Here too, the social context of a particular utterance is important: as Ungless et al. (2023) point out, while the use of an "ignore-list" seems useful in mitigating bias, it also means that a sophisticated sentiment analysis model cannot access a word like "gay" or "queer" in the utterance to contextualise the use of words which differ in their sentiment (and meaning) depending on who uses them (e.g., "sickening" or "fierce" used as positive terms within queer communities, (Calder 2019)). Which and whose language is considered "obscene" is an ideological choice imposed by the hearer (Spears 1998). Similarly, how words or sentences are labelled as "positive" or "negative" in training data for sentiment analysis tools is quite subjective (Mohammad 2017) and affected by annotators' biases (Sap et al. 2019).

Large language models are prone to reproduce structural oppression in a very direct way by "parroting" biases in the training data (Bender et al. 2021), for example islamophobic content (Abid et al. 2021; Dodge et al. 2021; Li et al. 2020). When large language models are used to

[1] A problem not limited to language, several art museums have had images of their exhibits, including the 25,000 year old Venus of Willendorf figurine, flagged as "pornographic"(Hunt 2021).

generate language, they have been shown to produce harmful stereotypes and "exclusionary social norms" (Weidinger et al. 2022). It is worth noting that broader (potential) harms of large language models, like those related to climate change, misinformation and radicalisation are also likely to disproportionally affect marginalised groups (Weidinger et al. 2022; Bender et al. 2021).

Similar problems also exist in machine translation. Gender neutral nouns or pronouns are often translated reflecting stereotypes or are simply ungrammatical (Savoldi et al. 2021; Dev et al. 2021). Machine translation also introduces stylistic bias, where translated text "sound[s] older and more male" than the original (Hovy et al. 2020). Gender bias is also a particularly persistent problem in word embeddings (Brunet et al. 2019), which can also interact with racial biases for example in embeddings of given names (Jiang and Fellbaum 2020).

Predictive bias in ASR

Performance of ASR systems is usually measured using the metric word error rate (WER) or character error rate (CER). WER captures the difference between the system output and a reference transcription and is defined as:

$$\text{WER} = \frac{S + D + I}{W}$$

where S, D, and I are the number of substitutions, deletions and insertions, respectively and W is the total number of words. CER is defined in the same way over characters, rather than words. As discussed in Section 3.4.1 and in Chapter 4, this metric, while established, has some significant shortcomings. Almost all prior work evaluates ASR systems in terms of word error rate or character error rate.

Adda-Decker and Lamel investigate the possibility of gender bias in ASR systems, noting that "in the early days of automatic speech recognition, female speech was widely considered as more difficult to automatically recognize than male speech" (2005, p. 2205). By 2005, they found that word error rates for broadcast news in French and English, no longer show worse performance for female speakers (Adda-Decker and Lamel 2005). On the contrary, they note that women's speech is assigned lower error rates in both languages. They find the same pattern in conversational telephone speech (Adda-Decker and Lamel 2005). Two years later, Benzeghiba et al. (2007) reviewed the existing literature and highlighted performance differences depending on speakers': first and second language varieties, age, speech style, speech rate, emotional state, gender, and physiology. Young and Mihailidis (2010) point out that ASR systems are also much less reliable for elderly speakers and those with dysarthria and other speech disorders. More recently, and after several key technical innovations improving overall ASR performance and integration of ASR into user interfaces, the focus of predictive bias research in ASR has shifted to commercial ASR systems targeted at end-users. Tatman (2017) evaluates automatic transcriptions of isolated English words on YouTube produced by 80 male

and female speakers across 5 dialect regions (California, Georgia (US), New England, New Zealand, Scotland). She finds higher error rates for some regions (notably Scotland), and for female speakers. While the gender effect was not replicated in a similar study of the same system, predictive bias was identified for different US varieties of English, and, in particular African American and mixed race speakers by Tatman and Kasten (2017). Koenecke et al. (2020) conducted one of the largest studies of predictive bias in ASR to date. They tested five commercial US English ASR systems (Apple, IBM, Google, Amazon and Microsoft) using two sociolinguistic corpora: the Corpus of African American Language (CORAAL) (Kendall and Farrington 2021) and the Voices of California Corpus (Stanford Linguistics n.d.). Both corpora contain sociolinguistic interviews, with Black speakers of African American Vernacular English[2] from three different US states, and with white speakers in two areas in California, respectively. As characteristic for sociolinguistic interviews, the speech style is relatively informal in both corpora. Koenecke et al. (2020) selected short audio snippets from the two corpora with similar distribution of age and gender to limit the impact of those factors in the analysis. They then analysed the performance disparities of the 5 ASR systems on the audio snippets and found that all systems had "substantially larger" error rates for Black speakers than white speakers (Koenecke et al. 2020, p. 7685). The systems performed particularly poorly for Black male speakers, showing a much larger gender gap among the Black speakers. Beyond the United States, some studies have considered predictive bias in ASR for second language speakers. DiChristofano et al. (2022) used the Speech Accent Archive (discussed in more detail below), to investigate differences in English ASR performance for speakers from different regions across the world. They also audit commercial systems provided by Microsoft, Google and Amazon and find that error rates are significantly higher for speakers whose first language is not English and those who were not born in the United States (DiChristofano et al. 2022). The age at which speakers started learning English and the number of years they had lived in an anglophone country also significantly affected error rates[3], with lower error rates for those who started learning English at a younger age and those who had spent more time in an English speaking country (DiChristofano et al. 2022). Chan et al. (2022) also analysed predictive bias in the English language Otter ASR system using the Speech Accent Archive, focussing specifically on the role of typological differences between speakers' first languages and found that it performs significantly worse for speakers whose first language was a tonal language. Age of onset of English acquisition was also a significant factor (Chan et al. 2022).

[2]Note that the terms African American Language (AAL), African American English (AAE), and African American Vernacular English (AAVE) are not necessarily interchangeable as used by different authors (as they relate to different positions on the origins of the variety and conceptualise the variety slightly differently). For a discussion of these terms, their use in sociolinguistics and their implications, see King (2020). In the context of this thesis, I always use the term used by the authors I am quoting.

[3]DiChristofano et al. (2022) do not use word error rate but word information lost (WIL) to measure performance.

3.2.2 Sociolinguistic approaches to predictive bias in ASR

An understanding of how language variation is distributed between and among different speaker groups, can be useful in illuminating the origins of predictive bias. In the case of the seemingly surprising "improvement" regarding gender bias in French and English ASR systems in the early 2000s reported by Adda-Decker and Lamel (2005), the distribution of speakers within the training dataset and their (to some extent gendered) differences in language behaviour turned out to be crucial. The French corpus featured interviews in which the interviewers were disproportionately women, and the interviewees disproportionately men (Adda-Decker and Lamel 2005). These different roles were reflected in the speakers speech styles with the women adopting a more formal style and the men's utterances being more unpredictable and informal (Adda-Decker and Lamel 2005). This tendency of the female speakers in this corpus to adhere more closely to the standard variety, also reflects wider, well-documented gendered differences where women generally speaking use more prestigious variants and varieties (e.g., Labov 1990). Koenecke et al. (2020) also explore the role of sociolinguistic variation to understand the observed differences not just between racial groups but also the notable gender difference between Black men and Black women. They use a dialect density measure quantify the use of syntactic and phonological features of African American Vernacular English employed by different speakers and find a positive correlation between the "dialect density" and ASR error rate. These dialect density measures are on average much higher for the Black men in the sample than the Black women, thus explaining the large gender difference in error rates (Koenecke et al. 2020). Also exploring ASR disparities affecting African American English, Martin and Tang (2020) take a closer look at how commercial ASR systems (fail to) handle AAE morphology and syntax. They find that utterances containing habitual *be*, a characteristic feature of AAE absent in Mainstream US English, are much more error-prone than those featuring other uses of *be*. Martin (2022) considers further examples of predictive bias in US English ASR systems related to AAE morphology. Chan et al. (2022) argue that the significant effect of whether or not a speaker's first language is a tonal language (on ASR performance in English) suggests that prosodic patterns may play an important role. This echoes Koenecke et al. (2020)'s finding that predictive bias persists even when the lexical content is controlled for, suggesting that only differences in phonology and prosody remain. Wassink et al. (2022) explore errors across speakers of four different ethnicities (and ethnolects) in a custom US English ASR model, paying close attention to individual phone errors. The painstaking analysis of errors shows that specific phonological differences between varieties can be pinpointed as sources for ASR errors. Choe et al. (2022) employ a similar phone-level analysis, relating typological characteristics of speakers' L1 to specific error types. For example, they note that speakers of languages with fewer phonemic vowel distinctions than English are likely to be affected by higher vowel substitution errors.

3.2.3 Linguistic variation and linguistic discrimination in Britain

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 (David M. Eberhard and Fennig 2021)[4]. English in the British Isles is also characterised by significant variation, conditioned both by region and social class (see e.g., Hughes et al. 2013; Foulkes and Docherty 1999; Wells 1982). 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 (Hughes et al. 2013). 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. A sizeable part of the population speaks English as an additional language. In 2011, around 7% of people living in England and Wales and 22% of London residents reported a language other than English as their main language[5]. Similarly, about 12% of people living in Glasgow, Aberdeen and Edinburgh reported speaking a language other than English (or Gaelic or Scots)[6]. As discussed in Section 2.3.2, varieties differ in social status related to the status of their speakers.

Linguistic hierarchies in Britain

Received Pronunciation In the British Isles (in particular in the UK), the classic example of a highly prestigious accent is Received Pronunciation (RP). 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 (Agha 2003; Fabricius 2018). 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) (Agha 2003). As Agha (2003) 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. More recently, the term Standard Southern British English (SSBE) has been replacing RP in the literature, and, as Fabricius

[4]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.

[5]`https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/language/articles/languageinenglandandwales/2013-03-04`

[6]`https://www.scotlandscensus.gov.uk/census-results/at-a-glance/languages/`

(2018) shows, like all language varieties, this "elite variety" too has been changing over time. The speech of upper-class social groups has had a curious position in the history of variationism as sketched out in Section 2.3.1. The speech of upper-class groups which is so often not just "the way upper-class people speak" but also very similar to the (idealised) "standard variety". As a result, researchers interested in the "vernacular", a kind of unaffected, "natural" way of speaking, have generally paid more attention to other social (class) groups. Fabricius (2018) argues this common conflation of the "standard variety" and the "elite sociolect", means that we overlook the fact that elite sociolects do change, and its speakers, just like all speakers, shift their linguistic styles contextually. To avoid this conflation, Fabricius (2018) suggests the distinction between "native-RP" and "construct-RP", where the former refers to the variety people socialised in a particular environment acquire, and the latter refers to the more abstract, idealised, standard variety transmitted through education and popular discourse. Agha is focused on this latter sense, showing how the variety has become "enregistered" and ideologically linked to a very specific social class position, and a "status emblem" (2003, p. 231). I highlight Fabricius (2018)'s work on variation and change in RP/SSBE here to avoid conflating these different senses of the term "RP", and emphasise that all "accents" are *both* collections of segmental and non-segmental phonological features (Crystal 2009; Trudgill 2006), *and* ideological constructs linking those linguistic features to a particular type of person (Agha 2003). I will revisit the question of what the term "accent" "means" in Chapter 6.

Language attitudes While the way RP and other accents actually sound has changed since the 1970s, public attitudes and linguistic ideologies, or the indexical links between language and perceived "systematic behavioral, aesthetic, affective and moral contrasts among the social groups" (Irvine and Gal 2000, p. 37) remain remarkably stable. How people in the UK evaluate regional and social accents and dialects has been a topic of study since the late 1960s. Giles (1970) conducted an early study where participants were asked to evaluate different accents (both vocal stimuli and simple accent labels) according to how "pleasant" and "prestigious" they sounded and how "comfortable" they would be interacting with a speaker of that accent. He found evidence for a linguistic hierarchy with RP at the top, and "industrial" accents (Birmingham, Liverpool, Cockney) and those associated with racialised minorities (West Indies, Indian) at the bottom (Giles 1970; Sharma et al. 2022). Coupland and Bishop (2007) conducted a replication study 35 years later using a much larger sample of UK respondents. Averaged across the entire sample, "Standard English" was still rated as most "prestigious" and most "pleasant" while "Birmingham" and "Asian" received some of the lowest scores on both scales (Coupland and Bishop 2007, p. 79). Perhaps unsurprisingly, social attractiveness and prestige are not strictly correlated for all varieties, with "Southern Irish English", "Newcastle English" and "Afro-Caribbean English" being rated higher for pleasantness than prestige, and "London English", "German-accented English" and "North American-accented English" being rated higher for prestige than pleasantness (Coupland and Bishop 2007, p. 80). The negative

attitudes towards varieties associated with working class speakers in the north of England, and with racialised minorities appeared unchanged then. Sharma et al. (2022) replicated this study another 15 years later using a participant sample representative of the UK population, thus being able to trace attitudes over 5 decades. Here too, the overall hierarchy is replicated, with RP being both "most prestigious" and "most pleasant", and Birmingham, Essex, Liverpool, Cockney, as well as "Chinese" and "Indian" towards the bottom of both scales (Sharma et al. 2022). Again some varieties are ranked higher on one scale than the other. Overall, the change observed over time is not in the ordering of the linguistic hierarchy, but in how "prestigious" and "pleasant" different accents are rated, slightly lower scores for RP and slightly higher scores for other varieties (Sharma et al. 2022). Particularly interesting for the purposes of this chapter, is the status of varieties which could be described as "migrant-heritage varieties" and "second language varieties". Giles (1970) included "West Indies", "Indian", "German", "Italian" and "French". The replication studies by Coupland and Bishop (2007) and Sharma et al. (2022) opt for "Afro-Caribbean" and "Asian"[7] and "Afro-Caribbean", "Indian" and "Chinese" respectively. Both also include "Spanish", "French" and "German". The evaluation by the participants (a demographically representative sample of the UK population) also shows that the "migrant-heritage" varieties receive low scores in terms of "prestige" (though Afro-Caribbean varieties are ranked as quite "pleasant") (Sharma et al. 2022). The way "migrant-heritage" and second language varieties are evaluated likely linked to wider attitudes regarding particular ethnic groups (and the way they are conceptualised by the British public). "French" and "Spanish" accents are generally evaluated as both pleasant and prestigious, while "German" accents are prestigious but not as pleasant (Sharma et al. 2022). Notably, however, all three "foreign" second language varieties (associated with Western Europe) are considered more "prestigious" than the migrant-heritage varieties. Evaluations are also affected by respondent age, with younger participants across all three studies "less conservative" and generally providing higher ratings for accents which aren't RP than older respondents (Sharma et al. 2022).

Linguistic discrimination It is worth highlighting that these attitude studies (as the authors all acknowledge) reflect attitudes which participants are not embarrassed to express, perhaps because they are expressing not just "their own" attitudes but trying to reflect the "correct", "common sense" attitudes towards these varieties. In other words, it is not at all clear that participants who indicate that RP is the "most prestigious" variety in the UK, also agree that it *should be*. Evaluations in terms of "social attractiveness" or "pleasantness" are perhaps more reflective of these personal attitudes. Rosa and Burdick draw a useful critical distinction between research on "language attitudes" and research on "language ideologies", noting that the former tends to "take at face value" what listeners *say about* language (2016, p. 105). They highlight instead the crucial importance of analysing language ideologies in a broader context

[7] Coupland and Bishop (2007) don't elaborate on their choice of these terms or how people likely interpret them. In the UK context, it is likely that most respondents associate "Asian" with L1 speakers of English of South Asian heritage.

and keeping in mind that all language and all identity is performative (Rosa and Burdick 2016).

Following Cushing and Snell (2022, p. 2), we can understand Standard (British) English (and, RP) as a "colonial and social construct which is designed by and based on the language of the powerful white bourgeoisie". Accepting this premise, the lower prestige associated with varieties associated with racialised minorities and working class speakers are somewhat intuitive. As Cushing and Snell (2022) highlight in their "raciolinguistic genealogy" of Ofsted, the British school inspectorate, language use of teachers and students is heavily policed in British schools. Following a deep-rooted pre-occupation with "Standard English", the inspectorate (and as a result, schools) positions any speakers outwith the "Standard English" norm as "deficient". The Standard English-focussed language policies and language policies in schools accounts perhaps for the persistence of the language attitudes first described more than 50 years ago.

Beyond the decontextualised evaluations discussed above, Levon et al. (2021) show that these attitudes are reflected in perceived "hireability" of (mock) job candidates who differ only in their accent, with study participants in England giving more favourable ratings to candidates speaking RP, and lowest to Multicultural London English and Estuary English. This effect is again shown to be affected by age with older participants much more likely to give lower scores to these two varieties associated with minority ethnic groups and working class speakers (Levon et al. 2021).

3.3 Predictive bias in commercial British English ASR

An abridged version of the contents of section Section 3.3 are published in:
Nina Markl (2022b). "Language Variation and Algorithmic Bias: Understanding Algorithmic Bias in British English Automatic Speech Recognition". In: 2022 ACM Conference on Fairness, Accountability, and Transparency. *FAccT '22. Seoul, Republic of Korea: Association for Computing Machinery, pp. 521–534. DOI: 10.1145/3531146.3533117*

The study presented in this section, consists of a quantitative and qualitative evaluation of two commercial British English ASR systems (Amazon Transcribe and Google Speech-to-Text), using two existing corpora of read speech: the Speech Accent Archive (SAA) (Weinberger 2015) and the Intonational Variation of English (IVIE) corpus (Grabe and Nolan 2002).

The research questions investigated in this study are:

1. What is the extent of predictive bias in British English commercial ASR systems for different regional varieties of English (as represented in the IVIE corpus)?

2. What is the extent of predictive bias in British English commercial ASR systems for second language speakers of English (as represented in the Speech Accent Archive)?

3. How does this predictive bias relate to existing linguistic discrimination and language ideologies in the British Isles?

The motivation for RQ1 was to fill a gap in the existing literature on predictive bias in ASR. Most of the research on predictive bias to date (and more so at the time of this study in 2021), focussed on the United States and US Englishes. As discussed above, varieties of English spoken in the British Isles differ dramatically by region and social class. As speakers from all backgrounds are increasingly likely to engage with ASR tools, understanding how well they work across different groups is important. The British Isles are also an interesting case study because of the well-documented public discourses and attitudes about language variation and in particular accent variation.

RQ2 is also seeking to fill a gap in the literature. Until very recently, discussions of predictive bias focused almost exclusively on "native speakers" (c.f., DiChristofano et al. 2022; Chan et al. 2022; Choe et al. 2022). The majority of people who speak English, speak it as an additional language learned after childhood. As ASR tools become ever more ubiquitous, the proportion of users engaging with them who are not first language speakers of English is increasing. Here, too, we can try to draw connections between ASR performance and broader attitudes and ideologies around language use and different speaker groups. As discussed in more depth in Chapter 6, second language speakers or "non-native speakers" are rarely the focus of ASR development or evaluation.

3.3.1 Data and methods

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[8]: Intonational Variation in English (IViE) (Grabe and Nolan 2002) and the Speech Accent Archive (SAA) (Weinberger 2015).

Read speech was selected to control the lexical content of the utterance to isolate effects of phonetic and phonological variation on ASR performance. Since ASR models also draw on a language model to decode utterances, factors unrelated to *accent* variation such as lexical frequency could influence ASR accuracy when comparing across utterances with different lexical content. Read speech does, however, differ from conversational speech and is likely to be more similar to a "standard" variety as produced by each speaker. Importantly for ASR (given the role of the language model), it is also more likely to consist of full and grammatical sentences than free conversation. The recordings used for evaluation here are not the most "realistic" examples for this reason. The word error rates observed in this study are likely lower than they would be for conversational speech by the same speakers.

[8]Data, code and analysis available at `https://github.com/ninamarkl/FAccT22_ASRBias`

IVIE corpus

The IViE corpus was collected around 2000 to study intonational variation in the British Isles. It contains audio recordings from 102 adolescent speakers from 9 cities in the British Isles: London, Dublin, Cambridge, Liverpool, Leeds, Bradford, Newcastle, Cardiff and Belfast. The IViE corpus does not contain any information 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 speakers reading the first two paragraphs of a longer retelling of the fairy tale Cinderella for analysis (about one minute per speaker). I validated all reference transcripts to ensure they accurately reflected the audio, including any repetitions or reading errors. This corpus was selected because of the range of regional varieties represented. One limitation of the corpus in this study is the absence of any Scottish speakers, as well as the lack of further demographic information, such as (caregiver) social class or ethnicity.

Speech Accent Archive

The Speech Accent Archive is a large 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[9] provided by first and second 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 languages as a first language (Arabic, French, German, Hindi, Italian, Mandarin, Portuguese, Spanish, Thai or Urdu) and vary in how, and for how long, they have been learning English. For each "native language subgroup", I selected up to 70 recordings. To ensure that any differences weren't due to systematic differences in recording quality, signal to noise ratio was measured using Praat (Boersma 2002), and recordings with a measure higher than 50dB were excluded from the subsequent statistical analysis. 445 recordings were retained. SAA was chosen as a corpus because of its unparalleled coverage of second language varieties of English. To compare the performance of British English ASR systems for L1 and L2 speakers, I selected all L1 speakers of British English. The Speech Accent Archive notes birthplace for each speaker. Since I am primarily interested in the differences between first and second language speakers, rather than variation among first language speakers here, I categorised the speakers according to the broad region of the British Isles their birthplace is located in, as shown in Table 3.1. This speaker group is very heterogeneous as shown in Table 3.2.

All audio recordings were converted to 16kHz FLAC files, uploaded to Google Cloud and

[9] Accessed here: `https://www.kaggle.com/rtatman/speech-accent-archive`

Table 3.1: Distribution of British English L1 speakers in the Speech Accent Archive.

Variety	Female	Male	Total
North of England	11	14	25
Northern Ireland	1	2	3
Scotland	3	5	8
South of England	7	13	20
Wales	1	2	3

Table 3.2: Distribution of L2 speakers in the Speech Accent Archive selected for the study.

Variety	Female	Male	Total
Arabic	15	34	49
French	31	27	58
German	20	11	31
Hindi	10	8	18
Italian	9	23	32
Mandarin	36	24	60
Portuguese	20	26	46
Spanish	23	40	63
Thai	6	7	13
Urdu	6	10	16

Amazon S3 Storage and processed using their Python APIs. Both corpora were processed with the default models for British English ('en-GB'). The generated transcripts were evaluated against the reference transcripts using `sclite` from the SCTK toolkit[10]. Further analysis of the evaluation outputs was conducted in R and Matlab to compare performance on speaker subgroups within each corpus.

3.3.2 Quantitative Results

I employ a mixed methods approach to analysing the experimental 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 standard 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. This qualitative analysis involves inspecting some of the most common errors to understand their potential origins.

[10] https://github.com/usnistgov/SCTK

Table 3.3: Speech Accent Archive: Distribution of mean word error rates for different groups and different systems.

L1	n	Amazon mean WER (%)	Google mean WER (%)
Arabic	49	21.79	32.26
English	59	14.08	23.13
French	58	19.10	26.84
German	31	16.44	25.43
Hindi	18	15.92	34.22
Italian	32	22.32	31.10
Mandarin	60	28.05	32.97
Portuguese	46	22.80	30.29
Spanish	63	26.79	32.86
Thai	13	32.53	38.35
Urdu	16	14.29	18.64

Table 3.4: Speech Accent Archive: Distribution of mean word error rates for different L1 groups and different systems.

L1	n	Amazon mean WER (%)	Google mean WER (%)
North of England	25	15.46	24.28
Northern Ireland	3	14.97	31.87
Scotland	8	15.39	20.27
South of England	20	11.13	22.02
Wales	3	17.87	19.80

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[11], 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. I provide mean word error rates of subgroups in Table 3.3 and Table 3.4.

For both systems, linear regression models show that word error rates are significantly higher for L2 speakers than L1 speakers.[12] Categorical predictors (variety: L1/L2 English, sex: male/female) are 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 (see Table 3.5a and Table 3.5b). Sex is not a significant factor for either model, and adding an interaction term of sex and variety does not improve model fit (according

[11] Recordings in the Speech Accent Archive are labelled by "sex" as either male or female. Acknowledging that sex and gender are separate if inter-related social constructs, I assume that "sex" in this context aligns with "gender" for most, if not all, speakers.

[12] Urdu L1 speakers form an interesting exception to this generalisation as seen in table 3.3 – while age of onset (as a proxy for English exposure and/or proficiency) was not investigated here, a qualitative analysis of the metadata and recordings suggests that the selected Urdu speakers have on average learned English at a younger age than most other L2 speaker groups.

to BIC/AIC). 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. This counter-intuitive result appears to be the result of exceptionally high WER for with very low speech rates (more than 1 standard deviation away from mean). Commercial ASR systems handle disfluent speech poorly. This effect is not observed for Google. Overall, Google produces higher error rates for both speaker groups.

Table 3.5: Speech Accent Archive data: Word Error Rate is **significantly (p<0.05)** higher for second language speakers of English than first language speakers of English for both ASR systems.

(a) Amazon: SAA – Reference L1 English

Variable	Estimate	Standard Error	t value
(Intercept)	19.58	0.84	23.37
English: L2	5.14	1.71	3.01
Gender: female	-1.53	1.12	-1.36
Speech rate	-4.40	0.60	-7.34
Age	-0.62	0.56	-1.08

(b) Google: SAA – Reference L1 English

Variable	Estimate	Standard Error	t value
(Intercept)	26.55	1.15	23.07
English: L2	7.91	2.34	3.37
Gender: female	-1.45	1.54	-0.94
Speech rate	0.09	0.82	0.11
Age	1.66	0.79	2.10

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 models. Amazon performs best on recordings from Cambridge, while Google performs best on those from London. Categorical predictors (gender: male/female) deviance coded and numeric predictors (speech rate in syllables per second) scaled and centered.[13] I provide mean error rates for different subgroups in Table 3.6.

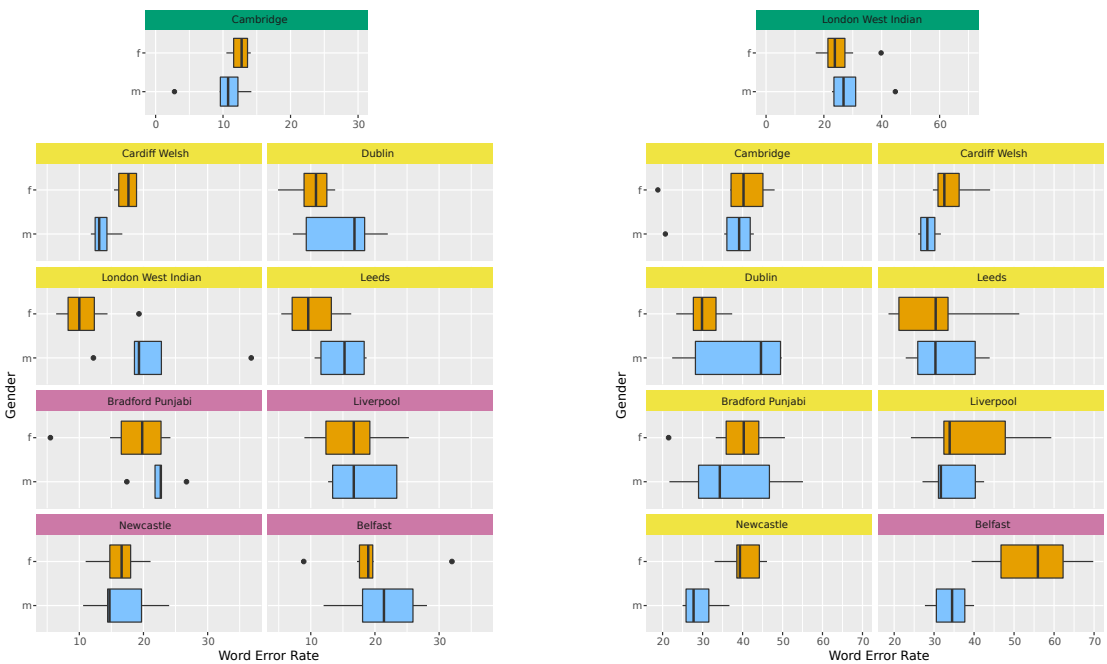
For speakers from Newcastle, Liverpool, Belfast, and Bradford, Amazon produces error rates which are significantly higher than those for speakers from Cambridge ($p<0.05$) (see Table 3.7a). There is also a significant main effect of gender, whereby recordings by female speakers show significantly lower 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 per-

[13] Recall that speakers were the same age and attended the same school. Other potentially relevant information is not recorded.

Table 3.6: IVIE: Distribution of mean word error rates for different groups and different systems.

Variety	n	Amazon mean WER (%)	Google mean WER (%)
Belfast	12	20.26	44.48
Bradford Punjabi	12	19.96	38.26
Cambridge	12	11.30	37.48
Cardiff Welsh	8	15.55	31.64
Dublin	11	12.31	34.23
Leeds	11	12.83	31.86
Liverpool	12	16.98	37.55
London West Indian	12	15.55	27.25
Newcastle	12	16.47	34.77

forms significantly worse for speakers from Belfast ($p<0.05$) (see Figure 3.1b and Table 3.7b). There is an interaction effect between variety and gender (which improves model fit): error rates are significantly higher for women from Belfast, Cardiff and Newcastle (see also Figure 3.1b).



(a) Amazon: IViE – Reference Cambridge

(b) Google: IViE – Reference London

Figure 3.1: 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) - as indicated by the green panel. Varieties with significantly higher ($p<0.05$) error rates than the reference level are indicated in pink. (Yellow panels indicate no significant difference).

Table 3.7: 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.

(a) Amazon: IViE – Reference Cambridge

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

(b) Google: IViE – Reference London

Variable	Estimate	Std Error	t value
(Intercept)	35.01	2.42	14.40
Cambridge	-0.07	3.38	-0.02
Cardiff Welsh	-0.12	3.45	-0.03
Bradford Punjabi	0.48	3.41	0.14
Leeds	-5.06	3.41	-1.49
Liverpool	2.82	3.22	0.88
Newcastle	-0.34	3.24	-0.11
Belfast	9.11	3.25	2.80
Dublin	-1.68	3.37	-0.50
Gender: female	-8.72	4.38	-1.99
Speech rate	6.41	0.96	6.67
Cambridge*female	8.31	6.08	1.37
Cardiff*female	15.55	6.83	2.28
Bradford*female	-2.21	6.24	-0.35
Leeds*female	9.68	6.31	1.53
Liverpool*female	10.48	6.12	1.71
Newcastle*female	12.73	6.09	2.09
Belfast*female	28.34	6.09	4.65
Dublin*female	1.35	6.28	0.22

3.3.3 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, 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. The ability to pinpoint which linguistic features "trigger" errors with the help of sociolinguistic expertise could be very useful in developing more robust technologies (Wassink et al. 2022).

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 (where the language model is "led astray" by an error early in the transcription and subsequently produces more errors) (Martin and Tang 2020, see also). Substitution errors can vary in impact: substitutions tend to be phonetically similar (but semantically unrelated) or morphologically related (but not necessarily phonetically similar).

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 /hɜːr/ is mis-transcribed as *her*. In RP, the sequence /hɜː(r)/ is indeed most likely *her*, while the actual target *hair* is produced as something like /hɛə/. Transcribing /hɜː(r)/ as *her* is therefore entirely expected if the system was trained on RP. However, Belfast English (like some other varieties of British English) collapses the distinction between the vowels in *hair* and *her* (Wells 1982). This is just one small example of a systematic difference in the phonology of different varieties, which can lead to predictive bias.

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).

3.3.4 Discussion

Predictive bias and 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 speech data tested: both corpora only contain read speech. This kind of careful speech is generally much easier to process for ASR systems than "real" conversational speech, as it is less affected by phonetic reduction and does not tend to contain hesitations or repetitions (Szymański et al. 2020). Because all speakers read the same passage, we can isolate differences in pronunciation, speech rate and prosody as the only sources of variation. The small but significant gender gap in Amazon's system with better performance for female speakers echoes findings by Koenecke et al. (2020) in the US. A quantitative analysis of the phonetic features (e.g. vowel quality) of the speakers is outwith the scope of this paper, but prior research in sociolinguistics has found time and again that women tend to avoid dialectal and stigmatised features more than men, speaking "closer to" the standard (Labov 1990). Google 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 WER of 9% for Amazon but 60% for Google with 80% deletion errors.

Overall, these findings add to the evidence that algorithmic bias not only extends to speech and language technologies but specifically reinforces and reproduces existing linguistic hierarchies and language ideologies. British English ASR systems appear to work best for prestigious varieties such as RP. Conversely, they work worst for second language speakers and speakers of (more or less) stigmatised regional varieties, groups who already experience (linguistic) discrimination (Sharma et al. 2022).

Origins of predictive bias

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, Koenecke et al. (2020) conclude that the higher error rates for African American English speakers are due to under-representation of AAE in the acoustic models. Martin and Tang (2020) further suggest that some AAE constructions are also under-represented in the language models used by commercial and open-source ASR. 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[14] 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 (Office for National Statistics 2020). The survey only considered variation in age and sex, with younger age groups being much more likely to use these technologies than older ones (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 (Pew Research Centre 2021). While 96% of households in the UK had broadband access in 2020, speeds vary by region (Baker 2021), 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 Lawrence 2021). Benjamin (2019b) 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 "desireable" 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" (Milroy 2001), the belief that there's an "ideal" or "standardised" variety of the language which is often the most

[14] https://www.amazon.co.uk/gp/help/customer, https://safety.google/assistant/

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 (Ardila et al. 2020) could in theory be more representative, in practice, they are also unbalanced (as discussed in detail in Section 5.4).

3.4 Discussion: Measuring predictive bias and classifying harms

3.4.1 Measuring performance and bias

Standardised evaluation metrics and benchmarks are central to all machine learning domains. Benchmarks are usually well-established and well-known "test" datasets which are used to compare the performance of systems and algorithms. Metrics are the "measurement tools" to quantify these performances. As Thomas and Uminsky (2022, p. 1) put it: "at their heart, what most current AI approaches do is optimize metrics." However, optimising metrics does not always relate to "better performance" if the metric does not accurately capture the task. As noted above, the standard procedure for evaluating novel ASR systems involves computing word error rates and/or character error rates over an established benchmark dataset. This process introduces two potential sources of "evaluation bias": the metric, if it fails to capture the error meaningfully, and the benchmark if it fails to represent the task well (here, speech recognition for a wide range of accents). Here, I will focus on the issue of meaningful metrics before returning to challenges related to benchmark datasets in Chapter 5.[15]

An obvious shortcoming of both word and character error rate as defined above is its complete disregard to the (linguistic and social) context of an error. Some errors obviously have a larger impact on the "usefulness" of the output. For example, the difference between *can* and *can't* obviously alters the meaning of the output completely. These kinds of errors may also be very difficult to disambiguate from context alone. This fundamental and obvious disadvantage of this edit-distance metric has long been noted in the speech recognition literature. One way to mitigate this problem is to assign different weights to different error types, with for example, a higher penalty for insertion errors than deletion errors. But this still sidesteps the issue that different errors of the same type can have different impacts. Morris et al. (2004) highlight two alternatives: match error rate (which represents the "probability of a given match being incorrect") and word information lost. While both of these metrics have clear advantages over WER and CER when it comes to evaluating ASR performance in application contexts like voice user interface, they have not been widely adopted. I have chosen to adopt WER despite its

[15] Here I distinguish between metrics (which we use to measure performance) and benchmark datasets which are designed as "test cases" for a particular system. See also Schlangen (2021) for a discussion.

shortcomings because it is so ubiquitous in the literature. As such it is easy to compare the results presented here, with results reported elsewhere. Unlike some of the alternatives, it is also relatively intuitive and simple to compute.

3.4.2 "Native speakers", "non-native speakers" and "accents"

There is an inherent tension in asserting on the one hand, that any group of language users is heterogeneous and that the boundaries between language varieties is fuzzy, and parcelling up speakers and varieties into neat bundles to evaluate a speech recognition system on the other.

As one reviewer of the study presented above generously pointed out, the unreflexive use of "L1" and "L2" as if they were homogenous groups runs counter to the nuanced sociolinguistic background I presented. The motivation for drawing on this distinction was to make a broad, simple point: people who acquired English as a first language, generally have an easier time interacting with ASR. However, in making that point, I obviously also glossed over a lot of variation within both groups. As highlighted in prior work and in my own work presented above, there are huge differences between varieties. Looking at the results from both studies we can clearly see that speakers of "standard", prestigious varieties like Standard Southern British English encounter better performance in commercial ASR tools, while those from less prestigious varieties might struggle. In this way, the existing linguistic hierarchies as documented though language attitude research and accent discrimination research, are replicated. Those who speak a "prestigious" variety are likely to be "heard more clearly", or at least more favourably, by both human and machine listeners.

3.4.3 Harms of predictive bias in automatic speech recognition

Following the taxonomy of harms laid out by Shelby et al. (2022), discussed in Section 2.4.3, we can begin to understand how both "biased" language technologies can create adverse outcomes for not just individual users, but larger social groups. In this section, I will discuss some of the harms which could and do result from the kind of predictive bias highlighted in the two studies presented here.

Alienation: othering and psychological harm

People who find that their interactions with technologies like smart speakers, voice assistants on a mobile phone or automatic captioning software are unsuccessful are likely to be dissatisfied with the technology. If they further notice that not everyone experiences these issues, but rather, that these technologies simply "work worse" for them, they might feel alienated by the technology. The psychological harms of predictive bias which affects already stigmatised varieties, go beyond simple inconvenience. Speakers of stigmatised varieties (L1 or L2), as well as other speakers whose voices are routinely subject to scrutiny or considered "deviant"

from the norm such as those with speech or hearing impairments, and trans speakers are likely already extremely aware of their own language production. As such they are more likely than more "normative" speakers (of high-status varieties, cis/gender-conforming, L1 speakers, no speech or hearing impairments) to be self-conscious about the way they speak.

Mengesha et al. (2021) and Wenzel et al. (2023) explore the psychological effects of inadequate conversational agent performance for speakers of African American Vernacular English (AAVE). As discussed above, predictive bias with respect to AAVE is will documented in US English ASR systems (Koenecke et al. 2020; Martin 2022). Mengesha et al. (2021) study participants report feeling "frustrated", "angry", "disappointed" and generally othered when American English conversational agents misrecognised their requests and they were unable to complete voice-based tasks. Many of the respondents expressed the belief that the technology was not designed with them or their language in mind, as they noticed, for example, errors in the recognition of "ethnic names" and "slang" (Mengesha et al. 2021, p. 5). While most attributed these errors to the system's limited abilities or flawed design, some did say that "the way [they] spoke or misspoke" caused them. Similarly, Wenzel et al. (2023) frame poor performance of ASR systems for speakers of African American English as "microaggressions". Using a Wizard-of-Oz experiment[16] simulating a voice assistant with controlled error rates, they show that higher error rates in voice assistants affect Black users' self-consciousness and self-esteem more than white users (Wenzel et al. 2023). The psychological harms of predictive bias are thus larger than the harms of low performance for speakers who are not usually marginalised or experience racial bias. As Wenzel et al. (2023) note, "linguistic and communicative misunderstandings are more systemic for Black individuals" and thus more harmful, while white users of speech technologies generally attribute errors to external sources, rather than their own linguistic, racial or ethnic identity.

Wu et al. (2020) show that both L1 speakers of English and L2 speakers of English (whose first language is Mandarin) carefully plan their speech and focus in particular on "clear" pronunciation, the way they reflect on speech recognition errors differs. The L2 speakers expressed frustration at having to restate whole queries if one word was misrecognised and the (perceived) inability of the conversational agent to provide some "guesses" based on the speech input (Wu et al. 2020). Importantly, again, some of them also seemed to rationalise these errors as the result of their own perceived linguistic "limitations" and place blame on themselves (Wu et al. 2020). An interesting insight from this kind of study focussed on an integrated voice user interface, rather than just an ASR system, is also that second language speakers would prefer longer pauses in between turns from the interface (Wu et al. 2020). Participants in both Wu et al. (2020) and Mengesha et al. (2021) also highlighted the need to be able to double check the output of the ASR system before a command is executed (e.g., an internet search, or sending a dictated text message).

Rincón et al. (2021) interview trans and non-binary users of voice user interfaces to explore

[16] A common experiment paradigm in human-computer interaction research where users interact with a device or user interface that is actually controlled by the researchers.

what their specific needs for these devices are and whether they are being met. While most of the participants discussed harms unrelated to predictive bias, as explored in more detail in Chapter 7, one person noted that the lack of sociolinguistic context meant that language related to gender and/or specifically to being trans might be misrecognised which could lead to frustration and alienation.

Increased Labour: correction and accommodation

Predictive bias can lead to increased labour for affected groups. For example, people who use automatic captioning of audio or video content may need to spend considerable amount of time correcting automatic transcripts. Considering that predictive bias disproportionately affects groups who might already be marginalised within the labour market or their workplaces, this is potentially compounding existing discrimination. Shelby et al. (2022) also mention linguistic accommodation as a type of "increased labor". This is an interesting framing, since research has shown that most speakers, regardless of their experience of predictive bias or their linguistic background, adjust their speech in some ways when interacting with an ASR system. For example, Branigan et al. (2010) have shown that people align their lexical and syntactic choices to computers. Similarly, both L1 and L2 speakers participating in Wu et al. (2020) discuss adapting their speech and planning their utterances carefully. However users who have experience of predictive bias, perhaps anticipate errors and adopt more mitigation strategies. These simplification strategies include changing pronunciation and speech rate and using "simpler" words and are also used by non-frequent users of conversational agents (Luger and Sellen 2016; Cowan et al. 2017; Wu et al. 2020). While linguistic accommodation also occurs (consciously and unconsciously) in interactions between people, any communicative challenges can be resolved through negotiation, for example by clarifying, or asking to repeat or rephrase. As Lawrence (2021) notes, unlike the accommodations made to be "understood" by an automated system, these interactions are not necessarily alienating because they "don't involve [...] taking on another's accent to be understood". As Lawrence (2021) puts it: "To put upon another an expectation of accent change is oppressive; to create conditions where accent choice is not negotiable by the speaker is hostile; to impose an accent upon another is violent."

Loss of benefit or service: accessibility

If the performance of a system is too poor, the benefit of that system is lost to the users. With ASR in particular, once a certain error rate is surpassed, repairing the transcript or interaction may involve more effort than using an alternative mode of transcription or interaction. The extent of the harm this causes depends on the specific use case. As ASR tools are increasingly used as accessibility tools (e.g., through automatic video captioning), this loss of benefits might not just affect the speaker but also the listener. For example, where automatic captions are

used to transcribe or live-caption business meetings or lectures, participants or audience members who require captions may be denied not just the benefit of the captions but, in essence, the benefit of the meeting or lecture. As Pradhan et al. (2018) highlight, voice user interfaces are used by people with hearing loss, speech impairments, visual impairments for a wide range of tasks (e.g., control smart home, playing music, looking up information, setting reminders) which would be much more difficult to achieve without voice recognition. Worse performance for these users who report increased independence and safety would therefore have a significant negative impact (Pradhan et al. 2018). As prior work shows, ASR systems do tend to perform worse for speakers with some speech and hearing impairments (Glasser 2019), as well as older speakers (Young and Mihailidis 2010) and children (Bhardwaj et al. 2022).

Opportunity loss and economic loss: high stakes contexts

One area where predictive bias in ASR could lead to opportunity (and economic) loss, is "algorithmic hiring". Automated systems promising to automatically pre-screen resumes and process video interviews to help employers select candidates to be interviewed by recruiters are increasingly popular (Raghavan et al. 2020). A central promise of many algorithmic hiring systems is to promote diversity and inclusion in the workplace by limiting the impact of unconscious bias of human recruiters (Garr and Jackson 2019; Sánchez-Monedero et al. 2020). Recent work has highlighted some risks of "outsourcing diversity work" in this way (Drage and Mackereth 2022), noted legal challenges in the UK (Sánchez-Monedero et al. 2020), discussed potential shortcomings of debiasing methods (Sánchez-Monedero et al. 2020; Raghavan et al. 2020) and pointed out specific instances of apparent bias in these systems (Rhea et al. 2022). Predictive bias in ASR could affect systems which explicitly analyse speech data. Higher error rates in the automatic decoding and analysing of responses to interview questions could disadvantage applicants, and result in them being ranked "lower" than other applicants. While many vendors, including Hirevue, complete internal and external bias auditing to limit the effect of bias in their rankings, it is unlikely that ASR systems perform equally well for everyone and quite likely that there are some systematic differences between speaker groups (see also Martin 2022, p. 56). Sánchez-Monedero et al. (2020) examine the way three large algorithmic hiring system vendors describe their "debiasing" processes, and point out that these rely on clear definitions of protected characteristics (a potential flaw of a lot of "fairness in AI" approaches, as discussed by Hoffmann (2019)) and that these characteristics do not include social class or disability. As these systems (many of which originate in the US), are adopted by employers hiring globally, auditing for and mitigating predictive bias related to language variation becomes a complicated challenge.

Cultural harms: language ideologies

Taking seriously the idea that these harms listed above matter, also means that there are wider consequences to predictive bias beyond the individual. In performing much better for some language varieties and some people, ASR systems (perhaps completely inadvertently) reproduce existing linguistic hierarchies. As discussed above and in Section 2.3.2, these hierarchies are expressed in attitudes about what is considered "proper English" or "correct pronunciation" or "clear speech", and which varieties are considered "prestigious" and "pleasant" and, consequently, who is considered "educated", "friendly", "trustworthy", "professional". These value judgments are not just rooted wider oppressive social structures such as racism, classism, sexism and ableism but also play a significant role in reinforcing them as they influence who is hired for well-paid jobs, who is taken seriously when speaking, and who is not. Disparities in ASR performance plays into this by further normalising and entrenching these hierarchies and value judgements. The "individual" harm of alienation discussed above manifests not only in frustration by individual users, but can also reinforce internalised insecurities about their speech which in turn also affects how they speak. Marginalised varieties like AAVE in the US have long been framed and heard as "deficient" (Rosa and Flores 2017; Rickford and King 2016; Lippi-Green 2012), and as highlighted above, negative attitudes towards varieties spoken by working class speakers and ethnic minorities in the UK also have a long history (Sharma et al. 2022). Predictive bias shows how "a machine" can also "hear" users as "deficient" in this way. It serves as a subtle (or not so subtle) reminder that their way of speaking is at best, not "appropriate" for some domains (like interaction with technology) or "difficult" for an ASR system, and at worst "incorrect". As Lawrence (2021) argues (in an essay titled "Siri Disciplines") the accommodation required to mitigate predictive bias in ASR, is not just alienating but is also a form of forced assimilation. As I discuss in Chapter 5 and Chapter 6, I believe that language technology developers have an important (if subtle) role in enforcing language policy and ideologies by shaping which varieties are "legible" to the "machine listener".

3.5 Conclusion

In this chapter, I explored predictive bias in two commercial British English ASR systems. To contextualise this work, I outlined prior work on predictive bias in ASR, which has generally found that ASR systems perform worse for already marginalised speakers. Similarly, my findings show significantly worse performance for second language speakers of English, and speakers of stigmatised varieties in the UK. To understand why this matters, I apply a framework of "harms" and consider how speakers are impacted by these different error rates depending on how and where they interact with ASR systems. Crucially, even speakers within the same "group" are affected differently depending on how "important" these systems are for them. As explored in the next chapter, one way to account for socially meaningful variation in language (which furthermore occurs at the group-level) is to develop systems specifically with particular groups

and use cases in mind, or enable easy fine-tuning with small datasets. In the following chapter, I consider how the specific application context of an ASR system in South Africa affects the kind of language datasets we should use in ASR development.

Chapter 4

Automatic Transcription and Standardisation

4.1 Introduction

To fully understand potential benefits and harms of automatic speech recognition it is worth reflecting more deeply on a specific task. In this chapter, I explore the way ASR is used to transcribe speech.

I first discuss some background on both manual and automatic transcription as a task. The way orthographic transcription and orthographic representations have been conceptualised in sociolinguistics, theoretical linguistics, language documentation on the one hand, and in ASR research on the other, is particularly instructive. The gap between these different perspectives may account for some of the limitations of current ASR development approaches. I present two studies in this chapter. The first is a case study of how ASR systems can be embedded in sociolinguistic research and analysis. A version of Section 4.3, including some additional background on algorithmic bias in ASR, and the discussion in Section 4.5.1 were published in Markl (2022a). It is an extended write-up of a talk presented at a sociolinguistics conference (NWAV 49) in 2021.

The second study presented in Section 4.4 focusses on designing and evaluating a transcription tool for and with the isiXhosa-speaking community in Langa, South Africa. This section is taken from an unpublished manuscript on which I am the lead author alongside an interdisciplinary group of collaborators. Some of our joint work discussed in this section builds on Reitmaier et al. (2023) (where I am a co-author).

Reflecting on these two use cases of automatic transcription, I discuss the way ASR can be used as a tool for both specialist and non-specialist purposes, and how it interacts with the wider sociolinguistic context once in popular use in Section 4.5.

4.2 Background: Transcription and Automatic Transcription as a Task

Here I briefly discuss scholarship around "orthography", as a set of conventionalised ways of representing lexical items, and "transcription" as a way of mapping speech sounds to graphic representations. Both terms denote complicated processes, yet both are somewhat under-theorised, in particular in relation to ASR where the process of mapping sounds to graphic representations is generally understood as conceptually (if not technically) straightforward.

4.2.1 Orthography as a "social practice"

Most literate speakers of standardised languages probably rarely reflect on the theories and ideologies underpinning "writing". The strict codification of standard varieties in prescriptive dictionaries and grammars means that for them, there is very little room for spelling variation in formal writing. However, as Sebba (2007) argues, orthography is a "social practice". Different "spaces" of writing are more or less regulated according to standard spelling norms (Sebba 2007, p. 43). Domains which Sebba (2007, p. 41) terms "marginal" literacies, like those facilitated and shaped by digital technologies, generally allow for more variation[1] (see also Androutsopoulos 2016). Variation in spelling, just like variation in speech, can carry social meaning. In addition to highly specific, recognisable social meanings (e.g., where the spelling reflects regional variation in spoken language), "misspellings" also carry an "oppositional" meaning as they defy the spelling norms (Sebba 2007).

Orthographies have an extremely important role in establishing and maintaining standard varieties. Developing, codifying and transmitting orthographic conventions helps establish the "standard" as a seemingly "invariant" variety. Discussions about orthography, just like discussions about other aspects of language use (Irvine and Gal 2000), are not about language per se but about larger issues of identity, culture and power. Decisions about what should be considered the "correct" way to spell are not made in a vacuum, but affected by existing power structures. This is particularly obvious in postcolonial contexts where tensions between orthographies that reuse spelling conventions used in colonisers' languages and those which fully reject colonial legacies are especially acute (e.g., Indonesian discussed by Sebba 2007).

4.2.2 Transcription as "theory"

As has been discussed in a number of linguistic subfields, transcribers have to *interpret*, often "underdetermined" (Himmelmann 2018, p. 35), speech signals based on their own linguistic and cultural knowledge. Ochs (1979) discusses "transcription as theory" in the context of linguistics (specifically, language acquisition research). In part drawing on Ochs, Jaffe (2000,

[1] Fifteen years later it appears debatable whether we should consider any language use facilitated by digital technologies as "marginal", but the general observation that they provide more space(s) for variation appears to hold.

p. 501) describes this as an issue of "selectivity": as she puts it "the *how* of *how people speak*" is so complex that representing it graphically always involves a trade-off between being readable to different kinds of audiences (e.g., a specialist or non-specialist one) and conveying the relevant semantic and social meaning. In this way, even the "simplest" orthographic transcription involves theoretically informed decisions. For example, transcribers may establish or choose between possible conventions to represent speech and non-speech sounds (e.g., laughter, background noise), as well as common features of spontaneous speech but not (formal) writing such as false starts and filled pauses. As Bucholtz (2000) notes, there are thus two levels to the interpretive work of transcriptions relating to both content and the form of the transcription, and, depending on what the purpose of the transcript is, these choices have far-reaching consequences.

As Himmelmann notes (in the context of language documentation), "consider[ing] transcription exclusively, or even primarily, a process of mechanically converting a dynamic acoustic signal into a static graphic/visual one" would therefore be "rather naïve" (2018, p. 35). Despite impressive advances in recent years, ASR transcriptions are not perfect, and, for most conversational speech, not on par with human transcription. This is, in part, because ASR tools, like other language technologies have only limited access to the linguistic, and perhaps more importantly, social and cultural context(s) humans can make use of when transcribing potentially ambiguous speech. They also assume that transcription *is* a straightforward mapping between a dynamic acoustic signal and a static graphic signal. And with the increased use of ASR, we perhaps risk further naturalising this idea of transcription as a "mechanic process", rather than a complex social process in which a particular kind of text is created by interpreting a different kind of text.

The relative inflexibility of ASR trained on standard orthography in the face of non-standard lexical items and pronunciations, is a particular challenge when we try to use it to create "full", verbatim transcripts. In Section 4.3, I discuss this in relation to sociophonetic corpora.

4.2.3 Orthography and transcription in ASR development

Established ASR development pipelines fundamentally depend on transcription for training and/or testing. Traditionally, supervised ASR systems rely on speech datasets paired with transcripts. These "gold transcripts", presumed to be "accurate", are used to train an ASR system, mapping speech sequences to character sequences. More recently unsupervised methods have alleviated the need for transcribed data in the training stage somewhat. For example, Baevski et al. (2020) leverage the ability of particular model architectures to learn representations from unlabelled data. However, in addition to requiring more audio data, these systems still require some (unpaired) text data to train a language model. While transcriptions may become less important in training ASR systems, they are still fundamental in evaluation. In order to assess the quality of the automatic transcription, we need a comparable "gold standard". This is particularly acute in cases where the task of the ASR system really is producing a

transcript, rather than, say, recognising a single voice command provided to a smart speaker.

Bird (2020) critiques this focus on "gold data", which all too often goes hand in hand with an understanding of transcription as "objective", "neutral" or "accurate" where any variation is due to "error" rather than a (potentially interesting) product of a complex social process. Traditionally, the work of creating "faithful" and accurate transcripts is assigned to linguists who are thereby positioned as "experts" often over and above members of the language community. In this way "data collection", whether for the purpose of language documentation, language technology development, or both, risks reproducing an extractive and colonial mode of research and development where the needs and desires of language communities are sidestepped to the benefit of the needs of outsiders (Deumert and Storch 2018; Bird 2020; Bird 2021). In addition to the harms of engaging with communities and their languages in this way (explored in more detail in Chapter 5), these approaches might also be ineffective in that the ASR tool developed at the end of the process is not of much use or appeal to the community.

If we take seriously the notion that both orthography and transcription are theory-laden, social processes situated in a wider sociolinguistic context, we need to think carefully about the kinds of orthographies we encode in language technologies. By "training" a machine learning model to render language according to a particular orthography, developers are interacting with existing linguistic hierarchies and potentially reinforcing the status of a widely accepted standard. In a language like English, where "the lid [on orthographic variation] is most tightly screwed on" (Sebba 2007, p. 32), most people using an ASR system likely expect output to follow the written standard very closely, though as discussed in the previous chapter, the inability of systems to transcribe "slang" can also cause frustration and alienation (Mengesha et al. 2021). Decisions around which conventions to encode in language technologies become much more obviously fraught when developing for and with minoritised varieties and in contexts where orthographies are not just contested but alienating for many speakers. In Section 4.4, I discuss the case of developing ASR for and with the isiXhosa speaking community in Langa, a township of Cape Town. The written isiXhosa norm is perceived as both a poor reflection of the variation in spoken and written language use in Langa, and furthermore strongly associated with oppressive colonial institutions which designed and implemented it. Here, adopting this written standard would likely result in poor performance of the system and either cause speakers to reject the system, or accommodate or adapt to it.

Before turning to the context of isiXhosa, I discuss the benefits (and pitfalls) of ASR in preparing sociolinguistic data for analysis and archiving. Here too, the focus on conversational, spontaneous speech and non-standard varieties exposes opportunities for improvement in ASR tools.

4.3 ASR in sociolinguistic research

The content of Section 4.3 is published in Nina Markl (2022a). "(Commercial) Automatic Speech Recognition as a Tool in Sociolinguistic Research". In: University of Pennsylvania Working Papers in Linguistics *28.2. URL: https://repository.upenn.edu/pwpl/vol28/iss2/11*

In this section, I discuss the potential advantages and problems of using ASR to facilitate transcription of sociolinguistic data, with a particular focus on commercial ASR engines provided by Amazon and Google. The specific application context is the Lothian Diary Project, an interdisciplinary research project at the University of Edinburgh, collecting audio and video diaries recorded by residents in the Scottish Lothians region documenting their experiences of the COVID-19 pandemic between May 2020 and July 2021 (Hall-Lew et al. 2022). As a member of the Lothian Diaries research team, I facilitated the transcription of the almost 200 English language recordings using ASR and manual correction.

4.3.1 Orthographic transcription in the age of "big sociophonetics"

Collection, compilation and storage of large speech datasets has, in general, become much easier in recent decades. Computational methods aiding acoustic analysis of larger datasets such as forced alignment, have also become more accessible and reliable (e.g., Mackenzie and Turton 2020; Reddy and Stanford 2015), and recent developments in the application of machine learning to phonetic analysis are promising (e.g., Villarreal et al. 2020). However, most sociolinguistic analyses, whether they involve acoustic analysis or not, also require orthographic transcripts. Preparing these orthographic transcripts can pose a bottleneck: manual transcription is a very laborious process (Bird 2021). Automatic speech recognition (ASR) coupled with manual correction could potentially alleviate this work load. While ASR has recently found wide use in applications ranging from voice user interfaces in mobile devices to automatic captioning in virtual classrooms and on social media, it is not yet a widely used tool in sociolinguistic research.

Orthographic transcriptions are important for a wide range of sociolinguistic research methods. In (socio)phonetic research, forced alignment of (partial or complete) orthographic transcripts is now a standard method to facilitate semi-automatic segmentation and acoustic analysis (Mackenzie and Turton 2020). Even simple orthographic transcripts allow efficient search and topic and corpus analysis, while more complex transcriptions can facilitate analysis of interactions. Many sociolinguistic datasets are also of potential interest to the general public (e.g., as oral history archives) and researchers in other fields (e.g., because of the topics discussed by participants). Transcriptions make these datasets more accessible, portable and, depending on the type of recording, durable. A transcript is also often easier to reproduce in a research publication or presentation, both for practical and ethical reasons, as the voice

recording, but not the transcript, may be considered "biometric data" which can be used to identify individuals (Information Commissioner's Office 2022).

With the development of new algorithms and architectures as well as advances in computing, performance of ASR systems, especially in (American) English, particular domains and speech styles, has steadily improved. Within sociolinguistic research, ASR has been incorporated in DARLA (Reddy and Stanford 2015; Coto-Solano et al. 2021), an impressive and popular tool facilitating fully automatic extraction of acoustic measurements from American English speech (using partial transcription). However, automatic accurate and full transcription of spontaneous speech is still very difficult, and not yet particularly widespread as a tool among sociolinguists (though some labs and research groups are taking it up: e.g., Wassink et al. (2022)).

4.3.2 A case study: the Lothian Diary Project

Between May 2020 and July 2021, the Lothian Diary Project collected 195 audio and video diaries (Hall-Lew et al. 2022). We invited residents of Edinburgh and the Lothians region in Scotland of all backgrounds to reflect on their experiences of the COVID-19 lockdowns. In addition to linguistic variation, we were also particularly interested in understanding and reporting on the ways people were impacted by the pandemic and the changes to day-to-day lives it brought, as well as attitudes towards government policy (which fed into a report for the Scottish Government[2]). Due to the scale and time-sensitive nature of the dataset and this interest in the attitudes and experiences expressed in the recordings, a relatively quick way of preparing full and accurate transcripts was crucial. We opted for a pipeline involving a custom, locally-run ASR system and manual correction.

Local setup

To facilitate faster transcription of the diaries, we developed a pipeline involving ASR and manual correction. We opted for the widely used open-source ASR toolkit Kaldi (Povey et al. 2011). The acoustic model and baseline language model were provided by colleagues at the Centre for Speech Technology Research at the University of Edinburgh. As most of our recordings are in a British variety of English, we used an acoustic model trained on recordings drawn from BBC broadcasts[3]. The language model was adapted and enhanced with (manually produced) transcripts of 37 Lothian diaries, as well as text scraped from the RSS feed of the BBC news service (published between March and July 2020) and posts relating to COVID-19 on the Edinburgh subreddit[4]. These additions allowed for better inclusion of terms related to the pandemic, local

[2] `https://lothianlockdown.org/parliamentreport/`

[3] Developed by the Centre for Speech Technology Research as part of the Multi-Genre Broadcast challenge in 2015: `http://www.mgb-challenge.org/`

[4] `https://www.reddit.com/r/Edinburgh/`: Posts containing the terms "Covid", "Lockdown", "Coronavirus" or "Quarantine" and all their comments: After removing duplicate posts, the corpus consists of 288 individual posts and 4833 comments. Only 2 were created before January 2020.

issues and place names and greatly improved performance of the ASR system.

Kaldi requires recordings and auxiliary files according to specific requirements (see Chodroff 2018, for an excellent tutorial). This involves fairly complex (but automatable) creation of utterance chunks and several text files linking each utterance to a time-stamp and a speaker. Once these files are prepared, Kaldi is run locally from the command line. To facilitate manual correction, the output of the ASR system can be saved as a tab-delimited file with time-stamps and imported to ELAN for further processing. Each utterance chunk can then be reviewed and corrected if necessary. Overall, this pipeline is more efficient than manual chunking and transcribing, though the quality of the automatic transcript differed greatly depending on accent, recording quality and background noise.

There are clear advantages to using a custom-built system, such as control over training data and a variety of system settings. Local processing also side-steps potential ethical or data protection issues associated with transferring recordings to a third party. However, using (and in particular, training) Kaldi is a non-trivial task requiring significant coding expertise, some computing resources on a unix system (Kaldi is not supported for Windows) and, of course, data to train the acoustic model and language model. There are also more user-friendly tools building on Kaldi, such as ELPIS (Foley et al. 2018)[5] which allow relatively simple training and use of custom ASR tools, as well as alternative open-source toolkits such as DeepSpeech[6].

Commercial setup

I considered the off-the-shelf British English ASR systems from Google (Google Cloud Speech-to-Text) and Amazon (Amazon Transcribe). These systems can be accessed through APIs in various programming languages or user-friendly interfaces. Using the ASR system does not require pre-processing of the data or programming skills. Like most commercially available ASR engines, they rely on cloud-computing. This can be advantageous as they do not require much local computing resources. As a result audio files need to be uploaded to the cloud storage of the respective provider (usually only for the duration of the processing).

Errors As noted above, ASR systems are usually evaluated using word error rate. This standard metric fails to account for the context in which errors occur: some errors distort the intended message more than others (e.g., substituting *can't* for *can*). In qualitative, "context-sensitive" exploratory error analysis of the commercial ASR systems, we identified several error types.

Some errors appear to be caused by phonetic differences between the training and test data. Figure 4.1 shows a comparison between three ASR models applied to recordings by two Scottish English speakers (a man from Edinburgh and a woman from Glasgow). Google only offers one model for "British English", Amazon offers "British English" and "Scottish English".

[5] Developed for language documentation: `https://elpis.readthedocs.io/en/latest/`

[6] `https://deepspeech.readthedocs.io/en/r0.9/`

While the models produce different errors, reduced phonetic forms appear particularly challenging for them. As can be seen in Figure 4.1a, all three models tend to delete tokens they identify as filled pauses. This affects actual filled pauses (here transcribed as *er* and *erm*) and reduced forms of word (e.g., *I* in *but I live alone*). In Figure 4.1a, we can also see how this interacts with phonetic variation: one of the filled pauses is identified as such and deleted by Google's model and Amazon's Scottish English model. Amazon's British English model, however, transcribes this this token, which is produced as /e:/ as *a*, presumably because this realisation is much more similar to, for example, Southern British English pronunciations of *a* than it is to those of *uh*. The reduced centralised vowel in *I* in *I live with a cat but I live alone* appears to be misclassified as a filled pause by all three models (and subsequently deleted). *Cat*, which is produced with final glottal stop, is also mistranscribed by all three models as *car* or *cap*. Word-medial glottal replacement in *isolating* appears less challenging for two of the systems, perhaps because there are fewer phonetically similar alternatives. Though notably, Amazon's British English system also fails to transcribe this correctly. This seems particularly surprising as glottal replacement is prevalent in most British varieties, including those in South of England which this system appears to be mostly trained on.

Substitutions are often morphologically related forms. Sometimes these are also phonetically similar and could be the result of reduction (e.g., *hardest* > *hard* in Figure 4.1b) and other times they are less similar (e.g., *found* > *find* in Figure 4.2b). These errors are particularly disruptive where they significantly change the meaning. For example, *We couldn't even go out* was transcribed as *We could even go out* by Google's ASR tool (see Figure 4.2b). These errors may be in part driven by the language model, which is used to decode the recognised sound sequences into utterances. When decoding a sound sequence using a language model, all words (types) present in the language model get assigned a probability. This probability is conditional on the sound sequence, (usually) some linguistic context (e.g., the words preceding and following) and the baseline probability of the word in the corpus. This means that, depending on the decoding algorithm, words which are very infrequent in the corpus used to train the language model[7] may be less likely to be accurately decoded than high-frequency words (which may, conversely, be assigned too high a probability). In the context of the Lothian Diaries, we can see this in frequent errors around terms like "lockdown", "Covid", "social distancing" and other words and names related to current affairs. We also see the influence of the language model in utterances with repetitions. For example in Figure 4.2c, the sentence *I miss my family, I miss my friends* is simplified to *I miss my family and my friends* by the Google ASR model. Both models delete a false start in that same utterance (*it's – I shouldn't be here* > *I shouldn't be here*). Another potential influence of the language model is the decoding of contractions as full forms (e.g. *I'm* > *I am* in Figure 4.2b). The deletion of filled pauses discussed above may also be exacerbated by language models trained on text which do not contain them.

[7] In the case of large language models like GPT-3, these corpora include petabytes worth of text crawled from the internet, e.g. CommonCrawl, a corpus which also contains a "significant amount of undesirable content, including hate speech and sexually explicit content" (Luccioni and Viviano 2021).

I've	missed it	it's	been	quite	a while	but	erm	myself I found	er	isolating a bit difficult
of		Mr has	been	quite	a while	but	***	myself I found	**	isolating a bit difficult
I've		Mr has	been	quite	a while	but	a	myself I found	a	a silly and a bit difficult
I've	missed it	has	been	quite	a while	but	a	myself I found	**	isolating a bit difficult

Er	I live alone	I live	with a cat	but	I live alone	no other adults	living in the house		
**	I live alone	* love	with a car	but	* live alone	no other adults	****** ** going out		
**	I live alone	* ****	with a cap	but	* let alone	lure that I don't	see you	in the	nose
**	I live alone	* love	with a cap	but	* live alone	no other adults	coming	in the house	

(a) Utterance produced by a male speaker from Edinburgh (born 1975).

The hardest	part of lockdown	for me was definitely missing out	on seeing loved ones				
The highest	part of what time	for me was definitely missing out	on seeing loved ones				
The hard	part of what done	for me was definitely messing it	on seen	loved ones			
The hard	part of lock down	for me was definitely missing out	on seen	loved ones			

(b) Utterance produced by a female speaker from Glasgow (born 1999).

Figure 4.1: Comparison of reference transcript (top) with automatic transcripts. For Scottish speakers, there are three models to compare: Google (British English) with errors highlighted in blue, Amazon (British English) with errors highlighted in orange and Amazon (Scottish English) with errors highlighted in yellow. Deletions are marked with asterisk (*).

Preparing transcripts with ASR assistance

Accurately transcribing spontaneous speech with filled pauses, overlaps, hesitations and false starts is a very challenging task even for trained human transcribers. Correcting incomplete or erroneous ASR transcripts of spontaneous speech can be similarly challenging. We hand-corrected transcripts by checking each utterance in ELAN, listening to each audio chunk, reading the corresponding transcript and making changes as necessary. While many errors are easy to spot and correct, others such as missing filled pauses and reduced forms can be trickier. These errors are also unlikely to be detected when reading through the transcript without the audio, since the resulting sentences may be perfectly grammatical and meaningful. In addition to errors humans might make too, there also errors no human transcriber would make (e.g., Figure 4.2a). ASR systems appear to be sensitive to issues in recording quality, noise and, as discussed above, language model biases.

There are trade-offs involved in whether and how to integrate ASR tools in transcription workflows. Especially with very diverse speech datasets like the Lothian Diaries (different accents, topics, recording conditions etc.) the same ASR tool may also perform very differently for particular speakers and recordings. Fully manual transcripts may contain fewer transcription errors (of some types: e.g., substitutions of nouns or verbs) than hand-corrected automatic transcripts. However, they also take an order of magnitude longer to prepare. At a bare min-

Hi my name is Rosa and today we'll be seeing how my lockdown
** ** how many murderers and today will be seeing how my lockdown
** ** how many mysteries on today we will be seeing how my lock down

(a) Utterance produced by female speaker born in 2010 in Italy who is growing up in Edinburgh.

we couldn't even go out more than once a day I mean I I'm very active I cycle and run everywhere
we could even go out more than once a day I mean I am very active I cycle *** run everywhere
we couldn't even go out more than once a day I mean I I'm very active I cycle *** run everywhere

and I found it really hard to only be able to go out once a day
and I find it really hard to only be able to go out **** * today
andi I found it really hard to only be able to go out once a day

(b) Utterance produced by a female speaker from London (born 1974).

I miss my family I miss my friends erm and i kind of feel like it's I shouldn't be here
I miss my family * and my friends *** and I kind of feel like **** I shouldn't be here
I missed my family and missed my friends *** and I kind of do you like **** I shouldn't be here

(c) Utterance produced by a female speaker from Lithuania (born 1994).

Figure 4.2: Comparison of reference transcript (black) with automatic transcript. For non-Scottish speakers there are two models to compare: Google (British English) with errors highlighted in blue and Amazon (British English) with errors highlighted in orange. Deletions are marked with asterisk (*).

imum we have found automatic segmentation of recordings into chunks (using voice activity detection) very useful in both manual and automatic transcription.

4.4 Automatic Speech Recognition in Langa, South Africa

Since January 2021, I have been involved in the UnMute Project, an interdisciplinary research project consisting of a team of speech technologists and human-computer interaction researchers and contributed to two collaborative research papers: Reitmaier et al. (2023) and a manuscript presented below of which I am the lead author.[8] In this section, I present this joint work on "Automatic Transcription and (De)Standardisation".

I am the lead author of the manuscript presented in Section 4.4. I designed, conducted and analysed the survey evaluation study with my co-author Electra Wallington, provided suggestions for the design of the transcription workshop and analysed recordings of that workshop conducted by my co-authors Thomas Reitmaier, Simon Robinson, Gavin Bailey, Jennifer Pearson and Matt Jones. Electra Wallington, Ondrej Klejch and Peter Bell designed the ASR systems discussed. All authors contributed to the write-up.

4.4.1 Introduction

Code-switching and variation are common features of spoken and written language in a wide range of contexts (Sebba 2007; Heller 1988). Most conversational user interfaces, however, are predicated on monolingual speakers of standard varieties (Bird 2022). Developers of Automatic Speech Recognition (ASR) systems, for instance, assume one "correct" "gold standard" of how particular words are spoken and transcribed to both train and test their systems (Bird 2020). Similarly, it is generally assumed that speakers speak one particular (named) language, rather than drawing on several languages in a single interaction or sentence.

In other words, there is a gap between real spoken language usage and the language usage assumed by most CUIs. This paper's contributions are thus as follows:

1. We illustrate this gap, and highlight why, in their current form, ASR development pipelines may not be compatible with such "real language usage"

2. We examine the dangers of this gap – in view of technical and practical considerations, and also from an ethical perspective – and propose some ways to address it through involving users in different stages of the technology development

We do this through the lens of one particular case study: developing an ASR system for the isiXhosa speaking community in Langa, South Africa. We focus on users' perspectives on how best to represent their language use in an ASR tool as explored at an in-person workshop and a remote survey-based evaluation of the tool.

[8] I was not involved in the development of Reitmaier et al. (2022) or any of the work that it was based on.

4.4.2 Context

Langa and isiXhosa: Multilingualism

The site of this study is Langa, a township in Cape Town, South Africa. In 2011, about a third of Cape Town residents spoke Afrikaans as a first or main language, while about 30% spoke isiXhosa, and 30% spoke English[9]. While other languages are also spoken in the city, these three languages are both official and dominant in the region, with the colonial languages of Afrikaans and English particularly common in official settings, on public signage and in writing (Deumert et al. 2021; Dantile 2015). In Langa, more than 90% of the about 50,000 residents reported isiXhosa as a first or main language in 2011[10]. During apartheid rule, the non-white isiXhosa speakers were prohibited from settling in Cape Town proper and were instead restricted to designated areas like Langa. IsiXhosa (also known as Xhosa in English) is closely related to other languages spoken in the region (e.g., isiZulu and isiNdebele) (Deumert and Mabandla 2017).

The variety[11] of isiXhosa spoken in and around Cape Town is characterised by code-mixing with other languages such as English (Deumert 2010). This "mixing" of seemingly distinct languages, is referred to as "code-switching" by linguists (Heller 1988), or described by the distinct but related concept of "translanguaging" (Otheguy et al. 2015). Many multilingual speakers, especially those living in multilingual communities such as Langa, draw upon all the languages or linguistic resources in their repertoire, often "switching" within the same conversation or even the same sentence.

Langa and isiXhosa: Orthography and "the standard"

Despite most speakers in Langa being multilingual, the established, formal written standard does not easily account for this variation, and more informal registers. Perhaps more critically, the written norm is also strongly associated with its colonial origins, and as such might alienate speakers.

The history of the standardisation of isiXhosa by Deumert and Mabandla (2017) is instructive here. Early dictionaries and grammars of the type recognisable (and necessary) to European linguists and missionaries, were produced by missionaries. By the mid 1800s, these early descriptions of isiXhosa and bible translations based on them, were subject to criticism by the "growing African intelligentsia", many of whom were (at least in part) educated in mission schools and in places like the United Kingdom (Deumert and Mabandla 2017, p. 206). The 19th century also saw the proliferation of isiXhosa newspapers, with the first outwith missionary editorial control appearing in the 1880s (Deumert and Mabandla 2017). As Deumert and Mabandla (2017) highlight, this isiXhosa publishing industry became an important locus for resistance to the colonial regime, literary writing and metalinguistic discourse, especially as

[9] https://www.statssa.gov.za/?page_id=1021&id=city-of-cape-town-municipality
[10] https://www.statssa.gov.za/?page_id=4286&id=318
[11] We use the term "language variety" as a more neutral descriptor of a "dialect".

literacy rates rose in the early 20th century. At this time, linguists and missionaries in Europe developed an orthographic system to be implemented across all African languages. This reform was met with vocal resistance by African writers and readers, but embraced and enforced by powerful publishers (who already ensured that nothing too critical of the colonial government or Christianity was published) (Deumert and Mabandla 2017). After further spelling reforms in the 1950s and 1970s, today's standard orthography remains alien(ating) for many speakers. As Deumert puts it: "[The written, school-taught isiXhosa standard norm] is perceived as an unchanging artefact which stands in strong opposition to the vibrancy an innovation of the spoken language" (Deumert and Mabandla 2017, p. 250). As a result local writing practices diverge from the formal standard, not just through the inclusion of words and phrases from other languages but also through extensive variation in spelling between different speakers (Deumert and Mabandla 2017).

An ASR System for Langa

The objective of ASR is to map a sequences of speech sounds to a graphic representation. Previous research involving people from Langa uncovered a key application context of an ASR tool, namely to enable the transcription of voice messages – an extremely popular medium of communication Langa (Reitmaier et al. 2022).

Traditionally, to develop such an ASR system for a new variety, we require "supervised'' speech data – that is, speech audio paired with "gold-standard'' transcription – along with sufficient amounts of text data in the target language (not necessarily needing to be tied to any speech). The choice of speech data in terms of domain and language variety dictates what kind of speech *input* our system will be able to process when deployed; the choice of text data (genre, topic, language variety) dictates what kind of textual *output* our system will produce. After training on this data, one would typically evaluate the system's performance on a held-out data-set also consisting of speech audio paired with "gold-standard" transcripts.

However, baked into this pipeline is an underlying assumption of monolingual speakers of standardised languages. "Standardised languages" are those varieties which are codified in dictionaries, transmitted through education systems, and often recognised as "prestigious" (Milroy 2001). By training a system using text (and often speech) data in the standard variety and/or formal domains, and subsequently evaluating system performance through comparison with transcripts following conventions of the standard orthography, the status of the standard variety and orthography is reinforced.

4.4.3 Part 1: The gap between "real language use" and existing language resources

Language resources

Available language resources for ASR development As introduced above, building a (conventional) ASR system requires labelled (i.e., transcribed) speech data-sets both for training and testing. While there are some data sets which fit this requirement for isiXhosa, it is apparent that this these resources are not representative of the way most isiXhosa speakers in Langa use language. The NCHLT isiXhosa speech corpus for instance (Barnard et al. 2014) consists of 56 hours of read speech (participants were asked to read out short phrasal prompts such as 'omnye ngaphandle kwesizathu'[12]); hence, this corpus contains none of the multilingualism or code-switching we may expect from spontaneous conversational isiXhosa.

Much of the available isiXhosa *text* data is similarly constrained. When Eiselen and Puttkammer (2014) compiled the NCHLT isiXHosa text corpus from (all) isiXhosa textual resources publicly available online, they commented on the lack of breadth of coverage in terms of domain, style and genre in those resources available. As a result the corpus is mostly drawn from "South African government websites and documents, with some smaller sets of news articles, scientific articles, magazine articles and prose" (Eiselen and Puttkammer 2014). This corpus therefore over-represents both the standard variety, and likely also technical and/or legal topics while, crucially, under-representing both informal language use (e.g., code-switching, spelling variation) and topics more commonly discussed among friends in voice messages (e.g., hobbies, conversations about families and friends, etc.). This is reflective of isiXhosa's limited presence on the internet (in 2014 and now), in part due to many multilingual users' preference to use English in this domain and in text-based mobile communication like SMS (Deumert and Masinyana 2008).

Compiling more appropriate language resources Whilst there are *some* resources more demonstrative of conversational speech – like the Soap Opera corpus (van der Westhuizen and Niesler 2018) discussed in more detail below – there are too few of these data-sets, and they contain too few hours of speech, for robust model training.

One way to supplement (speech) data in such situations is turn to the speech /user community and compile a new dataset with their help. Reitmaier et al. (2023) do this in two steps: first, a new speech dataset was collected using a public recording device which prompted local residents to share their experiences of the COVID-19 pandemic in short stories. Then, some members in the community were asked to transcribe these stories using a bespoke mobile app (Reitmaier et al. 2023).

Perhaps because of the personal nature of these stories, this speech data *does* contain clear examples of code-switching and has therefore potential utility for future ASR develop-

[12] English Translation: Another one without reason

Table 4.1: Examples from Reitmaier et al. (2023)'s study, illustrating transcription variability of samples taken from their COVID-19 stories dataset and from the soap-opera dataset (van der Westhuizen and Niesler 2018).

T	Transcript
1	kuthiwa abantwana abaninzi bakhula ngaphadle koTata especially boys are more p to the highest behavior...
2	Kuthwa abantwana abaninzi bakhula ngaphandle kotata, especially boys are more prone to high risk behaviour...
3	Kuthw'abantwan'abaninzi abakhula ngaphandle kotata especially boys are more pairing to harsh behavior...
1	then kengoku andakwazi ukuya eskolweni ...nda Quarantiner for ifourteen days ...
2	then kengok andakwazi ukuya eskolweni ...NDA quarantiner for 14 days ...
3	then kengoku andakwazi ukuya esikolweni ...ndakhwaratina for fourteen days ...

ment in the region. However, Reitmaier et al. 2023 also uncovered challenges, especially with respect to crowd-sourcing transcription from the community. They found extensive variation in the resulting transcripts: see the example in Table 4.1 for instance. This supports the notion that local writing practices diverge from the formal standard.

On a practical level, this variation is not amenable to immediate integration into ASR development pipelines; recall the requirement for just one "gold-standard'' transcription both during training and testing. However, there is no single "gold-standard" if users propose and accept different ways of transcribing the same utterance.[13] This variability, coupled with our prior knowledge about the status of the isiXhosa orthography suggests that users might not *like* a system which uses the formal written norm, both because it represents their language variety poorly and because of its associations with the institutions that shaped it.

While Reitmaier et al. (2022) made connections between transcript variability and language standardisation, they did not further systematically unpack these connections, nor involve community members further to understand these subtle variations and their implications for ASR system design. We set out to do just that by involving local residents exploring appropriate transcription standards for the ASR system.

Workshop: Understanding transcript variation

Here we present some insights from 5 participants who were asked to reflect on transcriptions variants provided by other Langa residents. Importantly, we have no reason to believe that the variation observed in the transcripts would be the result of transcribers misunderstanding or not attending to the task. As the discussions at the workshop confirm, the variation in spelling can also not be simply put down to "spelling mistakes", but rather represents a form of language variation (Bucholtz 2007). They were asked to comment on which of several transcriptions they deemed "the best" and elaborate on their reasoning for this evaluation. As expected, the participants did not agree on which transcript variant was "the best" for most examples. However, the discussions did surface some shared concerns around word segmentation, and

[13] Of course, "annotator disagreement" is not limited to transcriptions, but an inevitable feature of language datasets. NLP researchers have developed different approaches to account for this variation. For example, Plank et al. (2014) incorporate inter-annotator disagreement on to train a part-of-speech tagger.

the treatment of English and non-standard words.

Word segmentation One notable difference between many transcripts is how words are segmented (see Table 4.1). The same sound sequences may be chunked into shorter sequences by white space, joined with apostrophes or represented as one uninterrupted character sequence. The workshop participants do not always agree on the how words should be segmented, or how important "correct" segmentation would be. One participant explains that segmentation is highly contextual as it depends on speech rate: the same sound sequences should be concatenated when someone is speaking fast, but separated by white space when they're speaking slowly. This type of spelling variation would be quite unusual in (most varieties of) English. The typology of isiXhosa is also relevant here. isiXhosa is an agglutinative language. Relationships between words are indicated through stringing together several units, so-called morphs, which carry specific grammatical meanings (agglutination) and combining several grammatical meanings into one morpheme (fusion). As a result, word segmentation is more flexible as boundaries could be placed between individual morphs.

Non-standard speech Discussion with participants also surfaced the tension between the desire for transcripts which reflect exactly what has been said and those which follow prescriptive norms. As one participant notes when discussing the difference between <kengoku>[14] and <kengok>: "she's talking slang, she's not saying the full word" but "[some of the transcribers] added a <u> because how it should be said and written". Here there's a clear notion of how this word "*should be said*". However, this particular participant argues for a faithful transcript, i.e., one in which the phonetic reduction in an informal register is preserved in writing. Other workshop participants disagree, citing the "missing letter" as the reason why <kengok> is incorrect. This point of disagreement is particularly interesting as it highlights two important choices involved in transcription (manual and automatic). First, there's the question of how to deal with non-standard speech (or "slang"), and secondly, a question of whether the intended meaning or the verbatim speech is more important. This discussion also extends to the way hesitations, filled pauses and repetitions should be treated in transcripts.

Code-switching Like in many other linguistic communities, code-switching in isiXhosa is stigmatised (Dantile 2015; Deumert et al. 2021). Varieties of isiXhosa spoken in the Eastern Cape, an ancestral homeland of the amaXhosa people, have a lot of prestige on the other hand (Deumert 2010). These attitudes were also reflected in our workshop. One participant comments that he believes one of the speakers to be "from the Eastern Cape", adding "so he's speaking clear Xhosa". This speaker, he further says, would likely only use English to borrow words which do not exist in isiXhosa (like "[hand] sanitiser") rather than code-switch like speakers of "township Xhosa".

[14] We adopt the convention of representing writing in <angled brackets> (orthographic transcription) and speech in /slashes/ (phonemic transcription) .

In code-mixed utterances, English words are often carefully embedded into the utterance according to isiXhosa grammatical rules, for example through affixation of appropriate grammatical markers. One participant explains when discussing the difference between the spelling variants <for fourteen> and <fori ifourteen>: "that's what we do that's our English-Xhosa, so when we're mixing English and Xhosa [...] we add an /i/ just to make it sound as if it was Xhosa when it's not". Not all of this variation is equally accepted by all participants, however. One is unsure about accepting <fori fourteen> since "there's no word such as <fori> in isiXhosa but the person is actually saying /fori/" highlighting again the tension between verbatim transcripts and prescriptive standards. Another participant is certain that <fori fourteen> "it's completely correct because it's how she's saying it" but qualifies this statement with: "But then I'm Xhosa so I would understand it".

Another point of discussion was the word *quarantine* represented as <quarantiner> or <khwarantina>. One participant explained "the person just used the Xhosa way of writing it – or not the Xhosa way of writing it but how I'd write it if I was writing it in Xhosa without translating it". However, they also argued that "because it's an English word it makes more sense" to spell it according to English norms. Another participant disagreed, arguing instead that only the Xhosa spelling is correct and that the other variants are anglicised. A third participant is more diplomatic, noting that "it depends" and explaining that "Xhosa is more of a sounding language – exactly how it sounds is how you write it. [...] Both of them are correct – I'm not Xhosa though I can write like that but I'd use the <quarantine>."

Implications for ASR design and evaluation

The insights from this workshop confirm that the variation in spoken and written language use we observe among isiXhosa speakers in Langa must be considered during ASR system design and evaluation. The frequent code-switching in informal speech means that an ASR system needs to be able to handle input from isiXhosa and English[15].

Regarding orthographic variation, the workshop feedback somewhat confirms to us our above intuition: that users might prefer a system which does not strictly reproduce the formal isiXhosa norm. While adherence to the norm is important in some contexts, and they all make reference to some "spelling errors" in the transcript there is also quite a lot of awareness and acceptance of spelling variation. The lack of consensus on how words should be segmented and how to represent English words embedded in Xhosa utterances, is particularly interesting from an ASR design perspective.

These clear implications for the ASR development highlight the advantages of involving users, or, in the context of language technologies, speech communities, throughout the technology development process. Engagement with user communities is a key area where the technical aspects of CUI development, and the speech and language technology engineers

[15] Ideally a system would be able to accommodate even more local languages, but in part 2, we limit ourselves to isiXhosa and English.

responsible for it, can benefit from interdisciplinary interaction between human-computer interaction, linguistics and speech technology. This is particularly clear when designing alternative evaluation approaches which centre user perspectives, rather than "objective" metrics commonly used in speech technology development, as discussed below.

4.4.4 Part 2: Integrating language resources and user perspectives

From Part 1, we know that an ASR system which assumes monolingual speech as the input and one "standardised'' orthographic form as the desired output would emulate neither Langa's real-life language usage nor transcription preferences. Following the insights from Part 1, we aim to build an ASR system which can:

1. easily recognise English (and in principle other languages) in addition to isiXhosa, thus facilitating recognition of more complex code-switched and multilingual speech
2. represent the recognised speech to the end user in a manner which adheres less strictly to the "dictionary orthography"

Building ASR Systems

Using the Kaldi toolkit (Povey et al. 2011), we built a suite of ASR systems which vary systematically in accordance with the above two design features.

All systems are hybrid in architecture, meaning they comprise a neural-network based 'acoustic model' (AM) and an n-gram 'language model' (LM) component, united in a hidden Markov model (HMM) framework. Importantly, in this paper, it is during the LM building process that these varying design choices are realised. The training/building of the acoustic model was thus consistent across all systems.

For this acoustic model component, we employed transfer learning: a technique common to low-resource ASR [16]. That is, rather than training from scratch, we began with an existing pre-trained English AM trained on 500 hours of speech from the British English Multi-Genre Broadcast (MGB) (Bell et al. 2015)), which was then fine-tuned to the South African Soap Opera (SO) dataset (van der Westhuizen and Niesler 2018). This training method is integral given the insufficient size of the SO data-set; we could not simply train on it from scratch. The SO corpus consists of acoustic speech data and the corresponding 'gold-standard' transcripts from five languages: isiXhosa (xho), isiZulu (zul), Sesotho (sot), Setswana (tsn)—all belonging to the same Southern Bantu family of languages—as well as English (eng). The choice of an English pre-trained model was partly driven by practical concerns – the multi-genre nature of the data makes the model naturally robust to a wide range of acoustic conditions – and partly by the need to handle the English code-switching. Similar constraints are likely to face other ASR developers working to build systems for such low-resource languages.

[16] We use 'low-resource' here as it is defined in the ASR field – a label for a language or context for which there is little training or test data available, though see Bird (2022) for a critique

As noted in Section 4.4.2, the language model is the component responsible for dictating an ASR system's output vocabulary, that is, the words can the system actually recognise, and the orthographic form of their representation. We built a set of varying language models.

Baseline system We collated the following text data-sets: the isiXhosa component of the CommonCrawl corpus[17]; the NCHLT isiXhosa Text corpus; the isiXhosa component of the Leipzig Corpora Collection (Goldhahn et al. 2012); and the isiXhosa component of the Memat corpus (Tiedemann 2012). We then used the SRILM package (Stolcke 2002) to train a 3-gram language model from this training data. This model's main propensity is thus to output isiXhosa, although it may be able to generate occasional English words or phrases in specific contexts, if they happened to appear in the source training text.

Recognising more English We know code-switching between isiXhosa and English is common in Langa conversational speech. We also know from Section 4.4.3 that at least some end-users accept and even expect English vocabulary (with standardised English spelling etc) to appear within an ASR system's textual output. To boost the ASR system's capacity/ability to recognise lengthy spans of English as well as isiXhosa, we augmented our baseline LM with a second LM trained on English text data from the South African English component of the Leipzig Corpora Collection (Goldhahn et al. 2012).

Flexible orthography Recall from Section 4.4.3 and Table 4.1 the variation in how transcribers spelled and segeented words. This freedom in orthographic representation seems entirely at odds with the traditional approach to LM-building, where the graphemic output is constrained to a pre-defined, closed vocabulary list.

One way to better facilitate this freedom is to build our LM at a sub-word level: that is, teach our model to recognise and output sequences units smaller than words as opposed to full words. Sub-word sequences can then be combined/collapsed together in a post-processing step. This allows for a more flexible approach to orthography and word segmentation than strictly only modelling full words as seen in the training data. For instance, if our ASR system were to produce the sequence 'mbileyo be@@ si@@ bo@@ pha imi@@ khwa ko@@ profeti' (where '@@' is used to signify subwords that are not at a word boundary), this would be collapsed to 'mbileyo besibopha imikhwa koprofeti' before being presented to the user. Note that while the final transcript is at a word level as in our other systems, this method means that this model in effect operates with an open word vocabulary.

We employed Morfessor (Virpioja et al. 2013) to learn an optimal sub-word inventory from the same text data used to train our baseline LM. We learnt two variations: a 10k unit inventory, and a 2k unit inventory. Once learnt, we generated morphologically decomposed versions (2k and 10k respectively) of the LM training data, and used these to train our sub-word LMs.

[17] `https://commoncrawl.org`

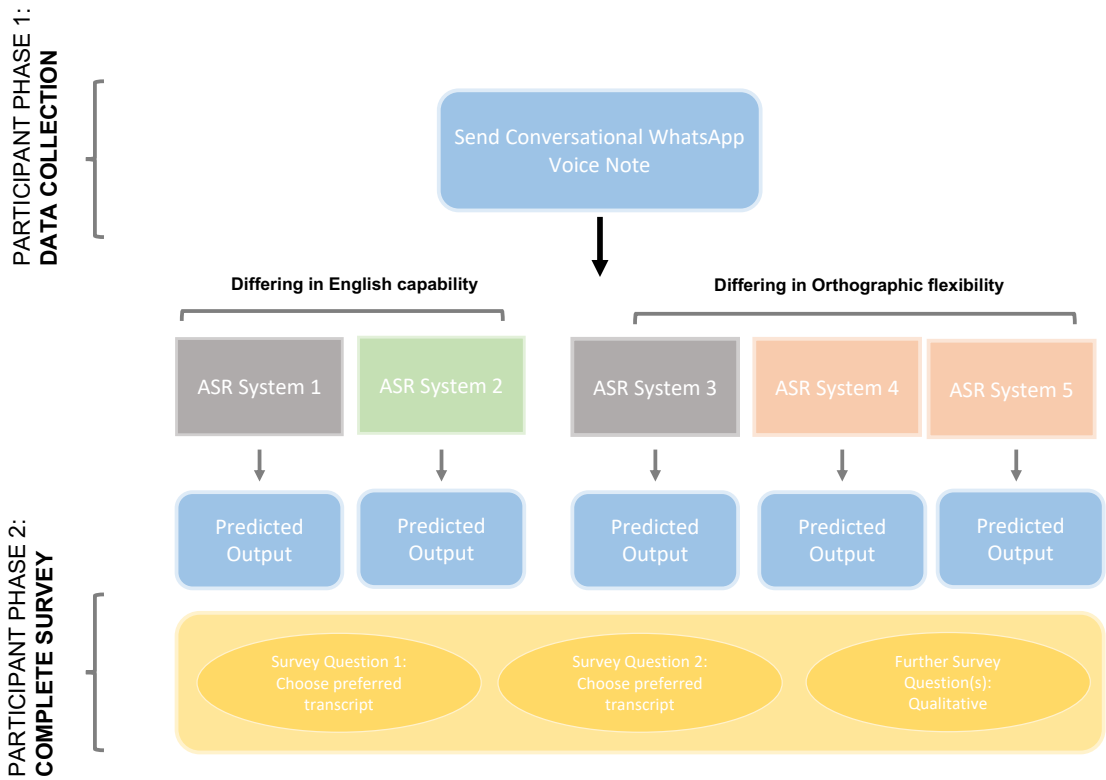


Figure 4.3: The study was conducted in multiple phases. Note the ASR system(s) shaded grey is that which uses the Baseline LM described in Part 4.4.4; the system shaded green uses the Extra-English LM; and those shaded orange use the sub-word LMs.

Note that, to produce sub-word sequences, we trained 8-gram rather than 3-gram models to maintain approximately the same modelling context.

ASR Systems evaluation

Survey The previous section detailed the changes we made in ASR development in line with Section 4.4.3's findings. In this section we ask: do these changes actually aid ASR performance in the eyes of the Langa community? That is, are the transcripts produced by these additional systems (generally) considered more acceptable or of better quality than those that would be produced by a naively-built system? And is there consensus on this or do opinions vary?

Paralleling Part 1, we asked the target users – Langa residents – to evaluate these systems via a remote study. The study involved 14 participants and consisted of two phases (see Figure 4.3). These 14 participants were recruited by our Langa contact, a market researcher we have previously worked with, from a pool of community members who have experience of such studies and, specifically, had participated in prior workshops regarding ASR development for isiXhosa.

In phase one, pairs of participants who were friends or acquaintances were asked to exchange short, casual WhatsApp voice messages about what they had done the previous weekend. They were also be asked to send a second WhatsApp voice note to one of the researchers (having been told that the researcher spoke isiXhosa). We then transcribed two messages per speaker (one for each recipient) using our ASR models, thus producing four "candidate'' transcripts per "message'').

These generated transcriptions were integrated into a bespoke survey for each participant, distributed (via WhatsApp) about 10 days after they had submitted their voice notes. The second phase of the study involved the completion of this survey by the participants. Specifically, participants would be presented with a reminder of a voice note they had submitted, alongside all four candidate transcriptions. The task was to rank these transcripts (and thus, by proxy, the ASR systems responsible for generating these transcripts).

Findings

The users' response to a system with increased English capability Figure 4.4a compares how participants ranked our Baseline ASR system with our Extra-English system; it illustrates that there was a slight preference for the latter, though opinions were mixed. If we focus on individual participants' preferences (and in parallel, listen to their submitted voice recordings) however, a clear correlation becomes apparent: those participants who included more English in their recordings – specifically, whole phrases instead of single, isolated words – nearly always preferred the Extra-English system over the baseline, and vice versa. For instance, consider the following two candidate transcripts generated in response to one participant voice note, which, from listening, we know to contain a significant number of English phrases:

(1) a. ...but most interest thing ingade kule weekend waza creativity iwebsite

b. but the most interesting ingade kule weekend was creating a website

Transcript 1a was produced by the ASR Baseline; we see that, whilst there are some English words (which are often close to the target semantically and phonologically), this transcript misses out on the correct syntax, morphology and function words, all of which are accurate in Transcript 1b (produced by the Extra-English system). This is because we explicitly trained the Extra-English system on more English data.

We can compare this to the following pair of transcripts, generated for another participant's voice note:

(2) a. abantu belizwe ndiyishiyile bathe bakufika ichiza

b. abantu beza beneficial him the bafiki cheese

Here neither transcript features the kind of long grammatical English phrases evident in the example above and, from examination of the audio, we can confirm that the voice note only contained isolated English words. The 'increased' English capacity of our 'English' model then actually hinders, more than helps here; in Example 2b, we see a transcription which includes English words even though the original audio only included isiXhosa words. This is intuitive: differences in the use of code-switching between users is related to differences in their opinions as to whether 'increasing' English capacity improves performance. Models with increased capacity for recognising English are helpful in utterances with long English phrases and for users who tend to code-switch more. Finding a way to estimate the probability of code-switching (by utterance interaction, or user), might be particularly useful in allowing the appropriate deployment of different models.

The variation in attitudes towards this model might also be accounted for through other factors. As highlighted in the workshop, users differ in their opinions on whether English words should be transcribed according to English or isiXhosa conventions. It would not be surprising if the participants had stronger opinions to deviations from the orthographic norm in English words, as they are frequently exposed to written English (e.g., on public signs and in education). Users might also be more sensitive towards errors where a language is effectively "misidentified".

User response to a system with more flexible orthography Figure 4.4b demonstrates that our pool of users preferred our sub-word systems to the baseline, despite its ability to generate unseen word forms. While this would be highly surprising in the context of a standard variety, these results are in line with our findings regarding variability in word segmentation. As we hypothesised above, just as variation in word segmentation and spelling is acceptable to speakers in manual transcripts, so is the output of the sub-word model. While this model still concatenates sub-words into larger units, it can produce a much larger number of different words, thus modelling the orthographic variation we see among the human transcribers.

Evaluation 'in the wild' versus Evaluation 'in the lab' Consulting the end user is not a typical method for ASR systems evaluation. The standard evaluation procedure involves distance-edit metrics which compare the generated transcript to a "ground truth", i.e., the aforementioned "gold transcript". The most common metrics are Word Error Rate (WER) and Character Error Rate (CER) which provides a ratio of erroneous words or characters respectively, and are usually reported as percentages. These metrics have clear advantages when comparing models or algorithms (on the same data): they are fast and easy to compute, and easy to interpret. In a development process which requires frequent evaluation this method is very useful. However, these metrics can fail to capture the context and impact of specific errors, and do not always correlate closely with user evaluations.

Considering the metrics' intrinsic bias towards outputs which confirm closely to a "gold

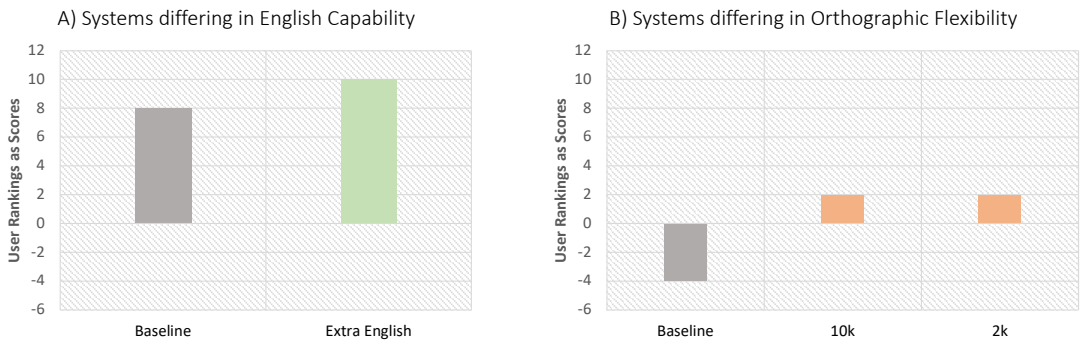


Figure 4.4: Comparing how survey participants ranked ASR systems differing in key design choices. To gauge overall user preferences (across all participants and all survey questions), we convert system 'rankings' into scores. When a transcript is ranked highest during any of the survey questions, the system responsible for that transcript gains a score of 1. For those survey questions asking the user to rank transcripts produced by the systems differing in 'orthographic flexibility' – a three-way comparison – we additionally assign a score of -1 to the system whose output was ranked lowest.

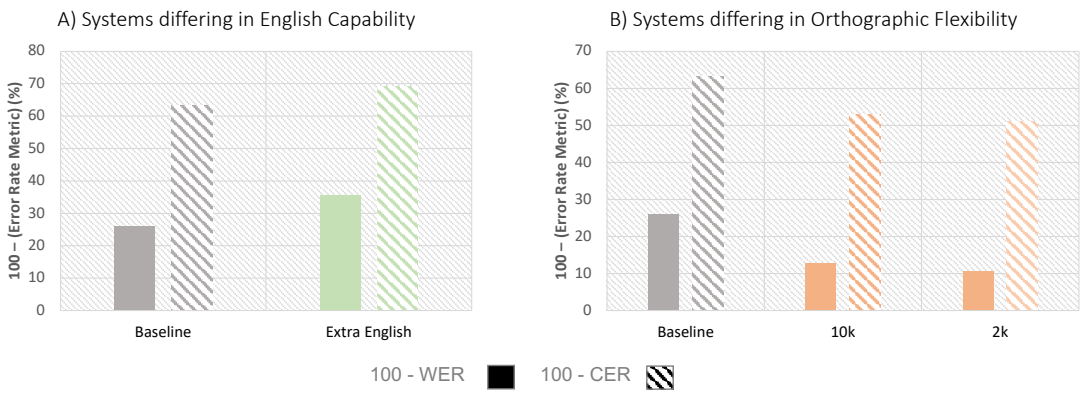


Figure 4.5: Comparing how WER and CER metrics ranked ASR systems differing in key design choices. Note we report 100 - (Error Rate) to facilitate comparison with the user evaluation (higher is better)

transcript" and our lack of such definitive "gold transcripts" in this context, we expected there to be a mismatch between user evaluations and these quantitative measures. To calculate WER and CER for each of our isiXhosa systems, we use the the test-set component of SO dataset (van der Westhuizen and Niesler 2018). Figure 4.5 presents these results. For all systems CER is lower than WER. This is unsurprising as WER captures word-level deviation from the reference (e.g., in segmentation or spelling), while CER captures character-level deviation. As expected, we do see a clear difference between the metrics and the user evaluation: the subword models perform poorly according to the metrics but are preferred by the users[18].

While some criticisms of standard evaluation metrics (like their potential disconnect from user experience) hold in many contexts, as we tried to highlight here, they seem especially inappropriate and misleading in contexts where selecting a "gold standard" accepted by all users is difficult. While it is perhaps not feasible to ask users about every single aspect of a model, we believe that user evaluation throughout the process has clear benefits. Decisions regarding "higher-level" design choices which clearly correspond to specific model behaviours such as whether to use a word or subword model or what kind of training data to draw on, perhaps should involve users.

4.4.5 Discussion

Reflections on remote survey-based evaluation of automated transcripts

Comparing between users We decided to present users with automatic transcriptions of their own voice message for two reasons. One, we wanted to ensure that users are providing us with their opinions on the transcription, not the content of the message or any judgements of the language used. Especially when probing for potential "mistakes", it was important to ensure that participants were not reporting stigmatised features they perceived in other people's speech – such as code-switching or the use of a particular accent or dialect – as "linguistic mistakes". We also wanted to give users an impression of what the technology could do for them personally. Performance of ASR systems can vary depending on a wide range of linguistic and non-linguistic factors such as recording quality and background noise. By encouraging people to evaluate the system in a way that is very close to the "real" use case, we hoped to elicit more useful evaluations.

Of course there are also downsides to this approach. First, users are generally unlikely to apply ASR tools to their own speech, but rather, that of their conversation partners. Second, being familiar both with what they said and what they intended to say, they might evaluate the system more harshly than if they were unaware of the intended message or unfamiliar with the other person's accent, for example. The performance of the our ASR model likely also differs between speakers, depending on their accent, the prevalence of code-switching and non-linguistic factors like background noise.

[18]Notably, there is not much difference in performance by the two subword models. This suggests that increasing the number of subword units does not necessarily improve performance according to users.

Language and questions We decided early in the study design process to distribute surveys in isiXhosa, rather than English. This was to ensure people felt comfortable using isiXhosa in their voice messages, rather than opting for English to accommodate us. The survey questions and response options were translated by our local facilitator. One limitation of this approach is – ironically given our broader aims – the risk of using an inappropriate register in both our questions and the participants' responses.

Presenting full transcripts in isolation We also chose to present participants with transcripts of whole voice messages, rather than short snippets. To ensure that all participants would be able to easily use the survey (regardless of mobile device), we opted for the simplest possible interface and did not ask participants to comment on each error separately, but to evaluate the voice message transcription as a whole. This has the advantage of being more realistic, as the ASR system would be used to transcribe whole voice messages in deployment. The accuracy or appropriateness of a transcript might also be more difficult to evaluate for a short utterance taken out of context. On the other hand, the evaluation is less precise as it is likely that each transcripts contains several errors, not all of which are related to the same design choice.

Sociolinguistically-informed considerations in ASR development

Based on our experience working on isiXhosa ASR, and working with isiXhosa speakers in Langa, we argue that taking account of language variation, and the way different spoken and written varieties are evaluated by potential users, is crucial for effective language technology design.

Consideration 1: "real-world" language use Perhaps the most obvious point to start is to establish how most people who might use the ASR system actually use language. In multilingual contexts, this is likely to involve code-switching. The specific application context might also affect the users' speech style and degree of accommodation towards the ASR tool. Informal speech, and speech directed at other community members, is likely to be characterised by phonetic variation and a higher density of "dialect" features.

A particular challenge here is that much of the speech data available for ASR development, is not similar enough to the "real-world" use. Successful ASR systems might therefore require the compilation of new, domain-specific speech datasets.

Consideration 2: local writing practises For systems designed to create transcripts of speech (rather than recognise single commands, for example), understanding if and how local writing practises diverge from the written standard is important. Generally, spelling variation is more likely in computer-mediated, informal writing like text messages and on social media. In contexts where the standard is very contested, not as widely used, or not as widely transmitted to

speakers, spelling variation is particularly likely. In this case, users might prefer systems which mirror this variation in their output.

The availability of appropriate datasets of unpaired text or manual transcriptions is again a limiting factor in accurately modelling this variation. Most text datasets, especially for "under-resourced varieties" are drawn from a small number of more formal genres (e.g., government documents, newspaper articles).

Consideration 3: appropriate evaluation strategies Involving users in the evaluation of ASR systems is particularly important in contexts where "real-world" language use diverges from the kind of language reflected in datasets which would traditionally be used for evaluation. In addition to in-person workshop, online surveys can be useful here. While we only involved a small number of participants in the evaluation process, this would be scalable to a larger sample size.

4.5 Discussion: Orthography, (De)standardisation and ASR in daily life

4.5.1 Benefits of ASR-based transcription in specialist and non-specialist contexts

Transcription in the age of big speech data

As the storage and collection of large amounts of speech data has become easier and cheaper, several subfields of linguistics, such as acoustic phonetics, psycholinguistics and sociolinguistics have embraced the idea of working with larger corpora of speech (Liberman 2019). There are some clear advantages to using larger datasets in the quantitative analysis of language variation and change. Rare phenomena are more robustly represented and variation and change across time and space may be more easily observable. In addition to established methods of data compilation in variationist sociolinguistics like sociolinguistic interviews conducted by fieldworkers with individual participants within a speech community, recent years have also seen the rise of new methods. While self-recordings created for either public audiences or just the researchers have been used in variationist and in particular sociophonetic research for several years (Schøning and Møller 2009; Hall-Lew and Boyd 2017; Leemann 2016; Leemann et al. 2018; Clark et al. 2016), and facilitating remote recordings with speakers became for many research groups the only way to gather speech data during the COVID-19 pandemic. Some projects developed specifically in response to the pandemic (Sneller 2022), including MI Diaries (Sneller et al. 2022) and the Lothian Diary Project (Hall-Lew et al. 2022). These new ways of compiling data to analyse variation appear to be here to stay. Developing efficient data processing and data analysis pipelines is essential to make good use of these incredibly

rich datasets, especially in interdisciplinary contexts where we (or perhaps our collaborators in different fields) might be just as interested in *what* is said as we are in *how* it is said.

As (socio)linguists, we should be particularly aware of the ways language technologies can fail. As noted above, algorithmic bias is a real and urgent problem in speech and language technologies (see also Blodgett et al. 2020; Koenecke et al. 2020; Bender et al. 2021). In practice, this means that speech and language technologies tend to be designed for prestigious language varieties and perform worse or not at all for marginalised groups. Another important shortcoming of ASR technologies today is their limited capacity to deal with conversational speech, noisy backgrounds, code-switching and multiple and overlapping speakers. In other words, many of the speech styles that we as sociolinguists are *most* interested in are also the *most challenging* for current ASR technology.

Nevertheless, our experience with the Lothian Diary Project suggests that incorporating automatic speech recognition can significantly reduce the time involved in transcription. As I've outlined, there may be practical reasons to opt for custom or off-the-shelf (commercial or open-source) systems. Important considerations here could be sharing permissions of the recording data, the varieties and speech styles used by the speakers, local computing resources and programming skills, cost of various options, among others. For example, for recordings of read speech in a "standard" variety of English, ASR transcripts might be very good. If the recordings contain multiple speakers with a significant amount of overlap, identifying and separating utterances by different speakers ("speaker diarisation") may be very challenging in and of itself. Of course, some of the contexts that are most difficult for an ASR system are also most difficult for human transcribers.

Embedding automatic speech recognition systems in (socio)linguistic research workflows could reduce the time involved in preparing orthographic transcriptions. Based on experiences with the Lothian Diary Project, I'd recommend working where possible with experts in speech technology to establish what the most appropriate tool is depending on available resources, skills and the particular dataset. As linguists we are particularly well-placed to use these technologies responsibly with a clear understanding of their limitations, and to contribute to improving them.

Transcription in the age of WhatsApp voice messages

As noted above, short voice messages sent via messaging services like WhatsApp are extremely popular in many language communities (see e.g., Matassi et al. 2019). As isiXhosa speakers in Reitmaier et al. (2022) note, voice messages are particularly useful for them because not everyone in their community is comfortable reading isiXhosa. In this way, they perceive them as more accessible than text messages (which is an interesting contrast to the perception of voice messages as more of an imposition in other contexts with different reading and writing practices, like among US English speakers) (Reitmaier et al. 2022). However, they also identify some disadvantages of voice messages which automatic transcription could alleviate:

it's much harder to retrieve specific information from multiple or long voice messages than text, sometimes listening to a voice message might be inconvenient (e.g., regarding privacy or noise), and because of their comparatively larger size, voice messages are more difficult to store on mobile phones long-term than text messages (Reitmaier et al. 2022).

4.5.2 Automatic transcription and (de)standardisation

Choices regarding which language varieties to support and how to represent them, are never neutral. In the case of isiXhosa, it is clear that the existing written standard is contested. Previous research has shown that many isiXhosa speakers, especially in Cape Town, feel alienated by the written norm. Unlike English and Afrikaans, most Langa residents do not frequently encounter isiXhosa in official writing or on public signs, or through formal education (Deumert 2010; Dantile 2015). The "urban" variety spoken in Cape Town is further considered to be very different from the "original" or "traditional", prestigious isiXhosa varieties spoken in the Eastern Cape in particular because of its characteristic code-mixing (Deumert 2010). Adopting the formal isiXhosa standard to transcribe, in particular, informal speech of Cape Town speakers would not only be inappropriate due to the inability to straightforwardly handle phonetic variation and code-switching but would also be a meaningful intervention in this sociolinguistic context. It would introduce the formal isiXhosa orthography into the new, personal domain of voice messages where it might be particularly unwelcome and alienating.

The choice of graphic representation here (and elsewhere) has several potential impacts, which I will discuss in turn. Firstly, where (any) written standard is selected, informal speech will be "standardised". Secondly, the selection will likely impact the status and perhaps use of the selected written standard. However, in acknowledging the agency of users, we can also imagine (and observe) resistance to these standards and the creation of alternative norms.

"Standardising" spoken variation

In all transcription, there is a central tension between faithful transcription of spoken language, and "more grammatical" paraphrases following a written standard. ASR tools are particularly prone to "standardising" spoken language according to the written norm. This is due to the fact that most systems are explicitly trained on text data which is furthermore often, especially in "under-resourced contexts", drawn from quite formal genres (like government documents or newspaper articles). As a result, artefacts of spoken language like hesitations, repetition and extra-linguistic sounds are often "corrected" by the language model (see also Section 4.3). This language model bias can disproportionately affect speakers of varieties that diverge from the standard, thereby causing predictive bias and attendant harms as discussed in Chapter 3.

Whether or not users perceive this to be a problem, depends on the use context. In our case study of applying ASR to facilitate transcription of sociolinguistic data, this "language model bias" frequently erased details we deemed important like repetitions and filled pauses.

As discussed above, these errors are particularly difficult (and time-consuming) to correct manually, and more likely to go unnoticed. Where more "appropriate" text data (e.g., transcripts of informal speech) is available, this language model bias can be alleviated to some extent. In the context of isiXhosa, our participants also discussed the tension between transcribing "what someone said" and what would be "correct". Here the specific use not just of the ASR tool (generation of transcripts for voice messages) but also the output (transcripts) likely informs user preferences. If transcripts are intended to "replace" the voice message (e.g., to save on a mobile device or read privately in a public space), preserving the informal nature of the speech is likely important to users. If, on the other hand, the main utility of the transcripts is the ability retrieve a voice message with a particular keyword, consistent (or predictable) spelling might be more useful. Since all of these applications are of interest to the users (Reitmaier et al. 2022), who, as we show above furthermore disagree on how "non-standard" speech and code-switching should be handled by the system, it is not clear that any *one* approach is appropriate. Rather, it might be preferable to develop systems which can represent speech in several different ways.

Boosting the status of the (written) standard

The isiXhosa case is perhaps particularly interesting because of the interaction between the standardising effect of the standard variety *and* the effect of the standard orthography. While these of course also go hand in hand in the English language ASR systems discussed in Section 4.3, the orthography of English appears so fixed to the vast majority of English speakers that there is essentially no room for variation in most contexts. Most English-speaking users of an ASR system probably expect the system to produce standard orthography. While the use of standard orthography in English language ASR reinforces orthographic conventions by exposing users to them, it likely does not introduce those conventions to users for the first time or produce output that is notably different (in terms of spelling) from what a human transcriber may have produced.[19] The ASR language model in particular, does however often have a standardising effect on any speech. In the isiXhosa context, there is much more scope for variation in orthography among speakers. This appears to be in part related to varying levels of literacy among speakers related to the way isiXhosa is (or is not) taught in schools, and a certain level of alienation many speakers experience with respect to the written standard and orthography. In this context, ASR is not just reinforcing the standard variety in terms of structure and lexis, but also introducing spelling conventions (including word segmentation) associated with that standard.

As discussed above, isiXhosa speakers in Langa also frequently draw on multiple languages within one interaction or utterance. The written and spoken isiXhosa standard varieties, how-

[19] Setting aside here differences in conventions between standards, such as differences between British and US English orthographies, which certainly are both reinforced and potentially introduced to speakers from different regions.

ever, do not involve this kind of code-mixing. As a result, unless the ASR tool is explicitly trained on text from several languages and/or code-switched utterances, it will further impose a monolingual standard. This monolingual standard reinforces the existing language ideologies held by speakers in Langa reported by Deumert (2010) (and discussed by our participants in the transcription workshop). It's the varieties spoken in the Eastern Cape, the ancestral homeland of the amaXhosa people, that are seen by speakers to be reflective of "tradition" , and, "a superior standard" (Deumert 2010, p. 251). These varieties, which are not characterised by code-switching, are the ones speaker consider to be "proper", "exact", "respectful" and "formal" (Deumert 2010, p. 251). As Deumert (2010) notes, these different "positionings" towards modernity, tradition, urbanity, and rurality (as Deumert (2010) calls them) are linked to social categories and identities dating back to the 19th century and are thus extremely deeply rooted within isiXhosa speaking communities. The key point here is that while the decision to train a language technology on a standard variety or orthography is *never* neutral, since it is as the very least affirming the status quo, it would be a particularly strong intervention in this sociolinguistic context.

Rejection and the creation of new standards

The adoption of a (formal) written isiXhosa standard in the context of speech technologies would correspond to a domain expansion. Recalling the agency of users in their engagement with technologies, there are different ways user could react to this expansion. For speakers who previously encountered it only in (few) formal contexts (e.g., translated government documents, some literature) and who would not use it themselves in writing, its sudden "appearance" in a domain perceived to be as personal as WhatsApp messages among family and friends might be alienating. As discussed above, a system too closely modelled on the standard likely also "misses" a lot of the socially meaningful variation in speech. As our interviews with isiXhosa speakers highlight, the transcription of code-switched utterances is a particularly complicated issue. Taken together, these shortcomings might lead to the rejection of a "standard" ASR tool by many potential users.

If the system is perceived to be "good enough" to be useful for a particular, important purpose, such as keyword search, errors may not matter as much. More frequent exposure to the written standard could lead to wider use and acceptance or even re-appropriation among users. As our interviews and survey study also show, isiXhosa speakers (at least those in our samples) are also open to new, "non-standard" ways of reading and writing informal speech. Among our human transcribers, variation in word segmentation was particularly common and other speakers were not in agreement which segmentation strategies were "correct". Similarly, the participants evaluating our automatically generated transcripts were open towards a system which generated "new" words based on the recognised sound sequences (rather than a dictionary). This points to exciting and interesting avenues in ASR development in contexts where writing practices are – for a variety of complex sociocultural and historical reasons

– more flexible than they are in many "monoglot standard language cultures" as Silverstein (1996) terms them. Here, speech communities (potentially with the help of developers) could create new conventions related to reading, writing and spelling specific to language technologies.

4.6 Conclusion

In this chapter, I considered how the task of producing automatic transcription of informal speech is shaped by (and shapes) its sociolinguistic context. I argued that this task, which could be formulated as conceptually straightforward, is complicated by the fact that neither "orthography" nor "transcription" are atheoretical or asocial processes. Rather, as (socio)linguistic work going back decades highlights transcription is a highly complex process involving numerous decisions with important theoretical and political implications. Similarly, orthography, though firmly codified for (some) standard varieties and in (some) standard language cultures, is characterised by (socially meaningful) variation. Arriving at and making use of an orthography is a social process which is part of larger standardisation processes. These complexities, while always in the background, are particularly obvious in the transcription of "non-standard" speech. It is also in these contexts, that development and deployment of ASR is most interesting and most challenging. This is in part because the availability of "gold standard" transcriptions, and by extension a standard orthography, is central to most current ASR development (especially for evaluation purposes).

Here, I considered two very different case studies. First I discussed the advantages and challenges of embedding ASR in a transcription workflow for sociolinguistic data (in English). I focussed on the trade-offs between "efficiency" (especially in terms of labour involved in creating transcripts) and problems biases introduced by acoustic models and language models. These can be mitigated to some extent through custom systems (which do, of course, bring their own technical challenges). I then turned to the development of an ASR system to transcribe informal voice messages of isiXhosa speakers in Langa, South Africa. Here the sociolinguistic context exposes many biases not just in the output of ASR systems but the way the task and the speakers are conceived of. Firstly, local linguistic practices are much richer than the "monolingual standard variety" on which systems are generally predicated, as speakers draw on multiple languages in their interactions. Due to the violent history and legacy of colonialism and apartheid in which Langa residents and isiXhosa are embedded in, many speakers are further unfamiliar or uncomfortable with the written isiXhosa standard. Building and evaluating an appropriate and useful system for this context requires significant input from the local speech community. I argued that while insensitive design has the potential to do harm (or be useless), the high acceptance for variation in writing we see also suggests opportunities for new, different ways of graphically representing speech.

In the following chapter, I pick up on the question of language data which has been raised

in the papers presented here: What is the language data we use to train and test ASR systems? And why is it that way?

Chapter 5

Language resources, data bias and power

5.1 Introduction

In previous chapters, I have explored the issue of predictive bias (Chapter 3) and considered how specific sociolinguistic contexts and applications (should) shape ASR development (Chapter 4). Language *data* and the way we understand it, is central to all of these discussions. In Chapter 4, I discussed the limitations of the available spoken and written "language resources" available for isiXhosa – and the ways we attempted to create datasets which more closely represent "real language use" by directly involving the intended user community as a speech community. Some of the challenges uncovered there are specific to isiXhosa, like the history of and attitudes towards the standard orthography. However, both application contexts discussed in Chapter 4, also surface much more general limitations of existing language resources. The poor performance on non-standard varieties and informal speech shown in Chapter 4 suggests that commercial ASR systems for British English are trained on speech and text datasets which over-represent standard varieties.

In this chapter, I focus on the role of this *data bias* as an origin of predictive bias in ASR tools. I am particularly interested in exploring how language ideologies and wider power structures shape what and who is and is not included in speech datasets used to train and test ASR systems. Understanding these "data gaps" enables the anticipation of predictive bias and attendant harms. Importantly, framing data compilation as a social process also highlights the agency of individual actors to mitigate bias by being more reflexive. I begin this chapter by briefly recalling the types of data involved in training and testing ASR systems and discussing the notion of data compilation as a social process.

I then present a paper I co-authored with Stephen Joseph McNulty, a sociologist of language working on language policy, which we presented to an audience of speech technologists at the Language Resources and Evaluation Conference in 2022 (Section 5.3). The aim of this paper

was to introduce the concept of *language policy* to developers who likely had not previously encountered it, and make the case that this is a useful lens to think about both data bias and the way language ideologies influence language resources. Specifically, it also allows us to understand the role and agency of language technology researchers and developers when compiling and using language datasets.

I then present a paper published at the workshop on Language Technology for Equality, Diversity and Inclusion at ACL 2022 (Section 5.4). The paper is aimed at an audience of language technology developers. I draw on *Data Feminism* as conceptualised by D'Ignazio and Klein (2020), as well as other feminist theory and critical research on datasets and dataset documentation to develop a documentation framework aimed at identifying and interrogating *gaps* in speech and language datasets. I apply this framework to popular licensed and open-source English language datasets.

Finally, in Section 5.5, I reflect on the problems associated with framing language data as a "resource" and some of the dominant values and goals underpinning much of current language technology development.

5.2 Background

5.2.1 Datasets in training and evaluating ASR systems

Most ASR systems are trained in a "supervised" manner and have two components: an acoustic model and a language model. These systems require labelled training data to train an acoustic model: speech recordings paired with "gold" transcriptions. The language model additionally requires unpaired text data. For example, the models developed for isiXhosa described in Chapter 4 have these requirements. As discussed in Chapter 4 and Chapter 6, these data requirements pose a particular challenge for "under-resourced" or "low-resource" varieties, where transcribed speech data and/or large amounts of machine-readable text in an appropriate domain is hard to come by. The most recent "state of the art" ASR models involve "weakly supervised" (Radford et al. 2022) and "self-supervised" (Zhang et al. 2023; Baevski et al. 2020) approaches to training ASR. These architectures can leverage large amounts of unlabelled (i.e., untranscribed) speech data and unpaired text, combined with much smaller labelled datasets in training. For example, Zhang et al. (2023) (Google) present a single ASR model which can be applied to over 100 languages. [1] At the cost of these very large data (and computation) requirements, these approaches generally outperform conventional models on both "high-resource" and "low-resource" varieties.

ASR systems, like other machine learning systems, are generally evaluated using benchmark datasets. Depending on the setup, these might be subsets of the training dataset not used during training (held-out test sets) and/or separate datasets only used for testing. Bench-

[1] See also `https://ai.googleblog.com/2023/03/universal-speech-model-usm-state-of-art.html` for a less technical summary.

mark datasets are meant to *exemplify* a particular machine learning task to enable reliable evaluation (Schlangen 2021). With the proliferation of different ASR use cases (e.g., dictation, transcription of meetings, transcription broadcast media) and a much wider range of users, acoustic environments, speech rates and domains, robust training and evaluation becomes more challenging (Aksënova et al. 2021), as models trained and tested on a single benchmark dataset tend to generalise poorly to other datasets – especially any that differ in domain (e.g. style, variety) (Likhomanenko et al. 2021).

Differences in style are extremely important here: word error rate for read (recited) speech can reach "human-like" performance for high-resource standard varieties, while spontaneous speech remains very challenging (Gabler et al. 2023; Szymański et al. 2020).[2] One key reason for these performance differences are the stylistic differences between recited and spontaneous speech which are well-documented in the sociolinguistic and phonetic literature: spontaneous speech involves more phonetic reduction, a faster speech rate, and, especially in relaxed conditions, speech production much closer to speakers' "vernacular" (Labov 1972) (see also Section 2.3.1). This is why sociophonetic data poses particular challenges to ASR, as discussed in Section 4.3. Recited speech further conforms (per definition) very closely to the written standard. Since almost all ASR architectures make use of a language model derived from *text*, not speech, there's an inbuilt bias towards the written standard in almost all systems. Many English ASR systems are furthermore (at least partly) trained on Librispeech, a corpus of recited speech drawn from public domain audio books (Panayotov et al. 2015). Results obtained for recited speech are therefore likely to be much better than those for transcribed speech.

But even "spontaneous" speech corpora have key shortcomings when used to estimate ASR performance "in the real world". As Gabler et al. (2023) point out, planned transcribed speech is easier to process than unplanned speech as it is more likely to adhere more closely to a prescriptive standard, less likely to contain repetitions and fillers. I will return to the challenges of compiling "realistic" corpora of conversational speech for language technology development in Chapter 7.

5.2.2 Data compilation and agency

While terms like "data collection" and "raw data" imply that "data" already "exists" to be "collected", the reality is both more complicated and more interesting. As Bowker famously put it "raw data is an oxymoron and a bad idea" (cited in Gitelman 2013, p. 1). Rather than being discovered or collected, data is socially constructed. In the papers presented below, I want to draw attention to the way the process of data compilation impacts data bias, which in turn can cause algorithmic bias. I follow Benjamin (2021) in using the term "data compilation". While Benjamin (2021) focuses primarily on the context of data protection and privacy, their

[2] Gabler et al. (2023) propose describing ASR training in terms of continuous axes: *planned* to *unplanned* speech on the one hand, and *recited* and *transcribed* speech on the other. This allows for a useful distinction between planned transcribed speech (e.g., news broadcasts) and unplanned transcribed speech (e.g., spontaneous conversation) as well as planned recited speech (e.g., reading of elicitation paragraphs).

critique of "data collection" can also be applied to machine learning datasets. The term "data collection" is linked to other metaphors used to describe "data", such as the framing of "data as a natural resource" (see also Stark and Hoffmann 2019; Taffel 2021) which can be "exploited" (Benjamin 2021). Especially in the context of scientific research, "data collection" is also linked to the notion of "objective discovery" as invoked in the natural sciences which conceals the "situated" nature of the knowledge and motives of the "collectors" (Benjamin 2021; Haraway 1988). Benjamin (2021) instead suggests we draw on the concepts of "creation", "curation" and "compilation", all of which foreground the constructedness of data and datasets. "Data creation" highlights the abstraction and labour involved in datasets but could also mean that ownership is solely attributed to the people assembling a dataset (e.g., linguists), rather than the people with whom the "data" originates (e.g., speakers) (Benjamin 2021). "Data curation" highlights the task and responsibilities of selection as separate from creation (Benjamin 2021). Finally, "[data] compiling provides an extra level of engaging with what is chosen for both inclusion and (perhaps more importantly) exclusion in the always-already-edited practice of defining the (power) structures of data sets" (Benjamin 2021, p. 18).

With the proliferation of machine learning research across different domains, and the increasing deployment of these technologies in high-stakes, real-world contexts, a growing number of scholars are paying close attention to the role of "data" and, in particular, the (social and technical) histories of datasets in machine learning (e.g., Crawford and Paglen 2021; Paullada et al. 2021; Koch et al. 2021; Raji et al. 2021; Denton et al. 2020). Following Denton et al. (2020)'s call for a "geneaology of datasets", I am interested in interrogating *why* we see data bias in order to understand *how* we could change our practices to avoid them in future dataset compilation.

5.3 Language technology practitioners as language managers: arbitrating data bias and predictive bias in ASR

Section 5.3 was published as:
Nina Markl and Stephen Joseph McNulty (2022). "Language technology practitioners as language managers: arbitrating data bias and predictive bias in ASR". in: Proceedings of the Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 6328–6339. URL: https://aclanthology.org/2022.lrec-1.680*

All language communities, even monolingual ones, show linguistic variation. The co-existence of multiple different ways of communicating the same meaning is a fundamental characteristic of natural language. Speakers can employ different words to refer to the same object (e.g. "film" and "movie"), pronounce the same word differently (e.g. "data" and "data"), address interlocutors differentially depending on the context (e.g. "you", "yous", "y'all" in many varieties of English), and even utilise different sentence structures (e.g. "data is" and "data are").

Despite the fact that this variation is inevitable, people still form judgements about them. Very often, these judgements reflect biases about (groups of) people, not language per se.

These social-linguistic judgements contribute to differential access to, and performance of, language technologies for speakers of the over 7000 language varieties[3] spoken in the world. Most language communities globally do not have access to them at all, and within those that do, performance for speakers of non-standard(ised) and marginal(ised) varieties is worse. For automatic speech recognition (ASR) systems, this "predictive bias", defined by Shah et al. (2020) as a systematic error disparity between different user groups, arises in part from data bias in the speech datasets used to train and test them. In this paper, we use the lens of "language policy" to understand the origins and consequences of this data bias, and to facilitate its mitigation. We contend that, perhaps unknowingly, organisations – and particularly individuals – involved in the design and creation of these datasets, whether crowdsourced or curated, perform the function of "language policy arbiters" (Johnson 2013). In their selection of widely spoken, prestigious (and often commercially-viable) varieties, these individuals effectively marginalise speakers of minority or lesser-used languages or forms of language. This marginalisation may take the form of limiting access to these technologies and exacerbating stigma towards some varieties in their wider application, thereby amplifying systemic discrimination against particular groups and their language(s). Yet, by recognising the need for proactive, diversity-oriented language management – and their role in engendering it – speech and language technologists can work to mitigate such harms, and work towards more equitable and inclusive technologies.

This paper illustrates the role of language policy in exacerbating predictive bias. While we use the example of English language ASR because of the outsize attention English has received in research on speech technology and algorithmic bias, the framework remains applicable in many contexts in which one variety is considered dominant vis-a-vis another.[4]

5.3.1 Predictive bias in ASR

Recent work shows that state-of-the-art commercial English language ASR systems display significant predictive bias for African American English (AAE) and some regional varieties of English. Koenecke et al. (2020) document dramatic racial error disparities for ASR systems sold by Google, Amazon, Microsoft, IBM and Apple, with much larger error rates for Black speakers of AAE than White speakers of Californian English. Overall, recent research suggests that this predictive bias is driven by under-representation of AAE in training data for both acoustic models (Koenecke et al. 2020) and language models (Martin and Tang 2020) used by commercial ASR systems. Koenecke et al. (2020) find error disparities based on pronunciation differences, while Martin and Tang (2020) show that Google Cloud Speech-to-Text handles AAE syntactic

[3] "Language variety" refers to languages (e.g. English), "dialects" (e.g. Scottish English) and accents (e.g. Standard Scottish English). The linguistic features characterising a variety are called "variants".

[4] This focus, in and of itself is, of course, also one of the central ways in which speech and language technologies are "biased" and deeply unequal (Joshi et al. 2020).

features such as "habitual be"[5] poorly. A slightly older set of studies has shown similar error disparities for different regional varieties of English, including Scottish English and Southern U.S. English in products sold by Google and Bing (Tatman 2017; Tatman and Kasten 2017). This apparent data bias is not limited to commercial ASR systems, as Mozilla's open-source DeepSpeech system trained on their crowdsourced CommonVoice corpus also performs significantly worse for AAE than Mainstream US English (Martin and Tang 2020). Other work has focused on the use of ASR as an assistive technology and found that most major systems perform poorly for Deaf and hard of hearing (Glasser 2019), and dysarthric users (De Russis and Corno 2019; Young and Mihailidis 2010).

Harms of predictive bias

While technical research on bias mitigation is important, especially because particular model structures can amplify data bias (Hooker 2021), it is crucial to consider the socio-historical origins of predictive bias and its consequences. Most obviously, speech recognition is a component of voice user interfaces which can be used as assistive technologies to access mobile devices and computers. As (mobile) computing becomes increasingly ubiquitous, predictive bias could severely disadvantage AAE speakers and other speakers of stigmatised and under-represented language varieties in completing everyday tasks such as making phone calls, searching for information on the web and sending emails, and engaging agents in private and public sector contexts. Recently speech recognition has joined other AI technologies in moving into very high-stakes contexts such as hiring and healthcare (Lee 2021)[6]. Companies like HireVue[7], for example, claim to use "voice data" including information about voice quality and lexical choice to pre-screen and rank job applicants (Raghavan et al. 2020). While HireVue has recently passed an independent audit of their algorithmic systems, according to which their training data is balanced by race, gender, region and job title, accent-based bias among first and second language speakers of English has not been studied[8]. HireVue and its competitors also offer customisation of training data for client companies, which makes identification and mitigation of data bias in practice particularly difficult (Raghavan et al. 2020). By disadvantaging marginalised speech communities in accessing technology and resources (up to and including employment), predictive bias can further reify and entrench existing linguistic, and by extension, social hierarchies.

[5] A common feature of AAE not found in Mainstream US English e.g. "I be in my office at 7.30" which is equivalent to MUSE "I am usually in my office at 7.30" (Green 2002).

[6] Amazon even sells an ASR system specifically for medical transcription: https://aws.amazon.com/transcribe/medical/

[7] https://www.hirevue.com/

[8] https://www.hirevue.com/blog/hiring/industry-leadership-new-audit-results-and-decision-on-visual-analysis

5.3.2 Language Policy

As Blodgett et al. (2020) show, discussions of "bias" in the language technology literature often lack grounding in the broader socio-historical context of users or the system, failing to spell out what exactly is meant by "bias," who is harmed by it and how it relates to larger power structures. In this paper, we use the sociolinguistic concept of "language policy" to understand both where data bias comes from and whom it harms.

Language policy relates to the rules, conventions, choices, values, ideas, or discourses which govern the way that we use or think about languages and their speakers (Spolsky 2003; Johnson 2013). These policies can be either explicit or overt, as is the case of language legislation or institutional language policy documents, or can be concealed, covert or de facto – often couched in decisions or actions not specifically related to languages, or in implicit judgements about them or those who speak them (Shohamy 2006). For Spolsky (2003) it is composed of three distinct but interrelated phenomena, which we will explain in turn. *Language practices* refer to conventionalised or patterned language behaviours; *language ideologies* are value-based judgements of specific language varieties and variants, and by extension their speakers and communities; and *language management* refers to attempts to modify language practice and language ideologies.

Language practices

As we have noted, language use is characterised by variation. In an influential formulation, Weinreich et al. (1968) refer to this variation as "orderly heterogeneity". That is, variation in language is patterned and rule-governed. Individuals and speech communities use this variation to construct social identities in interaction (Eckert 2012). These linguistic choices, which are constrained by the social norms transmitted within a community, make up the community's language practices (Spolsky 2003).

Understanding patterns of language variation is crucial to identifying the sources of predictive bias in ASR (and other speech and language technologies) and developing mitigation strategies. For instance, some earlier work on predictive bias in ASR noted apparent differences according to speaker gender (Adda-Decker and Lamel 2005; Benzeghiba et al. 2007; Tatman 2017). Adda-Decker and Lamel (2005) locate the source of better performance for women's speech as compared to men's speech in data bias in training and test datasets. In the English language broadcast news training and test corpora Adda-Decker and Lamel (2005) use, women are more likely to be newscasters and interviewers who adopt a formal speech style, while men are more likely interviewees whose speech is more often unplanned and conversational and thus characterised by repetitions, phonetic reduction, back-channels and filled pauses. In addition to the conversational roles of women and men in these datasets, they also attribute differences to broader gendered patterns of language use, whereby women tend to avoid stigmatised linguistic features more than men (Labov 1990). Overall, the more formal

speech styles associated here with women are easier to process for the ASR system[9]. This gendered pattern in language use is also reflected in Koenecke et al. (2020), who find that commercial ASR systems are more error-prone for men. An analysis of their test set shows that men are, generally speaking, more likely to use higher rates of non-standard forms (Koenecke et al. 2020). This "gender gap" in ASR performance and speech patterns is furthermore substantially larger for Black speakers, highlighting that race and gender as interacting axes of oppression cannot be considered separately, as has long been noted by Black feminist scholars like Crenshaw (1991) and Hill Collins (2000 [1990]). In addition to gender and race, other relevant social factors conditioning language variation are socio-economic class, educational background, linguistic background, disability and ethnicity (Van Herk 2018). Which of these factors are particularly important depends on the specific context. Generally, varieties spoken by powerful groups within a society or societal context (e.g. higher social class groups, White groups, particular geographical areas) become associated with prestige (due to their association with power). Often these prestigious varieties are also "standard varieties", codified in prescriptive (rather than descriptive) grammars and taught in the education system (Van Herk 2018). Poor ASR performance on non-standard varieties, then, is more likely to affect already marginalised speech communities.

Language ideology

From a linguistic perspective, no language variety is inherently "better" or "worse" than any other. However, because language (variation) is always situated within larger social contexts, specific ways of speaking can become indices of particular social identities. Language users create beliefs about language to explain and justify these (arbitrary) associations between speaker and form. As Irvine and Gal (2000, p. 37) put it, these beliefs "locate linguistic phenomena as part of, and evidence for, what [language users] believe to be systematic behavioral, aesthetic, affective and moral contrasts among the social groups indexed".

Language users (all of us) lean on these ideologies when we make judgements about other people (albeit often unconsciously). Like other ideologies, they often seem to reflect "common sense" and resisting them requires conscious effort. They also have real implications for, in particular, marginal(ised) groups. For example, many studies show that second language speakers of English are less likely to be hired (Hosoda and Stone-Romero 2010; Timming 2016) and are frequently rated "less credible" (Lev-Ari and Keysar 2010) than first language speakers. There are further very strong language ideologies around "professional", "educated" and "articulate" speech (Lippi-Green 2012; Baratta 2017). These ideologies are underpinned by broader structural biases within a society such as racism, classism and sexism. It is because of those broader structures that some social groups (e.g. White, upper and middle class, men)

[9]Garnerin et al. (2021) show that when women's speech is under-represented in training sets of read speech, performance is significantly better for men. Notably, adding more women's voices improves performance for women without degrading performance for men.

have more power relative to others (e.g. Black and non-white, lower and working class, women and non-binary people), and as a result, their speech becomes associated with power and prestige. Notably, language ideologies are not applied in the same way in every context. This makes sense if we recall that the supposed attitudes about particular linguistic features aren't about language per se, but about the social identities they are associated with. For example, some voice qualities like creaky voice ("vocal fry") are more stigmatised in young English speaking women than men (Anderson et al. 2014)[10].

Language ideologies feed into speech and language technologies in many different ways. As we explore in this paper, they influence which kind of language we use to train and test language technologies, and as a result, who is most impacted by predictive bias. But speech and language technologies also reinforce language ideologies. Better performance on some varieties emphasises their status. And even in cases where there's no predictive bias, they can reinforce existing ideologies. For example, HireVue (and similar models) may not show predictive bias as such, but since they are trained on interviews with successful job applicants, language ideologies around "professionalism" as expected during a job interview are encoded. At first glance it may seem fairer if these harsh prejudices are part of an algorithmic system since they are at least applied equally (e.g. "vocal fry" = "bad," "long sentences" = "articulate" = "good"). But more privileged people are much more likely to have access to "the right way of speaking" in an interview, for example because they have been taught how to speak during interviews and what kind of language (features) to avoid. According to the HireVue audit report, applicants from "minority" backgrounds are more likely to give very short answers which potentially puts them at a disadvantage as the system does not pose follow up questions.

Beliefs and ideologies about language varieties are inevitable and omnipresent. They simply represent "what people think should be done" with regards to language use within specific societal contexts (Spolsky 2003, p. 14). It is when these sets of preconceived judgments begin to affect language-related decisions that we enter the realm of language management.

Language management

Language management occurs across all spheres and sectors of society, and involves a wide and diverse range of actors (Spolsky 2003; Blommaert et al. 2009; Hornberger and Johnson 2007). What is crucial to remember, is that, even when pertaining to the most mundane, mechanical and technical actions or decisions, language policies are never neutral. By its very nature, language management involves taking a stance on language varieties and variation, by deciding which forms of speech are appealing, acceptable or correct, and which are unattractive, inferior or simply "wrong". Moreover, as Tollefson (1991) and Shohamy (2006) note, language management often serves to create, reify and reproduce unequal power divisions within society: privileging speakers of dominant, prestigious varieties (e.g. native speakers of a stan-

[10] Anderson et al. (2014) conclude that women should avoid creaky voice. We reject this conclusion, and point instead to Chao and Bursten (2021) for a detailed feminist critique of the response to women's creaky voice(s).

dard form of English) and further marginalising people who use stigmatised forms of language (e.g. non-native speakers of [a non-standard] English, or speakers of minority languages). Furthermore, as Wiley (2012) highlights, the absence of an official policy or non-consideration of issues related to equality and diversity in language, often serves only to reinforce the power and hegemony of prestige varieties, and marginalise others: "The lack of recognition of 'nonstandard' varieties of language [...] positions their speakers as merely 'substandard' articulators of English." Inaction, therefore, is action. As noted previously, language managers, planners or policy actors can take many forms. According to Johnson and Johnson (2014), however, certain individuals are endowed with a disproportionate amount of power within specific language policy processes. As a result of their position of influence within a given organisational, institutional or social hierarchy, these "language policy arbiters", through their interpretations and ideological reflexivity (or lack thereof), can influence how language policies are created or implemented (Hornberger and Johnson 2007). Given the possible impacts of their actions, if social inequalities are truly to be redressed, it is essential that these individuals recognise how much power they wield. The design and creation of speech technologies, we believe, constitutes a form of language management with consequences across societal scales, and its designers and operators perform the role of language policy arbiters for their end users, as well as for society more generally.

5.3.3 State-of-the-art: training & testing

An example of this form of language management would be the curation of speech datasets used in the training and testing of ASR systems. It is through this process that decisions about what kind of language to include or exclude in training and test datasets are made. These decisions then shape for which kinds of language, and therefore for which kinds of speakers, these technologies are useful rather than harmful.

Training ASR in industry

ASR systems by corporations like Amazon and Google, or large foundations such as Mozilla, are trained on very large datasets. In the case of commercial ASR these datasets consist (at least in part) of voice commands and dictation snippets which are collected from customers during their interactions with voice user interfaces and transcribed by employees[11].

Mozilla's corpora are made up of voice recordings which are submitted, transcribed and validated by volunteers via an online platform[12]. As explored in 5.3.1, both types of systems exhibit predictive bias towards less prestigious varieties, in particular African American English. In the following section, we explore how corporate language policies influence the apparent data bias giving rise to these error disparities.

[11] With consent of the users, as indicated in the privacy notices of e.g. Apple, Microsoft, Amazon and Google
[12] https://commonvoice.mozilla.org/

Corporate: Proprietary user data Corporations like Amazon, Google and Microsoft do not provide detailed model documentation for the ASR systems they sell to third parties (e.g. Amazon Transcribe, Google Cloud Speech-to-Text) or the ones embedded in their own products such as voice user interfaces (e.g. Siri, Alexa, Cortana) and video platforms (e.g. YouTube captions) but their privacy notices and academic publications suggest that large proprietary datasets which include data collected from users are involved. For example, Chiu et al. (2018) (Google) present a system which is trained on "representative voice search data" from their user base. Similarly, Facebook AI trained a multilingual ASR system on "publicly shared user videos" in 51 languages (Pratap et al. 2020). A fundamental problem with training on user data is that even if this data is "representative" of the user base, the user base is not necessarily representative of the population at large. According to a 2021 Pew Research Center survey, 85% of residents of the United States own a smartphone[13]. However, there are still quite big gaps between different age and social class groups. There are further even larger gaps in home broadband access depending on income in particular. As has been raised in the context of large language models, while digital spaces are in theory "open to everyone", participation in online communities is not equally accessible or attractive to everyone (Bender et al. 2021). Any dataset based on online communication, then, risks mis- or under-representing marginalised (speech) communities who may not be able or willing to participate (Bender et al. 2021). Indeed, the findings by Koenecke et al. (2020), Martin and Tang (2020) and Tatman and Kasten (2017) suggest that, in the context of US English, Black talkers in particular remain under-represented. To avoid predictive bias, data from different groups would have to be balanced rather than merely representative of the (skewed) population distribution (Suresh and Guttag 2021; Barocas et al. 2019)[14].

Big (speech and language) technology companies do not tend to have publicly available officially declared language policies. However, as alluded to above, just because there is no official document outlining a language policy, it does not mean that there is no policy in place. Some language policy scholars such as Schiffman (1996) and Shohamy (2006) distinguish between *de jure* and *de facto* language policies. Even in the absence of the former, de facto policies can still arise, often on the basis of what people in a particular context find to be sensible, convenient or common sense. In this context, beliefs about language (i.e. language ideologies) can be particularly influential (Shohamy 2006). A key aspect of language management is the selection of a particular language variety to be used in a particular context. In the context of speech and language technologies, this selection process includes the choice of a particular variety to train and test a system on, and consequently, develop for. For example, Benjamin (2019b) quotes a former Apple speech technology researcher working on Apple's voice assistant Siri asking their supervisor in 2015 why AAE was not a priority while support for other varieties of English such as Singaporean English was being developed. The response:

[13] https://www.pewresearch.org/internet/fact-sheet/mobile/

[14] As Hooker (2021), notes, the fact that most "real-world" data have skewed distribution is why it's important to focus on mitigating bias through model choice too.

"Well, Apple products are for the premium market." (Benjamin 2019b, p. 15). This statement expresses a language ideology held by (at least a part of) the corporation: AAE is not spoken by "the premium market" and AAE speakers do not (or cannot afford to) buy "premium products". Assuming that Apple's main goal is to attract (and keep) the "premium market" as is implicit in the quote above, only developing "premium" linguistic varieties is a good investment. This ideology is the company's de facto language policy: AAE is not supported by the company. By applying this economic reasoning to language varieties (and their speakers), Apple also reinforces existing "linguistic markets" (Bourdieu 1977). It's perhaps not surprising that Koenecke et al. (2020) found the racial gap in predictive errors to be largest, and overall performance on AAE to be worst for Siri (as compared to other systems tested). More broadly, selecting language varieties based on their perceived value on the (linguistic) market means that varieties spoken by marginalised or small communities are less likely to be supported. Differences in language policy between corporations are also reflected in the different sets of languages they select. Google has the largest range of language varieties, including national varieties for languages like Arabic, Urdu, English and Spanish[15]. While smaller national and regional languages spoken in Europe (like Macedonian and Basque) are supported, the same can only be said for languages with larger speaker populations outwith Europe like Uzbek, Zulu, Amharic, and Gujarati, highlighting a general global skew in speech technology availability. Similarly, Apple's Siri is offered in US Spanish and two post-colonial English varieties (India & Singapore) but does not support any languages indigenous to Africa, the Americas, Oceania or the Indian subcontinent. These choices do not just impact current and future customers of these technology corporations: Apple, Google and Microsoft sell their speech recognition services to third parties, and their choices (of data and algorithms) likely impact the way smaller companies act.

Open-source: Crowdsourcing The most obvious alternative to this purely market-driven model of technology development already in use today are open-source and crowdsourced technologies, such as Mozilla's DeepSpeech ASR system and CommonVoice collection of crowdsourced speech datasets[16]. The latter currently covers 76 languages. Volunteers contribute by reading out sentences which are recorded via an interactive interface and validated by other volunteers. All contributors can optionally provide information about their gender, age and accent. CommonVoice does not appear to have a top-down policy for selecting language varieties. Volunteers can request the initiation of a corpus for a new language. The accent labels available for volunteers seem to be selected by community members[17], with Spanish varieties defined in geographic terms while German varieties are defined as national varieties (eliding variation within nation states). Similarly, the English corpus contains "Scottish English" and "England English" alongside a very broad "US English" making comparisons of sampling bias very diffi-

[15] https://cloud.google.com/speech-to-text/docs/languages
[16] https://commonvoice.mozilla.org/en
[17] https://discourse.mozilla.org/t/spanish-accents/35638

cult. Mozilla is currently in the process of replacing this apparent "null policy" with a declared "languages and accent strategy' which has at least in part been crowdsourced in discussion with community members on a public Mozilla discussion forum (and seems to have also been informed by discussion with linguists) (Mozilla Common Voice: Discourse 2019; Mozilla Common Voice: Community Playbook n.d.; Mozilla Common Voice 2022).

For smaller or marginalised speech communities and/or those in the Global South in particular, this participatory framework of crowdsourcing both language and language policy appears a better strategy for speech and language technology development than relying on large for-profit corporations. Speakers can engage in "conscious data contribution" (Vincent et al. 2021), and (within limits) directly shape what kind of language(s) DeepSpeech will support. For some varieties, like Kabyle (a Berber language with 7 million speakers) or Kinyarwanda (a Niger-Congo language with 12 million speakers) this approach also appears successful as they have sizeable validated corpora. However, some varieties (regardless of speaker numbers) have only very small CommonVoice corpora, or corpora which are very unbalanced across varieties (most notably Arabic, which has a large number of distinct dialects spoken in different regions but is currently only represented by Standard Arabic), as well as age and gender groups. The majority of the contributors to the English CommonVoice corpus, for example, did not provide any information about their accent and only 15% identified themselves as female. Notably, in the context of existing research on bias in ASR, CommonVoice does not collect information on race or ethnicity, and "African American English" is not one of the possible "native accents". This lack of documentation makes evaluation of data bias difficult. Overall, while crowdsourcing can alleviate some of the data bias issues we see in commercial ASR, especially when done with an explicit focus on accent diversity, many representation issues persist. Recall Martin and Tang (2020) show that it performs worse for AAE speakers. As has also been discussed in the context of Wikipedia, who contributes to crowdsourced projects depends on many factors such as availability of free time, technical skills, access to digital technology and the culture of the crowdsourced project (Hargittai and Shaw 2015; Tripodi 2021). Crowdsourcing also places the onus to create data on potentially already marginalised speech communities who might furthermore disagree about how (and if) their language should be represented in these systems (e.g. which accents or writing systems) and how they would like any finished system to be used.

Testing

In speech and language technologies (and machine learning more broadly), benchmark datasets are used to evaluate the performance of new algorithmic systems (Schlangen 2021). While this focus on benchmarks has recently become the subject of critique (Bowman and Dahl 2021; Denton et al. 2020; Koch et al. 2021; Raji et al. 2021), they are still central to the way the field defines "progress". In the following section we explore how language ideologies shape the well-established academic benchmark corpora TIMIT (English) (Garofolo et al. 1993), Switchboard

(Garofolo et al. 1993) and CallHome (American English) (Canavan et al. 1997).

Data bias in academic corpora TIMIT (English) (Garofolo et al. 1993), Switchboard (Garofolo et al. 1993), CallHome (American English) (Canavan et al. 1997) are well-established licensed speech corpora which were collected in the late 1980s and early 1990s and are held by the Linguistic Data Consortium. TIMIT (English) was collected by MIT, SRI International and Texas Instruments to be used in speech technology development and acoustic-phonetic research. It features recordings of 630 speakers of 8 "major dialects of American English", each reading 10 phonetically rich sentences which have been phonetically transcribed and aligned. Switchboard contains 2,400 two-sided telephone conversations between 543 US American strangers on one of 70 pre-selected topics collected by Texas Instruments. CallHome features 120 unscripted 30-minute telephone conversations between friends or family members (all "native speakers of English" who grew up in the United States) and was collected by the Linguistic Data Consortium.

While all three corpora were carefully designed to capture some regional dialectal variation in US English, they are not balanced across gender groups. Further, most speakers appear to be White, though race is only recorded in the documentation of TIMIT. In the case of TIMIT this is perhaps due to convenience sampling of participants: most of the speakers were employees of Texas Instruments in Dallas which collected the corpus[18]. Demographic imbalances are potentially more critical for Switchboard, where only the topic of conversation, not the speech style was constrained and for CallHome where speech styles could also vary widely, and women are over-represented. As noted in 5.3.2, this gender imbalance could be indicative of a speech style imbalance. A recent analysis by (Martin 2021) further confirms that Switchboard and TIMIT under-represent AAE.

Evaluation bias & biased benchmarks Systems trained on biased datasets can exhibit predictive bias. But training is not the only context in which harms and biases can be introduced in the development and implementation of a machine learning system. Suresh and Guttag (2021) use the term "evaluation bias" to describe the bias which occurs when there's a mismatch between the benchmark data used for a particular task and the intended use population. As outlined above, some established benchmarks are unrepresentative of the potential user base of English language ASR, which include second language speakers, speakers of "non-standard" regional dialects and ethnolects and speakers who frequently code-switch between several varieties. These benchmarks are also in some ways misaligned to current ASR applications (Szymański et al. 2020). Today, ASR is widely used to transcribe conversational speech which is notoriously challenging for systems designed to recognise simple commands for virtual agents in human-computer directed speech.

Particular evaluation strategies can exacerbate this kind of bias (Suresh and Guttag 2021).

[18] https://nvlpubs.nist.gov/nistpubs/Legacy/IR/nistir4930.pdf

Computing an aggregate word error rate across these homogeneous and/or unrepresentative test sets hides predictive bias. If Koenecke et al. (2020), for example, had computed word error rate over all speakers, the overall higher than state-of-the-art word error rate would have perhaps been attributed to the conversational nature of the recordings, rather than significant difference by speaker race. As discussed in 5.3.1, and as the CORAAL (Kendall and Farrington 2021) recordings used by Koenecke et al. (2020) illustrate, race and gender interact in language variation. This is reflective of the concept of intersectionality originating in Black feminist thought (Crenshaw 1991; Cooper 2016), which recognises that interacting social categories (and axes of oppression) such as race and gender cannot be considered separately. Intersectional evaluation, then, is mindful of these interactions and can capture the differences in life experiences and linguistic behaviours between, for example, Black women and White women, rather than considering either only race or only gender. Within machine learning, this type of approach to evaluation has also been successfully applied in the context of facial analysis (Buolamwini and Gebru 2018).

It is difficult to ascertain how much language ideologies influenced the collection of these licensed corpora in the 1980s and 1990s. At the time, they were created for a relatively narrow purpose (to research speech technologies, particularly in an academic context). It is unlikely that the researchers designing the data collection expected these resources to still be used to benchmark state-of-the-art speech recognition systems thirty years later. While incorporating some regional dialectal variation was clearly a priority, ethnic diversity or the inclusion of African American English wasn't.

The decision to use these datasets as benchmarks in the 2020s despite these limitations is, however, a choice that constitutes language policy. Just as particular language varieties or datasets are "selected" in training, they are also selected in testing. And just as training is shaped by language policy, so is testing. At first glance, Switchboard, TIMIT and CallHome fulfil the primary function of a benchmark: to allow comparison with other systems. Following Schlangen (2021)'s definition of a benchmark, they should, however, also "exemplify" the overall task of interest. A mismatch between benchmark and real-world application is therefore undesirable. More importantly, a mismatch is unexpected, as there is an implied relationship between benchmark and real-life application. The selection of an unrepresentative benchmark is shaped by beliefs about what kind of speech (and by extension, what kind of speakers) speech recognition should (be expected to) work for. Due to the evaluation bias this application of benchmarks produces, these ideologies are then further reinforced. Failure to perform accurately on underrepresented speech not only goes undetected, but, perhaps more troublingly, is not penalised. Of course, the benchmark doesn't have to be representative of all application contexts if we choose to only use it to compare new systems to older systems. But nevertheless, the picture benchmarks provide are always partial and potentially very misleading, especially since they are almost never described in detail in the papers that use them to evaluate (Szymański et al. 2020).

5.3.4 Towards better practices

As we tried to highlight in this paper, both the curation and the use of particular speech datasets constitutes a form of language management, itself influenced by beliefs and ideologies surrounding language variation. Given the potentially far-reaching consequences of their decisions, practitioners working with speech datasets could be considered "language policy arbiters": individuals who "[wield] a disproportionate amount of power in how a policy gets created, interpreted, appropriated, or instantiated relative to other individuals in the same context" (Johnson 2013, p. 100). Who gets to select which data is used in training and testing obviously depends on the broader institutional context. In a commercial context, language policy appears to be primarily driven by (linguistic) markets, and may be decided by business strategists, rather than technologists. But even in commercial contexts, researchers can reflect critically on those policies and, as work in language policy highlights, often have some leeway in the way they implement them (Hornberger and Johnson 2007). This is not to say that we should only rely on individuals' sense of justice in the face of structural oppression, but instead to note that agentive and reflective work by technologists has potential to spark or enable broader discussion and change. In this final section, we also echo other critical work in machine learning (Paullada et al. 2021; Hutchinson et al. 2021) and argue that understanding (speech) datasets as increasingly important infrastructure is useful. It allows us to reframe the task of speech technology development from one primarily done by corporations for markets to one done by a wider range of actors for speech communities.

Speech technology design as civic design

A central obstacle to minimising predictive bias in commercial ASR systems appears to be a lack of incentive for corporations to do so. Smaller and more marginalised speech communities are unlikely to be seen as desirable markets by big technology companies, and curating very large datasets could be challenging and relatively expensive. Where proprietary datasets derived from user-data do exist, evaluating data bias is potentially difficult. It's unlikely that a technology company would (or should) be able to document or reliably infer important demographic information (such as accent, age, gender, race) about the speakers whose data is used to create a balanced dataset (Andrus et al. 2021). Curated licensed corpora could be combined to train complex systems (as was done by Microsoft in: Xiong et al. (2017)) but since current well-established corpora only represent a small section of all English speakers, new corpora would have to be collected for this purpose. Speech technology companies could, of course, do this themselves, for example by offering payment to users (or crowd-workers) who complete a survey about their demographic background and provide speech recordings of read or naturalistic speech (see Facebook AI's (Hazirbas et al. 2021) for one of the first attempts at this method). Ultimately, however, this approach would not solve the fundamental issues arising from designing for markets.

Alternatively, we could reframe speech technology as a kind of public infrastructure and its design as civic design. Mugar and Gordon (2020, p. 25) define "civic design" as an approach to design that "creates the conditions for a plurality of voices and interests to be represented, accounted for, and involved in shaping the outputs and effects of public life." Civic design is design for and with publics, rather than markets (Mugar and Gordon 2020, p. 53). The notion of a "public" as a collective of people which emerges through discursive circulation of shared interests with the purpose of influencing decision-making (Mugar and Gordon 2020, p. 66), has also been taken up in the analysis of language users (Muehlmann 2014; Gal and Woolard 1995).

Some linguistic publics intersect with the public of a nation (state), such as the Icelandic-speaking public or the Estonian-speaking public. In those cases, a (national) government (a traditional actor in language policy) shares the public's interest in the development of speech technologies which it understands as a type of infrastructure. It can steer (and pay for) corpus development. The governments of Iceland and Estonia have both overseen design and development of open-source speech and language technology resources (corpora and models) by private and public partners (Nikulásdóttir et al. 2020). Similarly, the Welsh government has prioritised speech and language technology development and is working with universities and private sector businesses to deliver it (Welsh Language Division 2020; Welsh Language Division 2018). Organisations like CLARIN[19] maintaining access to resources created in different contexts are also important here.

A civic design approach can also be useful for other kinds of diverse linguistic publics which do not necessarily form a "viable market". As digital devices are becoming crucial gateways to accessing public services, jobs, and media and predictive bias could exclude many people from using them. Civic design as something that is done by a public for a public also has the potential to resolve some of the current issues with crowd-sourcing speech datasets. By carefully (and meaningfully) engaging speakers, not just as anonymous data sources, but as co-designers who can shape the technology development process, following, for example, principles of design justice (Costanza-Chock 2020), technology developers (private or public) would likely be able to create more representative and ultimately more useful technologies, and move away from colonial frames inherent in many drives to "spread language technologies" (Bird 2020). With the proliferation of open-source speech technology toolkits and cheap(er) cloud computing, some publics may be able to build or modify these technologies without much or any support from governments of corporations (e.g. Masakhane[20], (Khandelwal et al. 2020)). Mugar and Gordon (2020) also emphasise that, in their vision, the aim of civic design by and for publics is care rather than innovation, and space for meaningful interaction between people. These values run counter to the ethos currently driving commercial technology development, but they are excellent principles in the context of technology designed fundamentally to facilitate communication.

[19] https://www.clarin.eu/
[20] https://www.masakhane.io/home

Speech datasets as infrastructure

Whether speech technologies are approached from a perspective of civic design or not, speech datasets are, like all datasets in machine learning, infrastructure. As Hutchinson et al. (2021) point out, curation and maintenance of this infrastructure is undervalued in the machine learning community and as a result, datasets are often poorly documented and precariously stored. Fundamentally, careful curation (following a civic design model or any other model) and good documentation of speech corpora is tractable, due to the comparatively smaller size of datasets compared to, for example, large language models (Bender et al. 2021). Documentation is essential in mitigating (or even simply anticipating) predictive bias. Speech datasets (like other language datasets) need not be static but rather, like physical infrastructure, require maintenance and updating. As language (both in use and form) continuously changes, static datasets will deprecate over time and an approach in which practitioners can add or remove data from training sets in deployment may be more useful (assuming any changes are documented).

5.3.5 Conclusion

Predictive bias in speech recognition technologies is an increasingly important problem as speech recognition systems get embedded into complex algorithmic systems, with harms disproportionately falling on already marginalised speech communities. We believe that language policy is a lens that can empower technologists to mitigate data bias and recognise potential harms of biased technologies. We want to encourage practitioners to adopt this reflexive approach to better understand how language ideologies affect speech technologies and their users, and to use this understanding to build better speech technologies.

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

Section 5.4 was published as:
Nina Markl (2022c). "Mind the data gap(s): Investigating power in speech and language datasets". In: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion. *Dublin, Ireland: Association for Computational Linguistics, pp. 1–12. URL:* `https://aclanthology.org/2022.ltedi-1.1`

5.4.1 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 2019b; 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.[21] 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,

[21] For an equivalent documentation framework for machine learning models, see Mitchell et al. (2019).

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[22]. 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 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 2021b). 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 2017; 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, p. 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; Graff et al. 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

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

development, while Switchboard-2 is a collection of telephone conversations collected by the Linguistic Data Consortium, an academic institution, to develop speaker recognition systems.

5.4.2 Background

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], p. 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).[23] 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

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

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 2021b).

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, p. 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, pp. 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, p. 23)[24]. 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.

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.

5.4.3 Power in language datasets

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

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

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 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, p. 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.

Who and what is missing?

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

lation) and what's not specified in the documentation can be revealing. These gaps provide insights into 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)[25].

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?

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, p. 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; Martin 2021) 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 Talat et al. (2021) point out, despite the "disembodied" framing 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

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

(2020, p. 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.

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, p. 18)

Once we have identified who and what is excluded from a dataset and what the potential or actual harms 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.

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, p. 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. 2022a). 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).

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.

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, p. 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 2021b). 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.

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, p. 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.[26] 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 2023). Second language speakers and multilingual speakers are also routinely excluded.[27]

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.

5.4.4 Examples

Below I include two examples of the kind of analysis I propose above. Each example is framed in a question-and-answer style to encourage direct engagement with the key questions I have raised.

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[28] and (where indicated) Ardila et al. (2020), which introduced the corpus.

Who and what is missing? **Q:** Whose voices & language varieties are missing?

[26] In addition to volunteer-based crowdsourcing platforms, micro-work platforms also have these limitations, especially as lack of demographic information about workers can make it difficult to annotate datasets appropriately and/or ensure data quality.

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

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

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 5.1 for full breakdown).

Accent

23% United States English, **8%** England English,
7% India and South Asia (India, Pakistan, Sri Lanka),
3% Canadian English, **3%** Australian English,
2% Scottish English, **1%** New Zealand English,
1% Southern African (South Africa, Zimbabwe, Namibia),
1% Irish English

Age

24% 19 - 29, **13%** 30 - 39, **10%** 40 - 49, **6%** < 19,
4% 60 - 69, **4%** 50 - 59, **1%** 70 - 79

45% Male, **15%** Female

Figure 5.1: Screenshot of Common Voice English release 7.0 documentation (Accessed 17/02/2022).

There are important gaps in documentation: 51% of recordings are not assigned an accent label. Although Mozilla allows users to choose the label "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[29], 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[30]. 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

[29] https://creativecommons.org/publicdomain/zero/1.0/
[30] https://github.com/common-voice/common-voice/tree/main/server/data/en

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.[31] Recordings are validated by other volunteers via an interface[32]: 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[33] 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 under-represented 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.

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 datasets such as second language speakers. However, like other crowdsourced projects, contributors are most likely to be young men[34], 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

[31] Since 2022 speakers can self-describe their accent (Mozilla Common Voice 2022)

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

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

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

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 2021a; Mozilla 2021b).

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[35] 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 under-representing 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.

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

Switchboard

Subsets of Switchboard-2 are well-established benchmarks for conversational ASR (e.g., Hannun et al. 2014; Tüske et al. 2020)[36]. All information here is drawn from the (more detailed) documentation of Switchboard-2 (Graff et al. 1998; Graff et al. 1999).

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 (under-representing 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.
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

[36]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

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.

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.

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.

5.5 Discussion: Language (technology) as a resource and infrastructure

A deep critical engagement with what linguistic data is, and what it means to compile and share it, is (increasingly) relevant to speech technology development. As the users and application context of ASR become more diverse, datasets need to reflect these changes and, in many cases, become more "naturalistic" especially where the aim is to transcribe human-to-human interactions. Some of the "new domains" ASR is embedded in have the potential to greatly benefit speech communities previously ignored by technology development by providing accessibility tools and new ways of engaging with technology and one another. At the same time, this expansion could be particularly risky for already marginalised speech communities, as ASR is entering high-stakes contexts like education, healthcare and hiring. Especially in the context of minoritised speech communities, these shifts raise important questions not just about data bias and predictive bias, but also about which data is used and who benefits from it.

5.5.1 Data gaps and language policy arbiters

The people involved in creating and using language resources play a meaningful role in mitigating predictive bias and intervening (consciously or not) in the broader sociolinguistic context by promoting some varieties vis-a-vis others. In Section 5.3, we use the term "language policy arbiter", as introduced by Johnson and Johnson (2014), to argue that language technologies are inherently governed by and reflective of language policies, and that individuals in language technology development have significant power in how these policies are "created, interpreted, appropriated or instantiated" (2014, p. 100). A key insight from studies of language policies in other domains like education or the family is that policies need not be declared in a document (Spolsky 2003). In most contexts, the "rules" or expectations regarding which (type of) language should be used are implicit rather than explicit and reflect a wider "default". Similarly, decisions in language technology development might not always be following a stated "language policy" but rather be reflective of wider attitudes or goals. For example, a technology company might be looking to cater to a particular territory and consequently prioritise an official language or national variety, as documented in Chapter 6. In the context of academic research, recruiting students as data subjects and annotators is a similarly "common-sense" but not neutral decision which might introduce biases. Even in crowd-sourced projects, existing structural biases in terms of participation may lead to the exclusion or under-representation of some (linguistic) groups.

Perhaps less obviously, choices about how varieties and speakers are categorised and described are also extremely important. Gaps in documentation make it more difficult to anticipate biases and interpret and replicate research findings. The absence of marginalised identities (among data subjects and documentation) also contributes to their wider erasure (e.g., in the case of non-binary or trans data subjects). On the other hand, the use of particular

categories also risks reifying them. As a result, reflexivity during the data compilation process is crucial. The "data gaps" framework presented in Section 5.4 can be used to identify these gaps.

5.5.2 Language "resources" and power

However, creating "more diverse" datasets is neither simple nor self-evidently mitigating harms. The harms and benefits of using data of "under-represented" groups depend entirely on what they are used for. But even more fundamentally, "diversifying" a dataset, requires us to make (political and ethical) decisions. While larger, crowd-sourced or web-scraped corpora can be more "inclusive" or "diverse", they are also more difficult to document, as has been explored extensively in the context of very large text corpora (Bender et al. 2021; Luccioni and Viviano 2021). Where language data is drawn from the web or sociolinguistic corpora, it also raises interesting legal and ethical challenges, in particular for indigenous and minoritised varieties.

These issues matter as the most recent "state-of-the-art" systems are shifting away from using smaller (public) corpora to train single-language systems, to using large web-scraped corpora to train multilingual systems. For example, Zhang et al. (2023) (Google) present their "Universal Speech Model" which is trained on over 12 million hours of speech data in 300 languages, 28 billion unpaired sentences in 1140 languages and "fine-tuned" on much smaller transcribed datasets. This approach is feasible because, as Zhang et al. (2023) note "while transcribed speech may be scarce in many languages, untranscribed speech and text data are practically unlimited". In addition to using (transcribed and untranscribed) speech data from public datasets (including CommonVoice and Librispeech), they draw (transcribed and untranscribed) speech from YouTube (Zhang et al. 2023). Radford et al. (2022) (OpenAI) present their "Whisper" model which is similarly trained on speech data from the web (though the paper does not provide more details about the exact sources)[37]. Many of these approaches to multilingual ASR are targeted at supporting "low-resource" or "under-resourced" varieties, where the availability of resources is reflective of the interests of those funding data compilation (e.g., public research bodies) and, more recently the ability of speakers to access and contribute to web-based platforms like Wikipedia (which is a popular source for text data) and YouTube (Joshi et al. 2020).

Where "low-resource" varieties are considered endangered, language technologies are also often positioned as "saviours". For instance, in 2019 UNESCO organised in partnership with the European Language Resources Association (ELRA) the "Language Technologies for All" conference where the demise of languages not supported by language technologies, was framed as inevitable: "Languages that miss the opportunity to adopt Language Technologies will be less and less used, while languages that benefit from cross-lingual technologies such as Machine Translation will be more and more used" (ELRA, 2019, cited in Bird 2020). While

[37] Unlike Google USM Zhang et al. (2023), Whisper is trained on labelled data ("weak supervision") (Radford et al. 2022).

this impulse is often well-intentioned, it reproduces the kind of "tech-solutionism" or "tech-chauvinism" (Broussard 2019; Greene 2021). Bird argues that this "language technologies for all" approach adopted by both large technology companies and (non-)governmental institutions is further "misguided" because it fails to engage with "the ecology of the world's languages" (2022, p. 7817). The impulse to apply the same standard to all languages, regardless of their historical, cultural and sociolinguistic context, and understand them all in the same way is an extension of the colonial approach to linguistic research and documentation (Deumert and Storch 2018; Heller and McElhinny 2017; Kuhn et al. 2020). While language technologies can be useful for communities speaking "low-resource" or "under-resourced" language varieties, it is important to interrogate who benefits from both the language data and the technologies built on top and it is absolutely crucial to involve language communities in the development process.

Data compilation can form part of a larger project Birhane (2020) terms "algorithmic colonisation". Both academic institutions and large technology corporations located and operating from the Global North seek in this way to extract (language) resources to develop tools, services and research which, ultimately, benefit them at least as much as they benefit the communities they're supposedly serving, both in terms of financial and cultural capital. As Hoffmann (2021b) highlights, discourses of "inclusive" and "ethical" development can be used by technology corporations (and academic institutions) to position themselves as responsible and "doing good" (see also Green 2019).[38] But, as Fuller Medina argues: "language data is patrimony" (2022, p. 2). Fuller Medina is talking about one specific sociolinguistic corpus which contains "disappearing cultural heritage" (the "Older Recordings of Belizean varieties of Spanish"), but since linguistic corpora often feature folklore or personal recollections of a particular time and place, her point is relevant to many datasets of "naturalistic" language use. To honour this patrimony (and the language communities) she calls for "repatriation [of sociolinguistic data]" (Fuller Medina 2022, p. 19). This framing raises important questions regarding the "ownership" of not just linguistic data but language varieties more broadly, which are particularly acute in language technology development.

One interesting case study here is the response by Māori speakers to efforts to create proprietary or open-source Māori ASR systems (Coffey 2021; Mahelona et al. 2023). Having compiled a transcribed speech dataset with Māori speakers, the Māori media company Te Hiku resisted requests to sell or license it to non-Māori developers (Coffey 2021). Instead they trained their own system (building on open-source architectures) to transcribe their own radio archive for the Māori community (Coffey 2021), a project they have since expanded into the Papa Reo project[39]. As one Te Hiko employee put it: "They suppressed our languages and physically beat it out of our grandparents. [...] And now they want to sell our language back to us as a service" (Coffey 2021). Importantly, these questions of data sovereignty and who

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

[39] `https://papareo.nz/`

should *own* language data are not limited to explicitly for-profit contexts. The recently released open-source multilingual ASR model Whisper (Open AI) (Radford et al. 2022) was trained on over a thousand hours of Māori speech data. As Mahelona et al. (2023) (Papa Reo) note it is not clear where exactly this data was drawn from (as Radford et al. (2022) provide no detailed description) but like Google USM Zhang et al. (2023), Whisper is trained on data from the web. While it is therefore likely not drawing on the *same* datasets Te Hiko compiled and tried to safeguard, it does represent language technology development without (meaningful) engagement or consent of the Māori community. As Mahelona et al. (2023) argue, indigenous language technology development should be led by indigenous language communities in ways which ensure that they retain control over both the technologies and the datasets they are trained on. Similar arguments are made by organisers of participatory projects like Masakhane NLP (Nekoto et al. 2020) who describe themselves as a "grassroots NLP community for Africa, by Africa" and have been working on a range of NLP tasks in a large number of "low-resource" African languages.

These discussions of course tie into broader debates on ethics of data sharing, especially in the Global South. Abebe et al. (2021) identify the same kind of "deficit narratives" we see applied to "low-resource" languages, applied to African societies more broadly. Folding African researchers, research institutions and governments into a global culture of (more or less open) "data sharing", is framed as a necessary aspect of "development", but as Abebe et al. (2021) highlight, "equitable data sharing" is challenging. It requires a nuanced understanding of the "data setting" (i.e., the context) (Loukissas 2019), local norms and interests and infrastructures which enable access for data subjects (Abebe et al. 2021).

5.5.3 Values and goals in ASR development

Questions of power in language technology development and use are particularly contentious in the context of indigenous languages and other "low-resource" languages, but the wider critique of the role of technology corporations raised by Mahelona et al. (2023) applies generally. As they highlight, the kind of models and datasets used by Google or Open AI are difficult to recreate, store and use without access to the right kind of (considerable) computing power, storage and expertise even if they are "open-source" (Mahelona et al. 2023). (Big) tech[40] companies play an increasingly important role in machine learning development, including in language technologies (as discussed in the context of Amazon and Google in Chapter 6). In an analysis of top-cited papers published at the two largest machine learning conferences (ICML and NeurIPS), Birhane et al. (2022a) show that 55% of the most highly cited papers (from 2018/19) have corporate affiliated authors. Whittaker describes the "steep cost of [industry] capture" of not just *what* we research but also *how* we research it: "[Big Tech] control[s] the tooling, development environments, languages, and software that define the AI research process – they

[40] In this context, Google, Microsoft, Facebook/Meta, Amazon and Nvidia are often referred to as "big tech" (Birhane et al. 2022a; Abdalla and Abdalla 2021).

make the water in which AI research swims" (2021, p. 53). As she points out, this dominance is similar to, and a continuation of, the role the US military used to play in scientific research – as highlighted Section 5.4, much speech technology research has been funded by (D)ARPA[41]. Coupled with the fact that user rights and ethical principles are not commonly cited values or concerns in machine learning research (Birhane et al. 2022a), this trend highlights the need for critical engagement with the wider context in which we develop speech technologies. This increasing influence of large technology companies is also reflected in research focussed on algorithmic bias and fairness (Abdalla and Abdalla 2021; Young et al. 2022).

Resource-intensive development aimed at (short-term) profits, and values like "generalisability", "novelty" and "efficiency", runs counter to the ethos of community-oriented development modelled by Papa Reo and Masakhane. In Section 5.3 we explore the possibility of framing speech technologies and the datasets they are built on as "public infrastructure". This is inspired in part by Denton et al. (2020) who lay out several ways in which machine learning datasets "act as infrastructure". They are a necessary foundation for the systems trained on them, have "an authoritative status" when used as benchmarks and are "likely to be treated as neutral or scientific objects" (Denton et al. 2020, p. 2). This framing makes it obvious why ensuring that these datasets are appropriate and well-maintained is essential. Taking this a step further, we suggest that we could consider speech technologies as a type of infrastructure. According to Larkin, "what distinguishes infrastructures from technologies is that they are objects that create the grounds on which other objects operate, and when they do so they operate as systems" (2013, p. 329). This framing of ASR as infrastructure is perhaps also useful to understand the complicated ways in which it is reliant on existing infrastructures (e.g., computer hardware and software, datasets), and enabling (or necessitating) other infrastructures (e.g., voice user interfaces, regulation). Importantly, we make the suggestion in Section 5.3 to consider the design of these speech technology infrastructures a type of *civic* design. Drawing on Mugar and Gordon (2020)'s work on *civic design* and Costanza-Chock (2020)'s work on *Design Justice*, we can imagine an alternative model of ASR design which is done by and for speech communities. In fact, speech technology development done by communities and collectives like Papa Reo and Masakhane, or under direct guidance of (national) governments like we have seen in Wales, Iceland and Canada (Kuhn et al. 2020) seems to already follow this alternative model.

5.6 Conclusion

In this chapter, I have explored how the datasets used to train and test ASR systems are shaped by technical and social structures. I have suggested that interrogating who and what is *missing* from speech and language data set allows us not only to anticipate predictive bias and

[41]Its "singular and enduring mission [is] to make pivotal investments in breakthrough technologies for national security" (DARPA 2023). Like other language technologies (potential) military applications appear to have been central in the development of ASR (Paullada 2020; Heller and McElhinny 2017)

potential harms, but also reveals those (to many oppressive) social structures. Here I chose the framework of data feminism to make these dynamics visible, interrogate why we see gaps in data and documentation and imagine alternative approaches which avoid such gaps.

Recalling the fact that "language data" is socially constructed, we see the role of individuals and the larger institutions they are embedded in compiling the datasets which form a core infrastructure for ASR. Crucially, these individuals have agency in the choices they (have to) make in the data compilation process. I argued that these choices can be understood as a type of language management as their outcomes affect both how people use language (at least when interacting with ASR systems) and what they think about language(s). I also suggested that understanding individuals involved in the development of "language resources" as "language policy arbiters" might be a productive way to think through the power they have in this process. A crucial insight from the language policy literature here is that even the absence of an explicit language policy represents an ideologically informed policy.

Turning to the wider institutional settings in which these datasets are compiled, I highlighted the role of linguistic markets in commercial ASR development. Within this model, language communities who are not perceived as valuable markets by developers are likely always "left behind". Alternative approaches such as crowd-sourced corpora and academic corpora also exhibit significant data biases which reflect deeply rooted inequities in access to the relevant infrastructures (e.g., internet). Moreover, regardless of how it is compiled, and whether it is intended to be proprietary, closed-access (e.g., licensed) or open-access, language data should be understood as a valuable cultural artefact and practice, not (just) as a "resource". In addition to asking questions around data bias, we also need to ask who should and should not benefit from those "resources" and how the needs and desires of language communities can be centred in this process.

In the following chapter, I build on this work by exploring how "language" and "language variation" are understood by academic researchers and constructed in promotional materials for commercial ASR. Exploring these discourses provides an insight into the language ideologies which shape technology development – and in turn how different language communities and language varieties are "supported" and framed in and through technology development.

Chapter 6

Language ideologies and language technology planning

6.1 Introduction

As discussed in Section 5.3, the decisions involved in developing language technologies can be understood as a form of language management. In addition to reflecting particular language policies, they have the potential to affect both language ideologies and language behaviours. In this chapter, I want to take a step back to consider not just how specific languages are framed by language technology developers, but how they appear to understand "language". This question perhaps allows us to get at the "infrastructural" lens highlighted by Ehsan et al. (2022), which considers the "social and political conditions that make the system possible in the first place" (2022, p. 1307). Here, these "conditions" include broad beliefs about what languages are and should be, in addition to the institutions developing speech technologies discussed in the previous chapter.

In Section 6.2, I present an analysis of the way the terms "accented", "accented" and "non-native" are used in papers published at the speech (technology) conference Interspeech. I am particularly interested in interrogating whose language is perceived as "accented", how language varieties are described and what the motivations for this research are in the first place. The paper has been submitted to Interspeech and is intended as a (hopefully constructive) piece of commentary for researchers working on speech (and especially speech technologies) who might not be familiar with sociolinguistic research. Rather than chiding researchers for their "wrong" use of these terms, the aim is to highlight limitations of broad terms like "non-native", "accented" or "foreign" and ways that some applications (such as "accent reduction" and "pronunciation assessment") can be damaging for speakers.

In Section 6.3, I discuss qualitative and quantitative analysis of the way language and languages are conceptualised in public-facing materials, including marketing and documentation of three ASR products: Amazon Transcribe, Google Text-to-Speech and Mozilla Common Voice.

Exploring which language varieties these providers support and how they describe them, as well as how they engage in wider discourses about the values of language(s) and language technologies, we can understand how language ideologies are reproduced through (commercial and open-source) language technology planning.

In Section 6.4, I discuss the key findings from both Section 6.2 and Section 6.3, with a particular focus on how researchers in academic and commercial settings frame language variation and linguistic diversity. I find that existing ideologies which co-construct languages and nations and confer differential (economic and cultural) value to different varieties appear prevalent. I also consider how language and language variation is conceptualised as a "problem", "resource" or "right" within language technology planning.

6.2 Everyone has an accent: "Accented speech" and speech technology research

This section has been published as Nina Markl and Catherine Lai (2023). "Everyone has an accent". In: Proc. INTERSPEECH 2023*, pp. 4424–4427. DOI: 10.21437/Interspeech.2023-1847*

6.2.1 Introduction

As speech technologies become increasingly ubiquitous, especially for some "high-resource" (standard) languages and their speakers, "accent variation" has become a popular research topic at speech technology conferences such as Interspeech. The recognition that both performance and availability of speech technologies is sharply unequal between different language communities, has further encouraged a focus on "inclusion" as highlighted in the special theme of this conference.

In this paper we want to draw attention to the motivations behind research on accent variation at Interspeech and the way researchers define and discuss "accented speech". Surveying papers published at Interspeech between 2004 and 2022, we notice that terms like "accent", "accented" and "non-native" are frequently used in under-specified ways that could hinder interpretability and reproducibility of research results. Drawing on a sociolinguistic perspective, we also make the case that some of the applications resulting from research on "accented speech", and the way researchers talk and write about accents could harm the very communities who are the intended benefactors of language technology development. We'd like to encourage researchers working on speech to think carefully about their research motivations and how that motivation affect what aspect of language variation really matters to their research question.

Note that we chose to reproduce quotes from Interspeech papers without attribution to

authors to avoid singling out individual authors.

What's an accent?

In linguistics teaching and research, *accent* is often distinguished from *dialect*: *accent* describes pronunciation (*segmental* and *suprasegmental* phonology), while *dialect* also encompasses syntax and lexicon, (see e.g., Trudgill in the Encyclopedia for Language and Linguistics (Trudgill 2006, p. 14) and Crystal in the Dictionary of Linguistics and Phonetics (Crystal 2009, p. 3)).

In everyday language, both *dialect* and *accent* are generally only used to describe some, non-standard varieties. Some speakers might be described as "having an accent" (implying that other people "don't have an accent"). In anglophone linguistics, "accent" and "dialect" are used as neutral descriptors – language varieties might differ in terms of phonetics, phonology, lexicon and syntax and they might differ in *social status* (depending largely on who speaks them) but they are all considered equally complex and rule-governed. As a result, **all varieties, including the "standard variety" could be described as dialects** and **everyone has an accent** (Lippi-Green 2012).

Linguistic variation perceived as accent variation is often tied to the identity of a speaker or speaker group, in particular in terms of geography and social class (Crystal 2009, p. 3). For example, in the context of British Englishes[1], "Received Pronunciation" (RP) is an accent associated in particular with upper class speakers, due to its use by upper class speakers and transmission in private schools (Agha 2003; Fabricius 2018). Other British English accents[2] are very strongly associated with particular regions, cities or areas, like Liverpool ("Scouse"), Manchester ("Mancunian"), Newcastle ("Geordie"), Glasgow ("Glaswegian") and (East) London (e.g., "Cockney" and "Multicultural London English"). Varieties associated with (post-)industrial areas like those, also retain strong associations with a lower socioeconomic status (similarly to how RP is associated with the upper class). In the context of the UK specifically, these associations still matter as some accents are considered to have "higher status" which can result in linguistic discrimination (Sharma et al. 2022; Levon et al. 2021; Craft et al. 2020). Studies also show that second language speakers and those perceived to have a "foreign accent" are affected by language-based discrimination, for example in employment (Timming 2016; Ramjattan 2019). Second language speakers, especially racial or ethnic minorities, are often framed as "deficient" speakers who *lack* linguistic skills which speakers of the standard variety have (Rosa and Flores 2017; Ramjattan 2022; Wright and Brookes 2018).

To summarise, we can draw on Agha (2003), who argues that the term *accent* is "neither

[1]For an excellent introduction to this issue in the context of the United States, see Lippi-Green (Lippi-Green 2012)

[2]Most regional and/or social varieties also differ to some extent in lexicon and syntax, making them *dialects*. However, many people frequently apply their "native" accent to the "standard dialect", adapting only the phonology, not syntax and lexicon. In sociolinguistics, the term "variety" is often used to encompass languages, accents and dialects.

very precise nor free of ideological distortion" (p 232). Firstly, *accent* often "implicitly presupposes a baseline against which some sound patterns — but not others — are focally perceived as deviant, foregrounded accents" (Agha 2003, p. 232). Secondly, accents do not *just* describe sound patterns in isolation but are inherently linked to a specific group of social identities (Agha 2003). Finally, accents are usually discussed as intrinsic features of a speaker (or their speech) which are either present or absent: some people don't have an accent, others do (Agha 2003). The reality is more complicated as the geographic or social descriptions of an accent depend on the listener's identity (Agha 2003).

6.2.2 Methods

We analysed all 94 papers returned by the search term "accented" in the ISCA archive. To understand the characteristics of "accent" research in Interspeech, we also provide a cursory analysis of the 319 papers published on "accent" in this period.[3]

We manually categorised how the word "accent" or "accented" was applied in the abstract and/or introduction of the paper each paper: prosodic prominence (e.g., "pitch accent")[4], first language varieties (e.g., "native accent"), second language varieties (e.g., "foreign accent"), or methods relating to first or second language varieties (e.g., "multi-accent").[5] We then qualitatively analysed how the papers discuss "accent" and "accentedness". Specifically, we look at how accents, speakers and listeners are described and what motivations researchers provide for researching accents in speech technologies.[6]

6.2.3 Who has an accent?

As shown in Table 6.1, half of papers about "accent" published between 2004 and 2022 focus primarily on prosodic prominence – we won't discuss those further in this paper. Of the remaining 161 papers, less than a third are specifically about first language varieties and speakers (L1), with the rest explicitly addressing second language speakers and varieties (L2) or theories and methods concerning accent variation more broadly (L1-or-L2). Conversely, less than a third of papers using the term "accented" discuss prosody, with the plurality focusing on second language speakers and varieties.

Of the 66 papers discussing accent variation, most provide some descriptions of relevant accents. The level of detail in these descriptions varies widely, with some just naming relevant varieties or corpora (n=19), while others provide specific phonetic characteristics of the

[3]The searches were conducted using the the "paper" search function in the ISCA archive. In February 2023, a search for "accented Interspeech" returned 96 results, with 94 available full papers."accent Interspeech" in the ISCA archive returned 320 results, with 319 available full papers published between 2004 and 2022.

[4]In linguistics, the term "accented" most frequently refers to prosodic prominence, i.e., lexical or pitch accent, as in "accented syllable" (Crystal 2009, p. 3).

[5]Papers investigating prosody in first or second language accents were categorised as L1 or L2 rather than prosody. 2 papers did not discuss any sense of the word "accent" and were categorised as NA.

[6]The annotations and tags are available here: `https://github.com/ninamarkl/Thesis-data`

Search term	L1	L2	L1-or-L2	Prosody	NA
"accent"	47	65	49	154	2
"accented"	15	37	14	28	0

Table 6.1: Distribution of Interspeech papers by topic. "Accent" is most frequently used in the context of prosody, but "accented" is most frequently applied to L2 speakers and L2 varieties.

variety and/or some demographic details about the speakers and listeners involved (n=40). It is notable that most papers focussing on specific L1 or L2 varieties use abstract terms like "foreign accent" or "accented" in the paper title and abstract. Only 12 papers specifically name the relevant accent in the title, and another 11 mention the language (but not the accent). This approach emphasises the broad applicability of findings or methods much of the research aims for, where specific varieties are meant to serve as examples. This can be appropriate if the methods or findings truly generalise beyond those specific varieties. In some cases generalisations about "foreign accents", do, however, inadvertently reinforce the idea that all "accented speech" is very similar, which as both linguistic and speech technology literature shows, is not the case.

Describing speakers or language use as "accented" as in "accented speech" or "accented English" also implies the existence of "unaccented" speech or speakers. This is an explicit assumption in some Interspeech papers which refer to "unaccented" or "non-accented" speech or speakers, or speakers who have "no accent" (as discussed below). From a (socio)linguistic perspective these labels are not particularly meaningful, if we assume that all speech is characterised by an accent of some kind. In the "lay" context discussed in the introduction, "no accent" or "unaccented" is a way of referring to the "unmarked" variety, usually the standard variety. This may be quite difficult to interpret for readers unfamiliar with the sociolinguistic context, and, as discussed below, study participants may also differ in the way they interpret these terms.

6.2.4 Who are the speakers and listeners?

As Cheng et al. (2021) highlight in the context of psycholinguistics, the vagueness of "non-native speaker" impedes effective study design, efficient recruitment of participants, clear interpretation of results, and, ultimately reproducibility. Who is considered a "native speaker" varies between researchers and, importantly in self-reporting studies, among speakers themselves (Cheng et al. 2021). There is also huge variability between "non-native" speakers. Baese-Berk et al. (2020) note that there are some "common aspects of non-native speech" across different target and first languages, such as generally slower speech rate compared to L1 speakers and specific target language features which can be challenging for learners with a range of different backgrounds (e.g., two features of English: voiced stops in word-final position or vowel reduction in unstressed syllables) (p. 3).

However, most models of second language acquisition (grounded in empirical studies) posit

that the phonology (system of speech sounds) and articulatory settings (the way speakers use their vocal tract habitually) of learner's "first" language has important effects on how they perceive and produce sounds in any additional languages (Baese-Berk et al. 2020). As the strong interest in "accentedness" in the Interspeech literature evidences, L2 speakers who share the same first language still vary widely in their spoken language production. Some individual differences like habitual speech rate appear to carry across languages (Bradlow et al. 2017). Furthermore, speakers also acquire sociolinguistic variation in their L2, depending on where, when and how they acquire and speak it (Hall-Lew and Elliott 2015; Meyerhoff and Schleef 2012; Nance et al. 2016). To complicate this even further, the context of the speech recording such as task (e.g., reading/conversation), style (formal/informal), topic (e.g., topics which are or are not emotive like work, family, memories), relationship to the interlocutor and accent of the interlocutor have all been shown to affect how people speak in their L1 and L2 (Baese-Berk et al. 2020; Labov 1972).

Grouping together speakers with different linguistic backgrounds – both in terms of L1 and in terms of their exposure and use of L2 thus risks obscuring a lot of variation. As Baese-Berk et al. (2020) note, perception crucially depends on the listener, not just the speaker. Listener expectations and (local) context such as the order in which stimuli of different speakers are presented, the degree of familiarity of listeners with different varieties and lexical frequency all affect perception tasks like accent classification or accentedness ratings (Baese-Berk et al. 2020; McGowan and Babel 2019). It is therefore particularly notable that many studies focusing on speech perception provide little or no description of speaker or listener demographics. Of the 17 perception studies, 7 only mention the L1 and gender of the speakers and 9 only mention L1 and gender of the listeners.

6.2.5 Why do we research accents?

The plurality of papers on "accented" speech focus on automatic speech recognition (n=27). Other popular topics are perception studies and phonetic description of different varieties (n=21) with fewer studies focusing primarily on language identification or speech synthesis.

A particularly interesting aspect of many perception studies are "accentedness ratings", which are employed in 15 papers. In these studies, listeners provide evaluations of second language speakers' accents (these ratings are not used for L1 speakers). As discussed above, the level of detail in the description of these listeners differs, but 14 studies confirm at least that they are native speakers and provide a gender distribution (one paper only mentions "human raters"), while some highlight relevant details such as familiarity with other varieties or residential history.

Most of these studies employ a scale ranging from "no accent" to "strong" or "heavy" accent. Two studies include scales of "no foreign accent" to "strong foreign accent". Three studies instead ask listeners to categorise short audio clips as "foreign" or "native". In this way almost all of the "accentedness rating" research explicitly invokes the notion of "unaccented

speech" or "unaccented speakers". The distinction between "foreign" and "native" is also particularly complicated in pluricentric, global languages like English which have a larger number of very different "native" varieties and many multilingual "native" speakers. One study employing accentedness rating makes the implicit hierarchy within different varieties explicit by asking listeners to describe speakers as "native: the speaker sounds native (e.g., US, UK, Australian)" or "non-native: the speaker sounds like a learner of English (e.g., Korean, Japanese, Philippine)", which was, for the purposes of that study distinguished from different "degrees" of "Indian accent" of speakers who "sound Indian": "subtle", "clear", "pronounced" and "very thick".

While there are different motivations for these studies including understanding human language processing, pronunciation assessment is a common theme. As shown in the quote below (drawn from the aforementioned study focussed at "quantification of Indian accent"), at the most extreme end, this research can frame "accents" as "inappropriate":

> "To be successful in this industry, there is an increased demand for employers to be able to detect the heaviness of an accent so that they can assign employees to appropriate job categories, or give them additional training to refine their accents as appropriate for their jobs."

In addition to framing some ways of speaking as (in)appropriate for specific jobs (in this case, in a customer-facing call centre), it suggest "training" as a kind of remedy to this linguistic deficiency. The pressure placed in particular on migrants and workers in and from the Global South to participate in "accent reduction training" is well documented (Ramjattan 2022; Cowie 2007). It is embedded in wider discourses around "appropriate" or "professional" speech, in which "accent" or language is often used to stand in for race and where linguistic discrimination is inextricably linked to racism (Rosa and Flores 2017; Craft et al. 2020).

A small number of studies focus on "accent reduction" from a technical perspective by applying "accent conversion" and on "speech error detection" using ASR. Both of these approaches are primarily motivated through use in second language teaching. Conceptually, this too relies on a notion of a "target pronunciation" or "correct pronunciation". While it is certainly the case that many learners of an additional language want to avoid miscommunication, not all pronunciation variation or even pronunciation "errors" lead to miscommunication (especially among human interlocutors who can often easily recover intended meaning by accessing the wider linguistic and non-linguistic context). Statements like the ones presented below reinforce notions of "one correct pronunciation", generalise across the extremely heterogeneous group of L2 speakers and "accents" as an impediment to (an undefined notion of) intelligibility "typical" (only) to L2 speakers.

> "Second-language (L2) English learners typically present accents and mispronunciations, which highly impact their intelligibility in practical communication."

> "Correct pronunciation is known to be the most difficult part to acquire for (native or non-native) language learners."

> "We focus on two major aspects of foreign accents: mispronunciations and improper prosody (rhythm, phonemes duration, and pauses)."

> "The goal of automatic pronunciation evaluation is to build an automatic system which can measure the quality of pronunciation given input speech."

These statements concerning "correct" pronunciation or "pronunciation quality" likely appear innocuous to most readers, including many linguists. Notions of "language proficiency" and "accentedness" as evaluated in comparison to some "ideal" or "prototypical" "native speaker" and the importance of standard varieties as linguistic targets for L1 and L2 speakers are deeply ingrained in language teaching (Cushing and Snell 2022). While acknowledging that these notions, as well as models of "target pronunciations" can be useful for learners to avoid miscommunication and feel confident in their L2, it is also important to note that not all speakers orient towards "native speakers" (Nance et al. 2016). Furthermore, as is particularly obvious in the context of (migrant) workers in international anglophone settings, "accent targets" are often imposed externally (e.g., by an employer) rather than freely chosen by learners (Ramjattan 2022; Ramjattan 2019; Cowie 2007).

6.2.6 Conclusions

Against this background, we would like to encourage researchers working on accent variation, especially in the context of speech technology, to think carefully about the underlying assumptions and motivations of their research.

Drawing on an ill-defined, or undefined notion of "accent" or "foreign accent" or "non-native speaker" risks erasing important variation in way that makes it much harder to solve real research problems like sharp differences in ASR performance for different varieties. While it is difficult to generalise about ASR performance about "foreign accents", or even seemingly better-defined notions like "British English", it is possible to identify performance differences between more narrowly defined varieties of the same language (Markl 2022b; Koenecke et al. 2020). Close examination of the language variation which "triggers" speech recognition errors could further be used to improve systems (Chan et al. 2022; Choe et al. 2022; Wassink et al. 2022).

While it makes sense to keep descriptions of speakers, listeners and corpora very brief in light of page limits at venues like Interspeech, we would recommend thinking carefully about these description and providing any details required to replicate the study or interpret results. Describing the variety or accent with relevant phonological and/or social details is often appropriate. One option could be to include standardised language variety tags (e.g., BCP-47) as recommended for natural language processing (Bender and Friedman 2018).

Being very specific in perception study design also aids reproducibility. "Accentedness" ratings, for instance, are not necessarily correlated with "intelligibility" (Baese-Berk et al. 2020). Alternatively, some studies investigating intelligibility ask listeners to write down or re-speak what they heard in addition to or instead of "rating" speakers.

Finally, accents are just as much about identity as they are about pronunciation. The way we speak is always shaped by who we are and how we want to be perceived. On the one hand, that means that the extent to which we can definitively "label" different accents is limited, and that is important to be very clear about how we go about naming speakers and their varieties. On the other hand, the connection between identity and accent also means that while all accents are "equally valid" from a linguistic perspective, they are not all "equally valued" in society.

6.3 Language management and (commercial) ASR development

The work presented in this section was conducted with Stephen Joseph McNulty. The methodology and framing of this paper was developed jointly through discussion. Data analysis was also conducted jointly. SJM is the lead author on a manuscript based on this work which will be submitted to a language policy journal.

6.3.1 Introduction

In this section, we are building on our exploration of the concept of "language policy arbiters" and different orientations toward language technology design, by exploring the language management of three different language technology developers. We do this by analysing public-facing materials (promotional materials, websites, documentation) related to Amazon Transcribe, Google Speech-to-Text and Mozilla Common Voice. Amazon Transcribe and Google Speech-to-Text are cloud-based, "off-the-shelf" ASR tools primarily targeted at businesses. Mozilla Common Voice is an open-source compilation of speech datasets which can be used to develop ASR tools.

On the one hand, we consider what these providers (claim to) offer: which language varieties they support likely tells us something about their motivations and interests. We focus on two kinds of data: firstly, we consider how each developer caters for linguistic diversity. In this context, we are also interested in how specific language varieties are named, labelled, and classified. Categories, even, or perhaps especially, as encoded in technology, are not neutral (Bowker and Star 2000). This is particularly clear when the "object" of classification and discussion is language or a language variety (Schneider 2019). Beyond this analysis of "what's on offer", we are also interested in the way language technology developers discursively frame language, language variation and linguistic diversity. As Stark and Hoffmann (2019) highlight, the "metaphors we deploy to make sense of new tools and technologies" are important as they both "highlight the novel" and "obscure away". In the context of language technologies, these new metaphors around technology interact with old(er) discourses around language and linguistic diversity.

Taken together, these data give us an insight into how they understand the services they are providing, the problems they believe themselves to be solving, and the their understanding of their role in the wider linguistic ecology. They also reveal something about the language ideologies which shape development – and the language ideologies they in turn transmit in and through their products.

6.3.2 Data

Data was collected in summer 2022 from public facing websites for Amazon Transcribe, Google Text-to-Speech and Mozilla Common Voice. The documents in Table 6.2, were analysed by

Source	Name of source
AMZ1	Amazon Transcribe Developer Guide
AMZ2	Amazon Transcribe Homepage
AMZ3	Amazon Transcribe Blog Post: "Break through language barriers with Amazon Transcribe, Amazon Translate, and Amazon Polly"
AMZ4	Amazon Transcribe FAQs
AMZ5	Amazon Transcribe Features webpage
AMZ6	Amazon Transcribe News article: "Amazon Transcribe now supports automatic language identification for multi-lingual audio"
AMZ7	Amazon Transcribe Language Support
GOO1	Google Cloud Speech-to-Text documentation: "Automatically detect language"
GOO2	Google Cloud Speech-to-Text documentation: "Speech-to-text basics"
GOO3	Google Cloud Speech-to-Text documentation: "Language Support"
GOO4	Google Cloud Speech-to-Text Homepage
MCV1	Mozilla Common Voice Languages webpage
MCV2	Mozilla Common Voice Governance Doc V1.0
MCV3	Mozilla Foundation blog: "MCV 'Our Voices' Model and Methods Competition – Taking Part"
MCV4	Mozilla Foundation Blog: "How we're making Common Voice even more linguistically inclusive"
MCV5	Mozilla Common Voice About webpage
MCV6	Mozilla Discourse post: "Apply to be a Common Voice Language Reps [Expression of Interest]"
MCV7	Mozilla Foundation blog: "Keeping language rights at the heart of Common Voice"
MCV8	Mozilla Common Voice powerpoint

Table 6.2: Data analysed for this study.

myself and SJM.[7]

6.3.3 Amazon Transcribe

Amazon is an important actor in language technology development and artificial intelligence research more broadly (Rikap 2023). For the purposes of this paper, we are focussing on Amazon's automatic speech recognition tool "Amazon Transcribe" which is offered through Amazon Web Services (AWS)[8]. Amazon Transcribe[9] is specifically targeted at businesses, with key applications being (customer service) call analytics, automatic captioning and clinical documentation[10]. Clients can generate transcripts using a user-interface or can integrate the service into their own applications in a variety of programming languages. Developers working with Amazon Transcribe do not need any prior experience with automatic speech recognition or linguistics

[7] Data is available here: `https://github.com/ninamarkl/Thesis-data`

[8] Amazon describes AWS as "Amazon Web Services (AWS) is the world's most comprehensive and broadly adopted cloud, offering over 200 fully featured services from data centers globally" – `https://aws.amazon.com/`

[9] `https://aws.amazon.com/transcribe/`

[10] For the latter, AWS has a specialised product Amazon Transcribe Medical which we do not analyse further here.

(though it is possible to modify the models somewhat). While Amazon's smart speakers tend to be marketed towards private customers, AWS machine learning and data analytics services are primarily targeted at corporate clients.

Language support

One way of understanding the Amazon Transcribe's (AT) language policies, is to consider which language varieties it supports. In April 2022, Amazon Transcribe provided transcription for 23 "languages", and 37 "variants". As discussed below, these terms are not explicitly defined. Figure 6.1 shows the way these varieties and their associated features are presented to clients (excerpt of AMZ7).

AT supports 9 variants for English, 2 variants of each Arabic, Chinese, French, German, Portuguese and Spanish and one variant for all other languages. 16 of the available varieties are predominately spoken in Europe and North America (and as discussed below, clearly associated with European and North American states). 11 of the varieties are associated with Asia. The only languages included commonly spoken or holding official status anywhere on the African continent are Arabic and Afrikaans, and English, French (though no variety explicitly linked to Africa).

Language variants are identified with a language code and label. The language code consists of the BCP-47 ISO code which combines a standardised language code (ISO 639-1) with a standardised country or region code (ISO 3166-1) (e.g., "en-AB"). The label consists of the English name of the language, optionally followed by a geographical or political descriptor (e.g., "English, Scottish"). The descriptors and the use of the ISO country codes evidences an implicit assumption that language variants "belong" to different political and/or geographical units.

For pluricentric languages other than English, an "unmarked" variant is clearly assigned to one political unit where it is an official language: German (de-DE), French (fr-FR), Spanish (es-ES), Portuguese (pt-PT). These are clearly contrasted from the "marked" variants associated with other territories: "German, Swiss", "French, Canadian", "Spanish, US", "Portuguese, Brazil". English is a notable exception in this pattern. Rather than providing one "unmarked" variant, all variants bear a geographical and/or political marker: English, British (en-GB), English, US (en-US), English, Australian (en-AU), English, Irish (en-IE), English, New Zealand (en-NZ), English, South African (en-ZA), English, Indian (en-IN), English, Welsh (en-WL), and English, Scottish (en-AB). Particularly interesting here is perhaps the absence of "English, England" which is instead subsumed in "English, British" and "English, Canada" which appears to subsumed under "English, US".

Not all labels are purely political or geographic: for Arabic and Chinese, the common descriptors "Arabic, Modern Standard" and "Chinese, Simplified" are used, but paired with country codes (Saudi Arabia and China, respectively). These are distinguished from "Arabic, Gulf" (country code: AE) and "Chinese, Traditional" (country code: TW).

Figure 6.1: Language varieties supported by Amazon Transcribe are identified by a label and language code. AT supports different kinds of features for different varieties (such as processing digits and acronyms, and training custom language models).

Supported languages and language-specific features

Language	Language Code	Data input	Transcribing digits	Acronyms	Custom language models	Redacting transcripts	Call Analytics
Afrikaans	af-ZA	batch	no	batch	no	no	no
Arabic, Gulf	ar-AE	batch	no	no	no	no	batch
Arabic, Modern Standard	ar-SA	batch	no	no	no	no	no
Chinese, Simplified	zh-CN	batch, streaming	no	no	no	no	batch
Chinese, Traditional	zh-TW	batch	no	no	no	no	no
Danish	da-DK	batch	no	batch	no	no	no
Dutch	nl-NL	batch	no	batch	no	no	no
English, Australian	en-AU	batch, streaming	batch, streaming	batch, streaming	batch	no	batch
English, British	en-GB	batch, streaming	batch, streaming	batch, streaming	batch, streaming	no	batch
English, Indian	en-IN	batch	batch	batch	no	no	batch
English, Irish	en-IE	batch	batch	batch	no	no	batch
English, New Zealand	en-NZ	batch	batch	batch	no	no	no
English, Scottish	en-AB	batch	batch	batch	no	no	batch
English, South African	en-ZA	batch	batch	batch	no	no	no
English, US	en-US	batch, streaming	batch, streaming	batch, streaming	batch, streaming	batch, streaming	batch

(a) Language varieties supported by Amazon Transcribe.

Supported languages and language-specific features

Language	Language Code	Data input	Transcribing digits	Acronyms	Custom language models	Redacting transcripts	Call Analytics
English, Welsh	en-WL	batch	batch	batch	no	no	batch
French	fr-FR	batch, streaming	no	batch, streaming	no	no	batch
French, Canadian	fr-CA	batch, streaming	no	batch, streaming	no	no	batch
Farsi	fa-IR	batch	no	no	no	no	no
German	de-DE	batch, streaming	batch, streaming	batch, streaming	no	no	batch
German, Swiss	de-CH	batch	batch	batch	no	no	batch
Hebrew	he-IL	batch	no	no	no	no	no
Hindi, Indian	hi-IN	batch	no	batch	batch	no	batch
Indonesian	id-ID	batch	no	batch	no	no	no
Italian	it-IT	batch, streaming	no	batch, streaming	no	no	batch
Japanese	ja-JP	batch, streaming	no	no	no	no	batch
Korean	ko-KR	batch, streaming	no	no	no	no	batch
Malay	ms-MY	batch	no	batch	no	no	no
Portuguese	pt-PT	batch	no	batch	no	no	batch
Portuguese, Brazilian	pt-BR	batch, streaming	no	batch, streaming	no	no	batch
Russian	ru-RU	batch	no	no	no	no	no
Spanish	es-ES	batch	no	batch	no	no	batch
Spanish, US	es-US	batch, streaming	no	batch, streaming	batch	no	batch
Tamil	ta-IN	batch	no	no	no	no	no
Telugu	te-IN	batch	no	no	no	no	no
Thai	th-TH	batch	no	batch	no	no	no
Turkish	tr-TR	batch	no	batch	no	no	no

(b) Language varieties supported by Amazon Transcribe (continued).

Discursive construction of "language"

Initial ethnographic analysis reveals that, despite advertising a product related to language technology, Amazon Transcribe does not appear to include any explicit definitions of "language", "variation", "variants", "dialects" or any other linguistic terminology on its website. Considering that these terms are used across the platform, they are framed as self-evident (not unlike the use of "accent" and "dialect" in academic research discussed above). As AMZ1 below shows, "accent" and "dialect", which are usually distinguished in linguistics literature, appear to be used interchangeably: en-GB, en-AB, en-WL represent British, Scottish and Welsh varieties of English respectively – it is not clear from the documentation if these models can account for lexical or syntactic variation (though results in Chapter 4 suggest that the en-AB model does perform better than the en-GB model on recordings by speakers from Scotland).

> Tip: Providing a subset of language options is particularly helpful in transcribing regional dialects. For example, if you know your media file contains an English dialect from the United Kingdom, but you're not sure which accent is present, provide Amazon Transcribe with en-GB, en-AB, and en-WL and omit the other English language codes. (AMZ1)

Language variation is generally presented as something that Amazon Transcribe can "handle" (see AMZ4). It is notable that in AMZ4, "accented speech" and code-switching are mentioned alongside "background noise" and "overlapping speakers". First, the uncritical use of "accented speech" implies the existence of "unaccented speech", which, presumably poses no challenges for the system. Secondly, while all four "factors" are genuinely challenging for many current ASR systems, two of them (code-switching or translanguaging and "accented" speech) could be resolved more easily through the use of different training datasets.

> Amazon Transcribe service is designed to handle a wide range of speech and acoustic characteristics, including variations in volume, pitch, and speaking rate. The quality and content of the audio signal (including but not limited to factors such as background noise, overlapping speakers, accented speech, or switches between languages within a single audio file) may affect the accuracy of service output. (AMZ4)

Particularly telling are the potential application contexts they highlight in this blog post targeted as potential (business) clients. Rather than focussing on the ways their tools could allow businesses to improve accessibility of their products, or communication between them and their customers, they highlight that it eliminates the "need of a human translator" (and, in the case of ASR, a human transcriber). The example of using automatic (speech) translation to facilitate conversations between doctors and patients expresses Amazon's confidence in the accuracy of their product (since this is an extremely high stakes context), and perhaps says something more profound about the broader technological future it is working towards.

> Break through language barriers with Amazon Transcribe, Amazon Translate, and Amazon Polly. Imagine a surgeon taking video calls with patients across the globe without the need of a human translator. What if a fledgling startup could easily expand their product across borders and into new geographical markets by offering fluid, accurate, multilingual customer support and sales, all without the need of a live human translator? What happens to your business when you're no longer bound by language? (AMZ3)

6.3.4 Google Speech to Text

Google and its parent company Alphabet has also been an important actor in language technology development, and AI development more broadly (Rikap 2023). Like Amazon, they provide Cloud Storage and Cloud computing and ready-to-deploy AI applications including machine translation, automatic speech recognition, speech synthesis and a range of data analysis tools[11].

Here we focus on their cloud-based ASR tool, "Google Speech-to-Text" [12]. Similarly to Amazon Transcribe, private and business clients can integrate the tool into their applications, or access it through an online interface.

Language support

In 2022, Google Text-to-Speech (GSTT) was available for 73 languages and 140 varieties. Similarly to AT, GSTT lacked support for non-European official languages in Africa only including Afrikaans, Amharic, Swahili and Zulu. Coverage across Asia was a lot stronger with 27 languages including English (several of which have multiple varieties). Figure 6.2 shows excerpts from GOO3 which lists all supported varieties and associated features for users.

GSTT also uses the BCP 47 language tags in combination with a language label. Unlike Amazon Transcribe, GSTT labels consist of the English name for the language followed by a (nation) state in brackets. Here too there is an implicit assumption that language varieties are tied to specific territories, and more specifically nation states.

Unlike AT, all varieties are marked with a territory description, including those for which only one variety is supported such as "Amharic (Ethopia)". For some pluricentric languages, GSTT lists a large number of varieties, e.g., Spanish (20), Arabic (17), English (16). The Spanish varieties correspond to 19 states where Spanish is an official language (including in South America, Central America, the Caribbean and Europe), plus the United States (which does not have de-jure official languages). Particularly interesting cases from a language policy perspective include the treatment of Basque, Catalan and Galician all of which are mapped to "Spain". Like AT, the description of Chinese varieties also stands out, with the options "Chinese, Can-

[11] https://cloud.google.com/
[12] https://cloud.google.com/speech-to-text

Figure 6.2: Excerpt from GOO3 showing some of the language varieties supported by Google Text-to-Speech. GSTT supports different kinds of features for different varieties (such as processing digits and acronyms, and training custom language models). Note that all varieties are mapped to a specific territory

Name	BCP-47	Model	Automatic punctuation	Diarization	Boost	Word-level confidence	Profanity filter	Spoken punctuation	Spoken emojis
Afrikaans (South Africa)	af-ZA	Command and search			✔		✔		
Afrikaans (South Africa)	af-ZA	Default			✔		✔		
Albanian (Albania)	sq-AL	Command and search					✔		
Albanian (Albania)	sq-AL	Default					✔		
Amharic (Ethiopia)	am-ET	Command and search					✔		
Amharic (Ethiopia)	am-ET	Default					✔		
Arabic (Algeria)	ar-DZ	Command and search					✔		
Arabic (Algeria)	ar-DZ	Default					✔		
Arabic (Algeria)	ar-DZ	Latest Long				✔	✔		

(a) Language varieties supported by Google Speech-to-Text.

(Ukraine)	UA	and search							
Ukrainian (Ukraine)	uk-UA	Default			✔		✔		
Ukrainian (Ukraine)	uk-UA	Latest Long				✔	✔		
Ukrainian (Ukraine)	uk-UA	Latest Short				✔	✔		
Urdu (India)	ur-IN	Command and search					✔		
Urdu (India)	ur-IN	Default					✔		
Urdu (Pakistan)	ur-PK	Command and search			✔		✔		
Urdu (Pakistan)	ur-PK	Default			✔		✔		
Uzbek (Uzbekistan)	uz-UZ	Command and search					✔		
Uzbek (Uzbekistan)	uz-UZ	Default					✔		
Vietnamese (Vietnam)	vi-VN	Command and search			✔		✔		
Vietnamese (Vietnam)	vi-VN	Default			✔		✔		
Vietnamese (Vietnam)	vi-VN	Latest Long	✔			✔	✔		
Vietnamese (Vietnam)	vi-VN	Latest Short	✔			✔	✔		

(b) Language varieties supported by Google Speech-to-Text (continued).

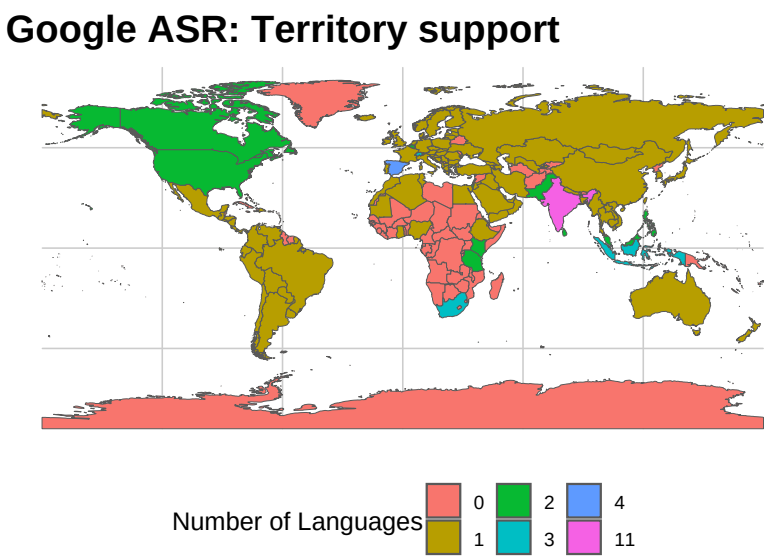


Figure 6.3: Google ASR language support for different territories. Map generated in R based on the information provided in the GSTT.

tonese (Traditional Hong Kong)", "Chinese, Mandarin (Simplified, China)" and "Chinese, Mandarin (Traditional Taiwan)".

Discursive construction of "language"

Like Amazon, Google does not offer any explicit definitions of "language". The terms "languages" and "variants" are used without much explanation, as in G004:

> Support your global user base with Speech-to-Text's extensive language support in over 125 languages and variants. (G004)

In contrast to Amazon, however, it appears to be more interested in supporting different varieties of the same named language. We see this not only in the large number of regional varieties (usually linked to a particular territory or region) but also in the way they advertise their "extensive language support in over 125 languages and variants."

In the documentation, of the system it appears that "variant" and "dialect" are used interchangeably – again tied to region or nation.

> Speech-to-Text's recognition engine supports a variety of languages and dialects. You specify the language (and national or regional dialect) of your audio within the request configuration's languageCode field, using a BCP-47 identifier. (G002)

6.3.5 Mozilla Common Voice

Mozilla differs from the other two companies discussed here in multiple ways. The Mozilla Foundation[13] is a not-for-profit organisation dedicated to open-source technology development. It presents itself as an organisation which "works to ensure the internet remains a public resource that is open and accessible to us all ", a pledge which is accompanied by a manifesto relating in particular to an open and accessible internet, and promoting community-focussed processes in software development[14].

While Mozilla is not primarily known for language technologies, the Mozilla foundation has been a particularly interesting actor in the context of ASR development in recent years. Unlike Amazon and Google, they do not sell an off-the-shelf speech recognition engine. We chose to include them in this comparison, however, because of their Common Voice project which compiles crowd-sourced speech datasets.

Common Voice[15] is a collection of publicly accessible speech datasets. The datasets are crowd-sourced by volunteers who can contribute by recording themselves reading sentences (also collected by volunteers) using Common Voice's web interface. Volunteers can also use a similar interface to validate that recorded utterances match the transcript. As discussed in more detail below, which languages are supported in this way is also in part dependent on volunteers outside the organisation who can request a new language to be added.

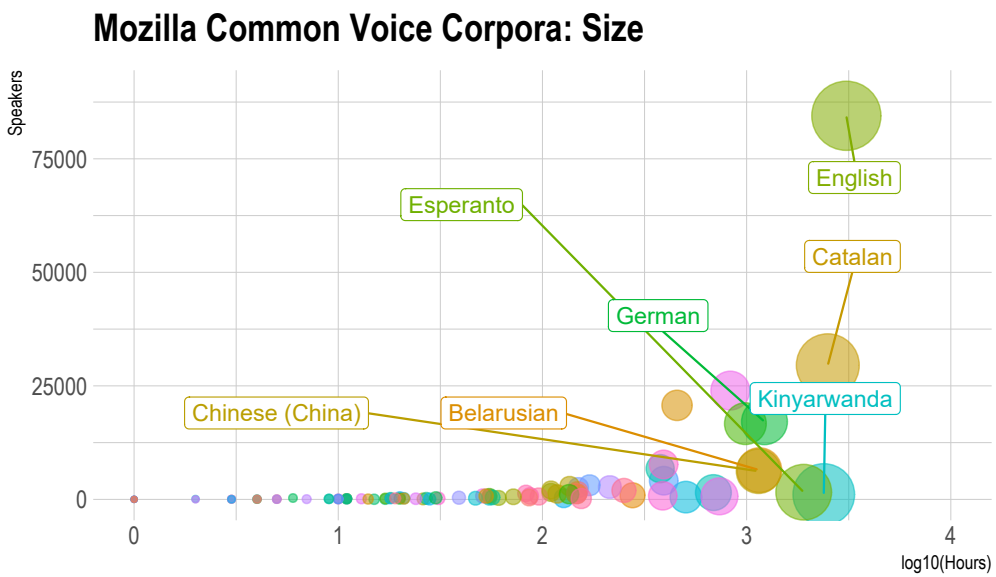
Language Support

As the explicit aim of Common Voice is to "[mobilise] people everywhere to share their voice" (MCV4), it is perhaps unsurprising that the dataset contains a large variety of languages. In July 2022, speech datasets were available for download for 99 languages. For a further 76, the required threshold of validated recordings had not been reached yet.

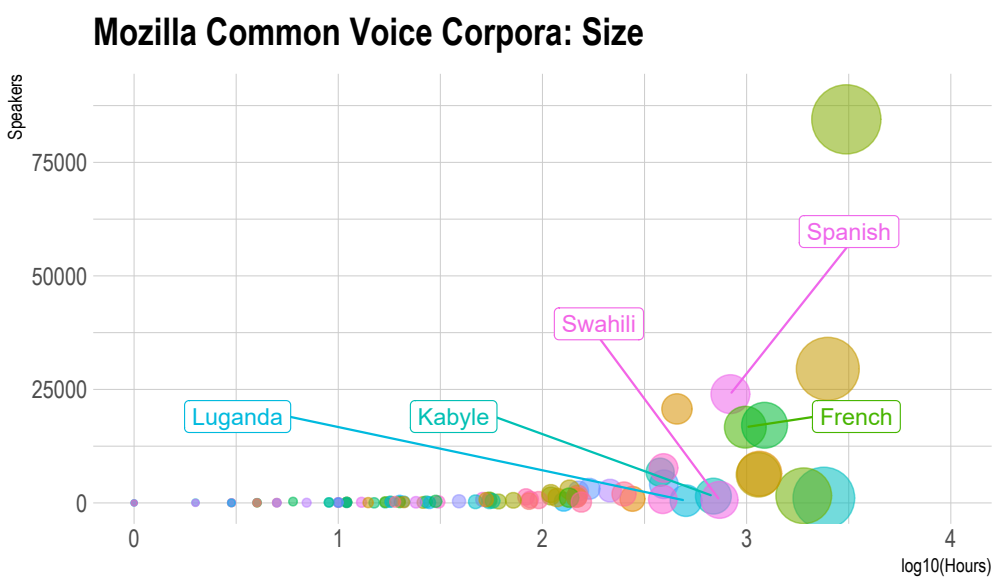
The size of these datasets varies enormously as illustrated in Figure 6.4. Datasets for 19 languages feature fewer than 50 speakers, including some which are supported by GSTT, like Azerbaijani and Macedonian and some of which are not supported by the other providers like Sardinian and Votic. At the other end of the scale, English, Spanish and Catalan boast tens of thousands and contributors. In terms of hours of speech data, English (3077 hours) and Catalan (2498 hours) contrast with the smallest corpora for Asturian and Nepali (both 1 hour). As Tyers and Meyer (2021) show in experiments with the Common Voice datasets, even very small datasets can enable the development of ASR systems which can be useful for (usually limited) tasks like recognising digits and keywords. The role of organised and engaged minority language communities is evident in the fact that for some of the corpora a large number of recordings were created and validated by a small group of speakers. For example, the Luganda corpus of 504 hours was compiled by 501 volunteers.

[13] https://foundation.mozilla.org/
[14] https://www.mozilla.org/en-GB/about/manifesto/
[15] https://CommonVoice.mozilla.org/

Figure 6.4: The size of Mozilla Common Voice corpora differs enormously.



(a) 7 varieties have a corpus size of over 1000 hours of data.



(b) 5 varieties have a corpus size between 500 and 1000 hours of data.

Discursive construction of "language"

In addition to the differences described above, discursive differences in the portrayal of languages and linguistic variation and diversity between Mozilla Common Voice and the other platforms mentioned are especially stark.

Unlike Amazon Transcribe and Google Text-to-Speech, Common Voice provides discussions of terms like "language", as seen in MCV5:

> What is a language on Common Voice? There are lots of ways to think about language. For the purposes of speech recognition models, Common Voice suggests focussing on 'mutual intelligibility', or 'can speakers of this language mostly understand one another if they try to?' (MCV5)

MCV5 is an excerpt from the Frequently Asked Questions section of the website. It is clear from this quote that they understand that "language" cannot be definitively defined across all contexts. Instead, they suggest an appropriate definition for the purposes of speech recognition.

We can see another acknowledgment of the contested nature of this definition in its Governance Document (MCV2), and invite language communities to challenge them if necessary:

> When we establish a guideline - for example defining language for the purposes of the dataset - we will always make space for open conversation about exceptions, for example to take account of political and social history. (MCV2)

In addition to clarifying its use of the term "language" on the platform, one Mozilla Common Voice blog post (MCV4) also acknowledges the non-communicative aspects and values possessed by languages for their speakers – a commentary that is notably absent from the other platforms.

> Language is more than sounds - it can be a sense of home, of belonging, the way we express our emotions and move others. In today's world, languages are not all treated the same. Many - even most - are ignored, threatened, exploited or degraded. (MCV4)

Moreover, it recognises that inequalities exist between different languages and language communities. In this context, they also provide a nuanced definition of "variants". In late 2021, Common Voice introduced the option for volunteers to describe the "variant" of a named language they speak. Up to 10 variants can be added to each language on the platform, in addition to multiple accent tags, with the possibility of adding custom descriptions.

> Variant: A specific form of a language or language cluster associated with a group of speakers, for example those living in a shared region or country, or who have a

> shared culture or heritage, and thus experience vocabulary, grammar and norms that differentiate their speech from others. *Note that we use the word variants rather than dialects because the latter has sometimes been used in derogatory ways, however variants are not limited to geographies. (MCV4)

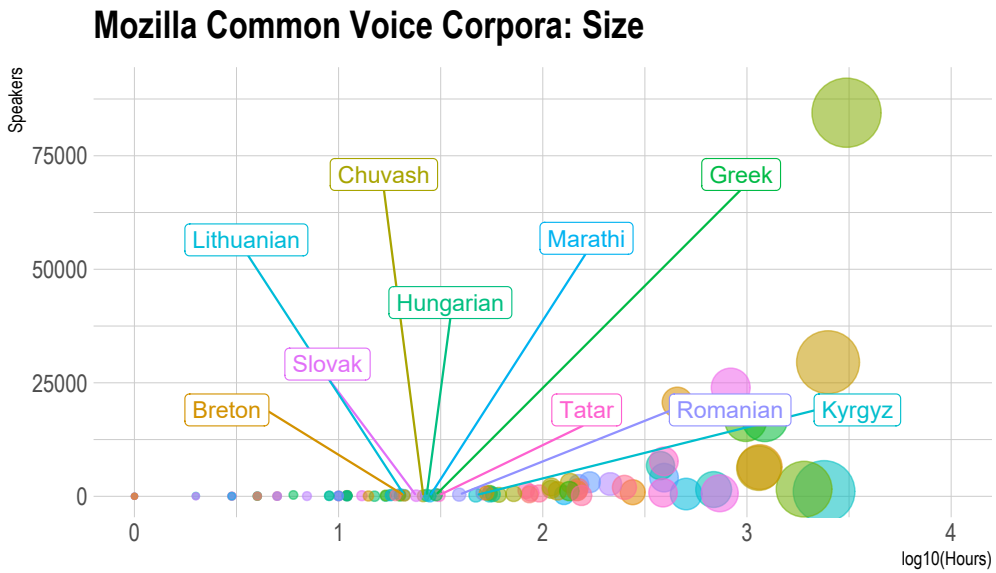
> Some language communities and contributors make use of accent tags, but can feel marginalised and undermined by this. Talking about language is talking about power, and some people want to have the ability to identify their speech beyond 'accent', in ways which respect and represent them. (MCV4)

In MCV4, they also discuss some of the fundamental motivations behind Common Voice – which include creating more inclusive and representative speech datasets in order to build technologies which eventually "work equally well for everyone".

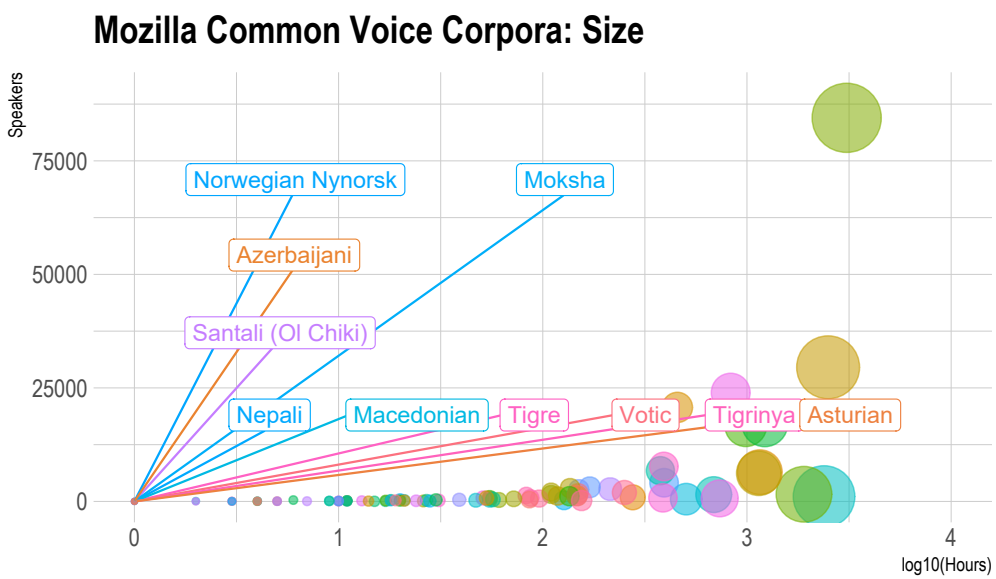
> At present, most voice datasets are owned by companies, which stifles innovation. Voice datasets also underrepresent: non-English speakers, people of colour, disabled people, women and LGBTQIA+ people. This means that voice-enabled technology doesn't work at all for many languages, and where it does work, it may not perform equally well for everyone. We want to change that by mobilising people everywhere to share their voice. (MCV4)

Interesting too are instructions (or lack thereof) for the volunteers providing speech samples. Rather than asking them to read carefully, a "normal voice" is encouraged. (Though this is slightly at odds with the fact that all volunteers are reading pre-selected sentences, rather than speaking freely.)

> Speak in your normal voice! The way you speak is welcome here - we want your accent as it is, and we want your usual volume, style and intonation. (MCV5)



(c) 10 varieties have a corpus size between 20 and 50 hours of data.



(d) 10 varieties have a corpus size of under 2 hours.

6.4 Discussion: Language ideologies in language technology planning

By examining how language technologies do or do not support language variation and linguistic diversity, and exploring how researchers and developers conceptualise language(s), we can glean underlying language ideologies. Here I want to focus on two questions raised by the data presented in Section 6.2 and Section 6.3:

1. How are languages and language varieties conceptualised by researchers?
2. How is "language" understood in language technology planning?

With respect to the first question, I am particularly interested in how languages and speakers are understood in relation to nation states, and how different varieties are constructed as "marked" or "unmarked". Regarding the latter, I am particularly interested in the different values language is afforded by researchers and developers, and how researchers and developers orient towards "language" (and language variation) more broadly.

6.4.1 What's a language (variety)?

Both academic and corporate speech technology development generally assumes "languages" as pre-existing, bounded, named objects. This is reflected in the fact that terms like "language", "accent", "dialect" and "variety" are rarely explicitly defined by language technology developers (in academic and corporate contexts). Mozilla Common Voice is a notable exception in the way it highlights cultural and social aspects of language, and the way languages and linguistic communities are constructed through discursive and social processes.

Generally, language varieties are framed as associated with *nations* and spoken by *native speakers*. Some varieties (and speakers) are further assumed to be *unmarked*, while others are *marked* (or accented).

Language(s) and nations

In the materials discussed in this chapter, language varieties are usually explicitly tied to nations. This choice on behalf of researchers and developers reflect existing language ideologies, according to which languages are "out there" ready to be classified (not dissimilar to the way "data" are understood to be "out there" ready to be "collected" as discussed in Chapter 5) – and more specifically that the appropriate way to classify languages and different language varieties follows national borders. This assumption is particularly obvious in the way standardised language descriptors which implicitly or explicitly link languages to nations and a particular set of formal features are used by developers. The fact that these standardised descriptors are promoted as best, "neutral" practice in documentation (including by myself) highlights just

how fundamental the notion of a "bounded" language or language variety is to all language technology development.

Schneider (2019) calls this bias towards nations as a natural(ised) locus for language and language communities "methodological nationalism". "Languages" as bounded entities, she points out, are "the outcome of national discourse", constructed alongside and as part of nations (Schneider 2019, p. 4). By explicitly associating language varieties to nations (e.g., "French (Canada)" or "French, Canadian") in materials for documentation and advertisement, the existing ideological linkages between language and nation are reinforced. Furthermore, by linking (some) standard varieties which already carry "official" status to the nation, but not others, linguistic diversity within nations (within and between "named languages") is erased (Irvine and Gal 2000), and existing nation-level language policies are reinforced. In the context of nation (states) which do not have a de-jure (declared) official languages policy, selective links between language and nation may reinforce de-facto language policies and/or be reflective of some (but never all) linguistic diversity. For example, both Amazon and Google support two language varieties they associate with the United States: "English (US)"/"English, US" and "Spanish (US)"/"Spanish, US". In addition to erasing diversity by constructing two homogeneous varieties from two groups of extremely heterogeneous varieties, this selection of "US" varieties also erases all other languages spoken in the United States, most notably indigenous languages.[16] It appears to reflect the fact that while the United States does not have a de-jure official language policy, varieties of English and Spanish are most commonly spoken and represented in the linguistic landscape (Spolsky 2003).[17] This popularity likely translates to a (perceived) market for language technologies. At the same time perceived linguistic (and perhaps cultural) similarity can lead to the erasure of national differences: Amazon Transcribe notably does not provide a model for "English, Canadian", perhaps because "English, US" is perceived as sufficiently similar.[18] A slightly different kind of erasure occurs when "English, British" is named alongside "English, Scottish" and "English, Welsh", and the English nation is conspicuously absent (or, most likely, understood as subsumed within "British"). More broadly, which languages, varieties and nations are missing reveals something about which linguistic markets are prioritised.

Marked and unmarked varieties and speakers

Where language variation is accounted for, some varieties (but not others) tend to be constructed as "unmarked" varieties spoken by "unmarked" speakers. Both marked and unmarked varieties are often under-specified through the use of broad short-hands. Particularly common (and notable from a sociolinguistic perspective) is the term "accented" which can be observed across commercial and academic speech technology research. In implying the ex-

[16] Though as discussed in the context of Māori, that may be preferable to the alternative.

[17] They are also framed as two distinct bounded language varieties – even though many speakers likely experience them as one linguistic system or repertoire (Otheguy et al. 2015).

[18] Similarly, we see "German, Germany" and "German, Swiss" – but not "German, Austria".

istence of a "unaccented" variety and "unaccented" speakers, it very explicitly constructs a distinction between marked and unmarked or normative and non-normative speech.

Within academic research on "accented" speech, the focus is generally on "non-native" speakers or those who "have foreign accents". This distinction between "native" and "non-native" too naturalises the link between nations and languages, positioning some speakers as "foreign" (to both language and nation). By under-specifying L2 speakers, variation within the category is erased. The lack of critical discussion or definition of the term itself, too naturalises the category (as it implies that it does not require definition or discussion).

These beliefs around marked and unmarked speech are also reflected in the way different varieties are constructed. Amazon Transcribe notably presents language varieties as named language followed by an *optional* national descriptor. The national descriptor, which, as discussed above, links the variety to a nation, is not used for *all* varieties and creates a hierarchy of markedness. For example, Amazon Transcribe lists "French" and "French, Canadian" (rather than, say, "French, French" and "French, Canadian"). In this way, some varieties are constructed as the "default" and presented as a "language from nowhere" (Woolard 2008) while others are highly localised. As Woolard (2008) notes, by removing the "social origins" of a variety it can be presented not just as anonymous but as authoritative.[19] Interestingly (and perhaps tellingly) this approach is not applied to English, Arabic, Chinese and Hindi.[20] In the case of English, each of the 9 varieties is explicitly marked with a national descriptor which refers to a nation where English is generally considered a first language (including Wales and Scotland which are nations but not independent nation states).

Linguistic markets and language ideologies

In considering which language varieties and speakers are supported in language technology development and how they are talked about reveals underlying language ideologies and wider motivations in research and development. The way varieties are framed as inherently tied to nations reflects the centrality of the nation (state) as a category, to both researchers and large, international corporations.[21] Which varieties (within and between nations) are prioritised, reveals the linguistic markets commercial developers target and deem valuable. We also see the re-production of hierarchies between varieties, with some varieties and speakers framed as "marked" while others are positioned as normative.

[19] Woolard is interested in the relationship between a "voice from nowhere" and public space and discourse. The "authority" these hegemonic varieties provide, is the authority to speak and more importantly be heard in public. In a way, ASR and other speech technologies confer a similar kind of authority to some but not other varieties and speakers – though not always through an explicit process of "anonymisation" as varieties can still be linked to, in particular, nation states.

[20] There is only one variety of Hindi which is presented as "Hindi, Indian". Chinese is presented as "Chinese, simplified" and "Chinese, Traditional". Arabic is presented as "Arabic, Gulf" and "Arabic, Modern Standard". Notably, all varieties are linked implicitly to nations or regions through ISO codes.

[21] A detailed exploration of the way Google and Amazon relate to nations and states is beyond the scope of this chapter – but it stands to reason that there are wider targetting particular linguistic and national markets for their products.

Crucially, we do not just see a "mirroring" of existing language ideologies but a re-entrenchment and perpetuation of them as they are "encoded" into language technologies. Gal and Woolard (1995, p. 129) name "translation, the writing of grammars and dictionaries, the policing of correctness in national standards, the creation of linguistic and folklore collections or academies" as practices which (re)produce "bounded" and "naturalised" languages. In the context of ASR, we can see the compilation, distribution and reuse of datasets as one such practice which reproduces bounded languages. We can think of ASR models and other language technologies as another such practice. Language technologies modelling a specific "variety" encode this variety in a similar way to grammars and dictionaries, delineating the "boundaries" of lexicon and variation based on the training dataset. Where speakers interact with them, they also perform the function of "policing", where variation outwith the "boundaries" is not recognised.

In the next section, I explore how the value of language and language variation is framed in discourses around language technology development.

6.4.2 What's the point of language and linguistic diversity?

Above, I considered how beliefs about languages as co-constructed with nations and the cultural and economic value associated with different varieties is reproduced in language technology development. In addition to these, we also see beliefs about the *purpose* and *value* of language and languages which inform how language technology is being designed and marketed.

As I argue below, one productive way of interpreting these beliefs (and the actions which follow) is through the frame of "language planning", an area of language policy research. Drawing on Ruiz (2016 [1984])'s "Orientations to Language Planning", I consider how academic and corporate language technology researchers and developers conceptualise language as a "problem", "right" or "resource".

Language (technology) planning

"Language planning" as a field, has historically been, as Ruiz (2016 [1984]) highlights (originally in 1984), focussed on "solving language problems". The "language problem" is often positioned as a necessary precursor to language planning. Much of this language planning work has been done in service of what we might term "nation planning": post-colonial national development (and modernity more generally) was generally seen to require the development of a national standard language (often alongside the coloniser's language) (Ricento 2000; Johnson 2013). Crucially, especially in this early phase, linguists positioned themselves as "experts" in this task which was framed as a purely technical, scientific process (Johnson 2013; Heller and McElhinny 2017).

There are clear parallels to the way language technologies are discussed. The impulse

to identify a "problem" which can then be "solved" with a technical (or technocratic) solution provided by experts, is foundational to almost all technology development but is perhaps particularly fascinating in contexts where the "problem" is about building machines which can "understand", "transcribe" or "translate" language. Additionally, language technologies are explicitly or implicitly positioned as solving "social problems" related to language – for example by enabling communication across "language barriers" through machine translation. Some of these social problems algorithmic systems which include language technologies supposedly "fix" are much broader such as employment discrimination (see Chapter 7) and language endangerment (see Chapter 5). In those cases it is particularly clear that the deep roots of these problems are not addressed by a technical fix (Benjamin 2019b; Drage and Mackereth 2022; Hoffmann 2021a).

Orientations towards language

Ruiz (2016 [1984], p. 14) defines three "dispositions toward language and its role, and towards languages and their role in society". They are "language-as-a-problem", "language-as-a-resource" and "language-as-a-right" and have been most productively applied to analysing how nations "manage" linguistic diversity. Over the last forty years, this framework has "been elevated to the level of paradigm" in the field of language planning (Hult and Hornberger 2016, p. 30).

Language-as-a-problem The "language-as-a-problem" orientation identified by Ruiz (2016 [1984]) arises from "a monolingual ideal and assimilationist mindset" (Hult and Hornberger 2016, p. 34). Within this perspective, speakers of minority languages are framed as deficient (see also Rosa and Flores 2017) and educational policy is focussed on transition from minority languages to majority languages (only), with the overall aim of limiting multilingualism which is assumed (or argued) to hinder national unity (Hult and Hornberger 2016, p. 33).

Transposing these frameworks from national (educational) policy to language technology development, we can ask whether speakers of "non-standard" varieties are also positioned as "deficient" and whether linguistic diversity and multilingualism is framed as a problem. In Section 6.2, I highlight that some of the research on "accented speech" focussing on second language speakers (of English, usually), explicitly reproduces discourses of "deficient" speakers through tasks such as "accent quantification" and "pronunciation assessment". Here, the "problem" the technology is meant to solve is both "classifying" speakers and, in the longer run, "fixing" their speech production. More subtly, much of the research invokes the notion of "unaccented" speech, which clearly positions many speakers, again in particular second language speakers, outwith the linguistic norm. Both in academic and non-academic work considered here, this non-normative speech is often framed not just as a problem to solve, but specifically as a problem for the already-existing technology, as highlighted in the quotes below. While this language can be very difficult to avoid, it does subtly (or not so subtly) shift the burden on the

speakers, rather than the system.

> "Automatic speech recognition is hindered by the linguistic differences occurring in accented speech." (Interspeech paper)

> "Non-native speech is well-known for reducing speech recognition performance." (Interspeech paper)

More broadly, we see the way (global) multilingualism and linguistic diversity is framed as a problem to be solved through language technology. This is perhaps most clear in the rhetoric employed by Amazon (reproduced here for convenience):

> Break through language barriers with Amazon Transcribe, Amazon Translate, and Amazon Polly. Imagine a surgeon taking video calls with patients across the globe without the need of a human translator. What if a fledgling startup could easily expand their product across borders and into new geographical markets by offering fluid, accurate, multilingual customer support and sales, all without the need of a live human translator? What happens to your business when you're no longer bound by language? (AMZ3)

Here language is framed as a "barrier" which "binds" businesses to particular linguistic markets. Language technologies (or more specifically Amazon's language technologies) can supposedly solve that problem. A notable difference between this notion of "language-as-a-problem" and the one set out by Ruiz (2016 [1984]), is that speakers here are not required to learn the majority language (or any other language). In this way, machine translation can support linguistic diversity (as communities and speakers in theory do not need to shift their language practices too much[22]), while fundamentally positioning multilingualism as a problem requiring a technical solution. Particularly telling in Amazon's example too is that the explicit problem machine translation is solving according to them is not just multilingualism, but specifically the cost of "human translators"[23]. Overall, language technologies, in particular machine translation, but also automatic speech recognition, are positioned as ways to "fold in" speakers of a wide range of (named, national, standard) languages into a wider global culture and economic system (Bird 2022). While this without doubt brings many benefits, such as access to information only available in English, it might also further normalise the dominance of English (and English speakers).

Overall, we see a "language-as-problem" framing in the way second-language speakers and speakers of non-standard varieties are framed as posing a "problem" to existing language technologies. Linguistic diversity is also identified as a "problem" which can be addressed through

[22] Of course, they might need to shift their practices to the standard of "their" named language to make use of these systems.

[23] On the value of professional, "human" translation and how this is changing in the age of machine translation, see Carmo (2020).

a technological solution. Importantly, these solutions can support linguistic diversity among speakers, but also lead to the devaluation of specific linguistic and (inter)cultural communication skills.

Language-as-a-resource In Chapter 5, I discussed the problems with understanding language data (only) as a resource to be extracted. Ruiz (2016 [1984]) proposed "language-as-a-resource" as an orientation in which languages are considered a slightly more abstract "personal and national resource" with intrinsic and extrinsic value (Hult and Hornberger 2016, p. 33). The speakers of different (minority) languages are further understood to have valuable expertise, and linguistic diversity is seen as strengthening the nation (Hult and Hornberger 2016, p. 33).

We can see this attitude reflected in Mozilla's framing of language and linguistic diversity. They specifically highlight the intrinsic value of language varieties, and speakers are positioned as authorities on their own languages (e.g., in naming varieties). Rather than "solving" the problem of language variation, Mozilla positions itself as solving the problem of data bias and under-representation in speech datasets. This focus coupled with an appreciation of the intrinsic value of languages could lead to the development of genuinely reliable and useful speech technologies in many different varieties which could meaningfully support linguistic diversity and multilingualism. Adopting a more complex model of linguistic ecology and understanding that different languages have different roles within it, especially for multilingual speakers, may enable communities to do this sustainably by prioritising varieties of "contact languages" used between communities over "low-resource" local languages (Bird 2022).

However, as we have seen in the context of Māori in Chapter 5, representation in open-source datasets also creates new problems, especially related to data sovereignty. If data is open-source, how can any community ensure that it is not used in ways that they do not want it to be[24]? Where data is licensed for commercial use (like Common Voice datasets are), the question of who gets to profit of these datasets becomes relevant. Similarly, despite Mozilla's stated desire for a very "diverse" dataset, (global and local) inequities regarding internet and broadband access, perceived and actual ability to participate in crowd-sourced projects, and the task set-up which involves reading sentences (requiring both literacy and some agreed-upon standard) still create barriers to entry. Regardless of community involvement in the selection and naming of language varieties, the final "arbiters" of what is and is not included in Common Voice still appear to be Mozilla developers rather than communities themselves, as the infrastructure for recording, validating and distributing datasets is ultimately controlled by them.

[24]This is, of course, not a problem limited to language technologies. Widder et al. (2022) provide an interesting case study of how developers of open-source "deep fake" models understand their ethical obligations and the potentials for abuse.

Language-as-a-right The framing of language as a right is perhaps particularly interesting as language technologies proliferate. It originates in a focus on legal protections for speakers of different varieties (Hult and Hornberger 2016). Within this perspective, language is understood as "mediating access to society" (just like in the "language-as-a-problem" frame) and as a result the right to use "one's" language in particular contexts is legally protected. None of the materials discussed above explicitly discuss the use of particular language varieties as a "right". However, as discussed in Chapter 5, discussions on "low-resource" varieties do invoke the idea of protecting the status of endangered languages and their speakers (Bird 2022).

With increasing deployment of ASR systems (and other language technologies) in "high-stakes" contexts, such as hiring and accessibility tools, "language-as-a-right" could be a productive concept to explore the risks and harms of predictive bias or lack of available speech technologies.

6.5 Conclusion

In this chapter, I explored how "language" and "linguistic" diversity appear to be conceptualised by researchers in academic and commercial ASR development. I first focussed on how research on "accented" speech published at a leading speech (technology) conference is motivated and framed. I found that speakers and listeners involved in these studies are often described in vague terms, hindering interpretability and reproducibility of some research. In addition to applying the term mostly to second language speakers, there is an underlying assumption in much of the work that accents can be (unproblematically) "quantified" or "detected" both by human listeners and automatic systems. Few papers acknowledge the social meanings (or personae) associated with those different accents and the variation in how listeners perceive them. In this way, some work reproduces existing discourses of "deficient" speakers and "unaccented" speakers. Similar notions of "marked" and "unmarked" speech are visible in the way commercial ASR engines discuss different language varieties. Generally, languages are understood as bounded and inherently connected to a nation and/or geographical territory. By extension, speakers are (again unreflexively) categorised as "native" or "non-native" speakers.

Descriptions of ASR systems (by researchers and companies) thus reproduce prevalent beliefs about language (variation). Choices regarding which language varieties to support furthermore often reflect the perceived "market value" of a variety and speaker group. Language and language variation is further often framed as a problem or obstacle to overcome. This can be contrasted with discourses taken up by Mozilla Common Voice which highlight extrinsic and intrinsic values of language(s). Unlike the other commercial and many of the academic materials discussed in this chapter, Mozilla also emphasises the complexities of "naming" language varieties. Applying the concept of "language planning attitudes" not only helps us understand what researchers and developers already do, but also what they *could* do. Par-

ticularly interesting here is the idea of "language-as-a-right" and how that could be extended to language technologies. Similarly, we could imagine language technologies which fit into a broader "language-as-a-resource" framing which welcomes and supports linguistic diversity in part through technology.

In Chapter 7, I explore how principles of "diversity" and "inclusion" could be implemented in ASR (corpus) development, and discuss some of the limits of this approach.

Chapter 7

Building better speech technologies

7.1 Introduction

Building on the work presented in this thesis, the question that remains is: "How do we build better speech technologies?"

One potential answer is to create benchmark datasets which more closely represent "real" speech and "real" speakers. These would allow for more reliable evaluation of ASR systems in realistic settings and highlight predictive bias. In Section 7.2, I present a paper published at the first Workshop on Bridging Human-Computer Interaction and Natural Language Processing at the European Association for Computational Linguistics 2021. In it, we highlight the advantages of using naturalistic speech data in evaluating ASR systems, and argue for the importance of compiling diverse benchmark datasets, conducting qualitative analysis and considering user experience(s) as well as the wider sociotechnical context of the ASR system. The qualitative analysis is presented using data from the Lothian Diaries Project (Hall-Lew et al. 2022), as also discussed in Chapter 4.

In part building on this call for action, I then present work on a new "diverse" speech dataset: The Edinburgh International Accents of English Corpus (EdAcc). In Section 7.3, I discuss the compilation process of this multimodal corpus of spontaneous English-language conversation involving speakers from a wide range of linguistic backgrounds. As a member of a larger team (consisting of speech technology researchers), I helped shape the data compilation with a particular focus on eliciting "naturalistic" speech (drawing on methods from sociolinguistics), collecting valuable metadata and documentation and ensuring participants were able to provide informed consent to not just the dataset but also potential uses of it. In Section 7.4, I include the data statement (Bender and Friedman 2018) associated with the Edinburgh International Accents of English Corpus. In Section 7.5, I present some of the results we obtained when evaluating different ASR systems with this dataset.

Finally, in Section 7.6, I return to the "three lenses" introduced in Section 2.5.1 and the critiques of a narrow focus on "bias" and "fairness". I take a step back to consider the limits of this approach to "inclusion" and the potential harms of "unbiased" ASR technologies.

7.2 Context-sensitive evaluation: Considering user experience & language variation

This section was published in: Nina Markl and Catherine Lai (2021). "Context-sensitive evaluation of automatic speech recognition: considering user experience & language variation". In: Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing. *Online: Association for Computational Linguistics, pp. 34–40. URL: https://aclanthology.org/2021.hcinlp-1.6*

7.2.1 Introduction

Automatic Speech Recognition (ASR) has become a common tool in human-computer interaction, enabling, for example, voice user interfaces and (imperfect) automatic captioning of multimedia content. As with other language technologies (e.g. Sap et al. 2019; Blodgett and O'Connor 2017), rapid improvements in performance have not been equal for different user groups. As Blodgett et al. (2020) show, discussions of this "bias" are often poorly defined, not grounded in explicit normative judgments and divorced from socio-historical contexts, origins and harms of the system behaviours. In this paper, we argue that researchers at the intersection of speech and language technologies (SLT), human-computer interaction (HCI), and sociolinguistics are well-placed to consider the experiences and social context of different speaker/user groups in critical quantitative and qualitative evaluations of ASR systems. Knowledge about language variation and its relation to society coupled with expertise from HCI allows us to understand how predictive biases reflect larger social structures and ideologies about language, and how they affect users.

After presenting prior work on language variation and ASR, a case study of self-recorded audio diaries collected for the Lothian Diary Project[1] highlights the need for a context-sensitive approach to ASR evaluation which we outline.

7.2.2 Language variation, bias and ASR

Blodgett et al. (2020)'s critique notwithstanding, predictive bias, defined here as error and outcome disparities for different user groups (Shah et al. 2020), has become a research focus in SLT and other machine learning fields as applications are extended to high-stakes contexts such as hiring, policing and banking where they have been shown to (re)produce structural inequalities (see e.g. Benjamin 2019b). Predictive bias also appears to be prevalent in commercial ASR systems for English[2].

Recent work describes stark racial bias in commercial American English ASR systems (including Google's Cloud Speech) (Koenecke et al. 2020), with much higher word error rates

[1] https://lothianlockdown.org/

[2] The focus here is on English, but predictive bias is likely to affect stigmatised and unstandardised varieties vis-a-vis standardised varieties of other languages too.

(WER)[3] for speakers of African American English (AAE) than white speakers of (Californian) American English. Notably, these types of error disparities appear to be driven by under-representation of AAE training data both for the acoustic modelling (Koenecke et al. 2020) and the language model used to decode sequences of phones into utterances (Martin and Tang 2020). "Regional" variation has also been reported as a source of unequal performance, with particularly high error rates reported on YouTube's captions for speakers from Scotland and (the US state) Georgia (Tatman 2017). Similar to more recent work, YouTube captions have been found to perform worse for African American speakers (Tatman and Kasten 2017). These problems are not limited to proprietary systems, as Mozilla's open source system DeepSpeech performs significantly worse for speakers of Indian English than "American English"[4] (Meyer et al. 2020), and also fails to transcribe AAE morpho-syntactic variation correctly (Martin and Tang 2020). While some early research has suggested ASR performance differences based on (binary) speaker gender (Adda-Decker and Lamel 2005; Benzeghiba et al. 2007; Tatman 2017), it is unclear that gender by itself is a significant factor in recent systems (Tatman and Kasten 2017; Meyer et al. 2020). Koenecke et al. (2020) suggest that the interaction of gender and race is significant, with differences between Black men and Black women being more significant than between white men and white women or men and women across race[5]. These results appear to be linked to speaker's speech styles (e.g. in Adda-Decker and Lamel 2005) and use of dialect features (Koenecke et al. 2020), both of which have long been documented to pattern with gender (see Labov 1990, for a classic paper) and could be correlated with gender in training and test sets. Other work in this space has focused on the potential of ASR to improve accessibility of audio media and digital technologies, looking at experiences of Deaf and hard of hearing users (Glasser 2019) and dysarthric speakers (De Russis and Corno 2019; Young and Mihailidis 2010). For both groups commercial ASR systems perform quite poorly, though the severity and amount of errors varies by speaker. Research on predictive bias in commercial ASR for regional varieties of English beyond the United States and in the context of systems not exclusively trained on American English, as well as experiences of second language learners of English, and other groups who are potentially particularly reliant on ASR to access computing technologies such as elderly people, is sparse.

From a linguistic perspective, no language variety or speech style is inherently more difficult, incorrect, or inappropriate than any other. There are, however, powerful ideologies regarding the relative status of different varieties and styles which are rooted in broader socio-historical contexts and reflect the social status of the groups who speak them (Woolard and Schieffelin 1994). In addition to being stigmatised in "traditional" contexts of power in society, varieties spoken by marginalised communities appear to be (not coincidentally) under-represented in

[3]WER is an edit-distance measure capturing the number of deletions, substitutions and insertions required per word to match a reference transcript.

[4]Meyer et al. (2020)/Mozilla do not specify speaker race or region within the US.

[5]A finding which echoes work in other ML domains and other areas of SLT highlighting the way that multiple demographic axes linked to interacting structures of oppression (e.g. gender and race) cannot be considered separately (Buolamwini and Gebru 2018; Jiang and Fellbaum 2020)

the data we use to build and evaluate speech technologies, leading to substantial predictive biases making speech technologies less accessible to already marginalised groups.

7.2.3 Lothian Diaries: A case study

The Lothian Diary project is an ongoing interdisciplinary research project inviting residents of the Lothians region of Scotland to contribute self-recorded audio and video diaries about their experiences of the COVID-19 pandemic. The more than 120 diaries collected so far are highly variable in recording quality, number of speakers and topics discussed, and participants are diverse[6] in terms of age, gender, linguistic background, ethnicity, socio-economic class and level of education. Edinburgh and the surrounding Lothians region are of particular interest for sociolinguistic research because of the capital region's status as a centre for higher education, finance and tourism. In addition to the variation within Scottish English[7] between different areas and different socio-economic groups within the city, there is also a wide range of other first and second language varieties of English, as well as other languages. The Lothian Diary project also includes many of these other varieties of English, rather than focusing on speakers with long residential histories in a particular area (as is often the case in sociolinguistic work) or first language speakers (as is usually the case in SLT evaluation). The recordings form a highly naturalistic and exceptionally varied data set. ASR is used here to facilitate social science research which requires accurate and complete transcriptions (achieved through manual correction).

So far, 13 diaries submitted by participants who agreed to have them made public, have been processed with the Google Cloud Speech-to-Text API[8] (GC STT). Diaries (16 kHz FLAC files) were processed in their entirety using the model used for long audio files which uses asynchronous speech recognition. WER was computed separately for each speaker using `sclite`[9]. In the following section, we present a brief qualitative error analysis.

WER for individual speakers varies dramatically (see Table 7.1). Some of these errors appear to be related to accent differences. For example, Scottish speakers' pronunciations of *I* or *I've* are frequently mistranscribed as *ah* or *of* and other accent-based errors include: *cat* [kaʔ] $>$ *car*, *living* $>$ *leaving*, *hating our* $>$ *heating are*. However, there is also significant variation within each accent group. GC STT fails to transcribe filled pauses (*uh, um*) and word fragments and occasionally deletes false starts and repetitions. Furthermore, errors appear to be more prevalent in the vicinity of hesitations and repetitions. As a result speakers who produce more hesitations and repetitions tend to have higher error rates, while people who appear to read from prepared notes tend to be more fluent and have lower error rates. The highest WER in this sample derives from a recording by a Scottish English speaker who produces many false starts,

[6] though not representative of the Scottish population

[7] "Scottish English" is used here as a broad term including the continuum between Scots and Scottish Standard English (see Stuart-Smith 2004)

[8] `https://cloud.google.com/speech-to-text`

[9] `https://github.com/usnistgov/SCTK`

ID	G	Variety	WER
RF	F	Scottish English	46.4
La	F	Scottish English	35.7
CE	F	Scottish English	20.9
AA	M	Scottish English	29.8
MR	M	Scottish English	55.3
DL	M	Scottish English/Scots	88.9
Li	F	Southern British English	29.5
JW	M	Canadian English*	25.3
L	F	L2 English (L1: Lithuanian)	31.5
S	F	L2 English (L1: Cantonese)	27.7
MG	F	L2 English (L1: Italian)	35.3
JL	M	L2 English (L1: Filipino)	40.8
A	M	L2 English (L1: Chinese)	70.2

Table 7.1: Word Error Rates for different participants vary widely both across and within groups (lower is better). *Decoded using 'en-US' language option, for all others 'en-GB' was used

word fragments and a number of Scots words (which the system likely would not recognise under any circumstances). Words are also often substituted by a wrong (but often grammatically appropriate) inflectional form (e.g. past tense $>$ present tense).

All of these errors are particularly challenging for the accurate and complete transcription of spontaneous and conversational speech, especially for social science research where researchers (users) might consider hesitations, false starts and filled pauses important as they convey pragmatic information. Considering impacts of this predictive bias, transcripts of speakers who produce more "fluent" speech are much more easily interpretable. Retrieving speech content and speech style of less fluent speakers as well as some second language speakers, on the other hand, requires more labour and time, potentially negating any benefits of ASR.

7.2.4 Proposed methods

To document predictive bias in ASR in a way that is mindful of 1) user experience, 2) socio-historical and (socio)linguistic context, 3) (potential) harms (re)produced by the system, and 4) technical aspects of ASR, we need to draw on methodologies and knowledge from HCI, sociolinguistics, research on fairness in AI, and SLT.

Intersectional benchmarks

ASR systems are usually evaluated in terms of their WER, for one or more unseen test sets (often including well-established benchmark sets). As seen in the case study above, word error rates vary strongly across individual recordings and speakers, and (benchmark) test sets (e.g. Barker et al. 2017) are becoming increasingly naturalistic and (potentially) diverse; a recent

state-of-the-art system by Google (Chiu et al. 2018) was trained and tested on "representative" data drawn from Google's voice-search traffic. However, even assuming that the test sets are representative of the developer's users, it is 1) not clear that the intended or current user base is reflective of all use cases or potential users (especially if the system is sold to third parties as with GC STT), and 2) possible or even likely that significant variation in performance between user groups is hidden by reporting an average across all tested recordings. Importantly, as Black feminist scholarship has pointed out, multiple demographic axes linked to interlocking structures of oppression (e.g. race and gender) cannot be considered separately (Crenshaw 1991). It is thus important that in addition to disaggregating by language variety to also consider, for example, gender to create an "intersectional" benchmark (see also Costanza-Chock 2020). This approach has been successful in highlighting disproportionate predictive bias for particular subgroups in other ML domains (e.g. darker-skinned women in facial analysis: Buolamwini and Gebru 2018; Raji and Buolamwini 2019), and SLT (Jiang and Fellbaum 2020).

To apply an intersectional benchmark to a larger sample of the Lothian Diaries, we intend to match short audio snippets with the same reference transcript produced by different speaker groups to isolate pronunciation effects, and look systematically at potential differences in content and speech style (following Koenecke et al. 2020).[10]

Qualitative error analysis

Intersectional benchmarks alone are not enough however, as WER does not account for the context or effect of an error. Understanding the context of errors is useful since errors are both more likely to occur and to be severe in particular phonetic, prosodic and lexical contexts. Like us (though working with a very different system and data), Goldwater et al. (2010) find that words before or after hesitations, repetitions and word fragments, turn-initial words and infrequent words are more likely to be misrecognised and that erroneous substitutions are often different forms of the same lexeme (e.g. *ask/asked*). While some of these errors can be easily disambiguated through context, others (e.g. *can/can't*) could be quite disruptive to communication. Word errors can also lead to domino effects, where one wrongly decoded word feeds into further erroneous predictions (Martin and Tang 2020). While metrics which are more sensitive to the type and context of the error or directly model human evaluations have been proposed (Nanjo and Kawahara 2005; Morris et al. 2004; Mishra et al. 2011; Kafle and Huenerfauth 2020) they are not widely adopted and extensive qualitative error analysis is rare. A context-sensitive approach would be particularly interested in the type of error and its effect given the linguistic context.

[10] Note that this is no longer forthcoming work.

User experience

Evaluations of SLT systems rarely reflect explicitly on how users interact with them[11]. However, because both (perceived) severity and impact as well as prevalence of errors depends on recording and task, understanding how people use ASR-based technologies in their daily life is important. Future work concerning predictive bias in ASR would benefit from incorporating HCI methodologies like interviews, ethnography and qualitative surveys to gain a deeper understanding of users' experiences. So far, researchers in HCI have been particularly interested in how people interact with voice user interfaces (e.g. Porcheron et al. 2018; Luger and Sellen 2016), though little attention has been paid to the role of accent and dialect. Furthermore, especially given the context of the recent shift to increased remote work and education, applications of cloud-based speech recognition for personal or business use extend beyond voice user interfaces to automatic captioning of audio and video lectures and meetings. Domain-general and naturalistic recordings of continuous spontaneous speech pose a particular challenge to ASR systems, and insights into what types of errors users perceive to be particularly disruptive and common depending on their linguistic and demographic background should inform development and evaluation of ASR systems. For example, in the context of the Lothian Diary Project the goal of ASR is to produce transcriptions which can be used by linguists and other social science researchers to analyse both what participants are saying and how they are saying it. Every aspect of their speech, including disfluencies and repetitions as well as specific lexical choices (e.g. past tense vs present tense) are relevant to this analysis and should as such be preserved in a transcript. Furthermore, because most speech in this context is largely unplanned, higher error rates around disfluent or informal speech are particularly disruptive. When applying the proposed methodology to other use cases (e.g. automatic captioning of video lectures or business meetings) interviews with stakeholders can clarify what types of errors are particularly disruptive.

Considering context and impacts

Considering the broader societal context in which an ASR system is developed and implemented allows us to identify the specific harms it could inflict on users and (sometimes at least) see the underlying societal structures giving rise to predictive bias. Identifying risk and causes in turn allows us to mitigate harms (and, in future systems, bias). In the case of commercial ASR (in English), research suggests that predictive bias is a result of under-representation of varieties of marginalised speaker groups in proprietary training and test sets. For many open source and licensed corpora used to train and benchmark ASR systems, incomplete documentation makes it difficult to estimate representation; the commonly used Switchboard (Godfrey and Holliman 1993) and TIMIT corpora (Garofolo et al. 1993) (both US English) and Mozilla's

[11]Though intended use is sometimes implicit in the choice of training and test data: e.g. Google's use of voice search data (Chiu et al. 2018)

recent open-source Common Voice corpus[12], for example, do not record speaker race. The speaker characteristics of training sets depends on the broader societal context. For example, use of commercial speech recognition (e.g. in the case of Google's system) and participation in scientific studies (e.g. the licensed corpora) or crowd-source tasks (e.g. Mozilla Common Voice) differs across demographic groups (for example based on income and education). Imbalanced corpora are also tied to ideologies around whose ways of speaking are considered "legitimate", "correct" or "native".

Some of the more obvious specific harms of predictive bias include difficulties using voice user interfaces, which for some users are crucial assistive technology. As ASR spreads into high-stakes contexts such as hiring, substantial harms could be incurred if systems perform worse for already marginalised groups, effectively encoding "accentism" and linguistic prejudice in automatic systems. Even assuming no prediction bias across different speaker groups, the use of ASR in automatic analysis of video interviews to recommend or rank applicants (e.g. HireVue[13]) risks real harm in the case of even small recognition errors and potentially entrenches existing language ideologies around "professional", "fluent" or "competent" speech patterns. For example, *HireNet* (Hemamou et al. 2019) extracts information about prosody and speech fluency to predict "hireability" (as annotated by recruiters). Other harms include less usable automatic captions and potential downstream effects as described in our case study.

7.2.5 Conclusion

We have proposed an approach to ASR evaluation which considers the experiences of different user/speaker groups, sociolinguistic context and potential impacts of predictive bias. We argue that this interdisciplinary approach is necessary to significantly advance our understanding of ASR usability. We particularly invite perspectives from the fields of human-computer interaction in order evaluate speech and language technologies as systems situated in specific sociolinguistic and socio-technical contexts which perform specific tasks for specific (language) users.

[12]available here: `https://commonvoice.mozilla.org/en/datasets`
[13]`https://www.hirevue.com/`

7.3 Compiling a dataset of "real conversations": The Edinburgh International Accents of English Corpus

Some of the contents of Section 7.3 were published in: Ramon Sanabria, Nikolay Bogoychev, Nina Markl, Andrea Carmantini, Klejch Ondřej, and Peter Bell (2023). "The Edinburgh International Accents of English Corpus: Towards the Democratization of English ASR". in: ICASSP 2023*, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10095057*

7.3.1 Introduction

A central barrier to the development of ASR systems which robustly support a wide range of varieties is the absence of suitable speech datasets to train or even reliably evaluate these systems. Advances in English language ASR are usually reported as word error rates (WER) on established benchmark datasets. These datasets are unrepresentative of many real-life ASR applications in terms of both *style* and *speakers*, focusing overwhelmingly on read (recited) or planned spontaneous speech by speakers who tend to be white, highly-educated, "native" speakers of (standard American) English (see also Section 5.4). Unrepresentative test datasets risk "evaluation bias" where predictive bias is simply missed because the dataset used to evaluate the system only represents one type of speaker (Suresh and Guttag 2021).

The current state-of-the-art on Switchboard Godfrey and Holliman (1993) – an US English spontaneous (transcribed) speech dataset – is 4.3% WER (Tüske et al. 2020). Recent systems (Baevski et al. 2020; Baevski et al. 2021) produce word error rates as low as 1.4% and 3.2% respectively on the US English read speech dataset Librispeech (Panayotov et al. 2015). The robustness of these results has been questioned as these systems perform less well on other varieties of English (Lin et al. 2022; Hsu et al. 2021).

As discussed in depth in Section 5.4, Switchboard, like other established benchmark corpora, only represents a particular subgroup of US English speakers and contains significant "gaps". While there is very little metadata about the speakers included in Librispeech, the focus on recited speech from English language public domain books suggests that standard American English is likely over-represented[14]. Martin (2022, p. 180) further confirms an almost complete absence of features of African American English. Second language speakers are not represented in either dataset. Beyond questions of diversity, Switchboard is, despite being "spontaneous" still a bit "unnatural": it contains telephone conversations between people who generally did not know each other and which focus on some predefined topics.

To enable better evaluation of existing ASR systems on conversational speech data from a wide range of English speakers, we compiled a dataset consisting of audio and/or video recordings of English language conversations between friends with diverse linguistic backgrounds. We introduce the first release of an ongoing project Edinburgh International Accents of English

[14] In the paper introducing the dataset, Panayotov et al. (2015, p. 5208) note that they used a "simple automatic procedure" to select recordings "with accents closer to US English".

Corpus (EdAcc) in Sanabria et al. (2023). Our dataset contains almost 40 hours of video call dyadic English language conversation between speakers who know each other. The conversations range in duration from 20 to 60 minutes. EdAcc contains more than 40 self-reported English accents from speakers from 51 different first languages. We also collected their linguistic background including any languages they speak, how long they have spoken English, and places where they have lived for extended periods of time. We release our dataset with the responses from each speaker publicly[15], and a Data Statement can be found in Section 7.4.

7.3.2 Dataset design

My role in the interdisciplinary team was to shape the design of the data collection and data annotation processes. I was particularly concerned with 1) designing a process closely modelled on sociolinguistic interviews to elicit relaxed, informal, "naturalistic" speech, 2) ensuring that participants would be able to self-describe their linguistic background and provide rich documentation for the dataset, 3) thinking critically about the ways "accent labels" were employed throughout this process. In the design of the EdAcc dataset, I tried to ensure that all possibly relevant information about speakers' linguistic backgrounds was included including the languages they used in their daily lives, age of onset of English and their location history.

Our data collection process is designed to provide a simple framework for eliciting naturalistic speech by allowing speakers to record relaxed conversations using the Zoom video call software. A questionnaire distributed to participants further enables the curation of a well-documented and diverse dataset. To capture participants' linguistic backgrounds we asked them about: any first languages (acquired before age 5), any second languages, when they started learning English, which language they mostly use in different domains (work, friends, family) and any places where they have lived for more than three years. We also asked them how long they have known their conversation partner and whether they usually speak English with them. Finally, we asked them to self-describe their accent in English. To capture their social background we also asked about their age, gender, ethnic background and education.

I was also very concerned with the way different varieties would be labelled in the dataset. Rather than specifically recruiting speakers of a named variety, we asked speakers to self-describe their accent (see Section 7.4) for the exact wording of the questionnaire). Unsurprisingly this approach yielded a lot of different accent labels, and included both geographic descriptions (e.g., "Scottish English", "Italian") and more evaluative comments (e.g., "fluent"). While we draw on a standardised accent category in Sanabria et al. (2023) (see Section 7.5), the publicly released dataset only contains the descriptors provided by speakers. This was very important to us, to ensure we do not misrepresent or (mis)categorise participants without their consent or introduce bias to this dataset by attributing the "wrong" label.

[15] `https://groups.inf.ed.ac.uk/edacc/`

Recruitment

Participants were initially recruited through the authors' personal and professional (local and global) networks. As the data collection progressed, speakers were also recruited through online micro-work platforms[16]. Each participant was compensated with 10 GBP for every 15 minutes of conversation.

Recording procedure

The use of video call software made our approach scale-able and simple: conversations could be recorded at the same time in multiple places, allowing us to reach speakers in different parts of the world. Due to software limitations, only one audio channel could be recorded – instead of one channel for each speaker. Conveniently, this setting also replicates real-world acoustic conditions where ASR engines are usually deployed. The contributors were provided with detailed instructions on how to record the conversation (on audio, and if they wished to do so, video[17]). We provided some discussion prompts about topics such as hobbies. This design is inspired by data collection procedures in sociolinguistics (Van Herk 2018; Schilling 2013), where an engaging topic can reduce self-consciousness and promote more natural speech patterns. Informal speech is further promoted by the interlocutor – all participants talked with friends or acquaintances. To a certain extent, this design may also limit linguistic accommodation effects, though it should be noted that all interactions involve some accommodation or alignment (Giles et al. 1991). Finally, the "observer's paradox", where participants in an experiment feel self-conscious and adjust their speech, is further reduced by asking participants to self-record their conversations (Schilling 2013; Van Herk 2018). Before starting the conversation, each speaker was also asked to read the same control passage[18] to allow evaluation on a controlled domain and enable detailed linguistic analysis.

Anonymity and ethics

EdAcc is designed specifically to make our data compliant with CC-BY-SA so that data can be fully shareable. Each speaker was given an speaker ID, and asked to identify themselves with it during the conversation. We manually verified that no sensitive data about the speakers or anyone else was shared during the conversation. In the final step of the data collection, participants signed a consent form with the data protection statement and confidentiality policy, and then received their compensation for their contribution. We believe that the transparent design of our collection and distribution pipeline makes it secure for data subjects.

This project has been approved by the University of Edinburgh, Informatics Ethics Board – Ref. 49776. It has been funded by the Institute for Language, Cognition, and Computation at

[16] `https://www.fiverr.com/` and `https://www.upwork.com/`

[17] We are not publishing video data for this version of the dataset. We will do it in next releases.

[18] The "Stella" passage designed to elicit a wide range of features of different English accents. It was developed and used by the Speech Accent Archive (Weinberger 2015).

the University of Edinburgh.

Transcription

All conversations were manually transcribed by multiple professional transcribers. Each turn was manually segmented and orthographically transcribed, including overlaps between speakers, noise, laughter and hesitations. The transcription company removed the stored data ten days after receiving it.

7.4 Being more transparent: EdAcc Data Statement

7.4.1 Introduction

This document serves as documentation for the DATASET RELEASE 1. The format is adapted from Bender and Friedman (2018).

We explain *why* we compiled the dataset, *what* is in the dataset, *who* contributed to the dataset, *how* it was annotated and compiled and funded. This description applies to the current, fully transcribed dataset and is a draft. A final version will be released with the dataset (and for each future release a new version will be supplied).

7.4.2 Curation Rationale

The key aim of this compilation process was to create a well documented, diverse dataset of English language natural conversations between friends. By *diverse* we mean:

- including second language speakers and first language speakers and their varieties of English
- including speakers of different genders
- including speakers of different ethnic backgrounds
- including speakers of different socioeconomic backgrounds
- including speakers of different ages

By *well-documented* we mean a dataset which contains information about:

- speakers: gender, age, ethnicity, socioeconomic and educational background, speech and language impairments, first language(s), second language(s), language(s) used in different domains, residential history in and outside English-speaking countries
- recordings/conversations: duration of relationship between conversation partners, language(s) they use with one another

By *natural* we mean relaxed conversations between people who know each other.
This corpus is intended to be used for:

- training and testing of automatic speech recognition systems
- researching language variation and change, social interactions and conversations

7.4.3 Language variety

The data set comprises about 35 different varieties of English (including second language varieties). There are 122 speakers who list a total of 51 different first languages. 45 speakers speak English as a first language. The second language speakers of English in our data set have started learning English at different times in their life (recorded as the year they started learning English) and speak it in different domains (recorded in the data set: family, work, friends).

The dataset also contains information about the current place of residence of speakers (in/outwith Scotland), their residential history (places where they have lived for more than three years), and any second languages they speak.

7.4.4 Speaker demographics

Speakers completed a questionnaire about their demographic background. Here we describe some of the most important demographic questions we asked them, report their answers and comment briefly on how the responses affect the overall demographic distribution of the dataset.

Gender

Question: What is your gender?
Question type: *drop-down and open-ended text box*
Response options: Female, Male, Non-binary, Other (text box)
Description: This information will help us create a more representative sample of speakers. Feel free to self-describe if you prefer.

Responses 62 described their gender a *female*, 59 contributors described their gender as *male*, and 1 described their gender as *demiboy*.

26 conversations are between two speakers of the same gender (12 male-male, 14 female-female), and 35 between speakers of different genders.

Comments While male and female contributors are almost equally represented in this dataset, there is an obvious under-representation of other genders. All conversation pairings were self-selected by the contributors.

Age

Question: What's your year of birth? (e.g., 1992)
Question type: *open-ended text box*
Response options: text box
Description: No description

Responses The oldest contributor in this dataset was born in 1956 and the youngest in 2004. The median value is 1995 and mean 1992. In 2022 at the time of data collection, the oldest contributor was 66 years old, the youngest was 18, and the average contributor was 30 years old.

Comments While this dataset has a wider age range of contributors than others, older speakers in particular remain under-represented.

Race/ethnicity

Question: What's your ethnic background?
Question type: *multiple choice (pick one) with open-ended text box*
Response options: Asian, Black, Mixed, White, South Asian, Prefer not to say, Other (text box)
Description: We will use this information to create a more representative and diverse sample of participants.

Responses 57 speakers identified as *White*, 27 as *Asian*, 20 as *Black*, 10 as *South Asian*, 4 as *Mixed*, 1 as *Latina*, 1 as *Jewish, White*, 2 as *Latin American*.

Comments These categories are loosely based on the UK census questions about ethnicity. We acknowledge that this is a highly simplified scale and that any categories of race and ethnicity are socially constructed and might differ across cultural and historical contexts. As racial biases are a particular problem in machine learning systems and linguistic research, we wanted to include some information about contributors' ethnic and racial background.

White speakers form the plurality but not the majority of speakers.

Native language

Question: What is/are your first language(s)? (Languages learned before the age of 5) separate them with commas (e.g., Mandarin,Catalan,French)

Question type: *open-ended text box*
Response options: open-ended text box
Description: No description.

Responses Speakers listed 51 different first languages. 9 languages were spoken as a first language by at least 5 contributors: English (45), Spanish (12), Vietnamese (10), Catalan (7), Italian (6), Hindi (5), Mandarin (5), Bulgarian (5), Urdu (5).

Comments Many speakers listed multiple first languages. We also collected information about the different languages speakers' use in different contexts (work, family, friends), their second languages and how they would describe their accent in English. We would like anyone working with the dataset to keep in mind that descriptions of languages and accents are as much about identity as they are about phonology.

Education

Question: What's your highest level of education?
Question type: *multiple choice (pick one) with open-ended text box*
Response options: No qualifications, National 5/GCSE or equivalent, Highers/A-Levels or NVQ/SVQ 3 or equivalent, Undergraduate degree or NVQ/SVQ 5 or equivalent, Postgraduate degree, Other (text box)
Description: Pick the closest equivalent if you've completed your education outwith Scotland.

Responses 46 contributors hold a *postgraduate degree*, 53 contributors hold an *undergraduate degree or equivalent*, 17 have completed *Highers/A-Levels or equivalent*, 4 have completed *National 5/GCSE or equivalent* and 2 have *no qualifications*.

Comments Since our data collection initially focused on people currently living in Scotland, we opted to use Scottish educational qualifications (or equivalents). Highers/A-Levels are qualifications awarded at successful completion of secondary school and are required for admission to UK universities; most UK students sit GCSEs or National 5 exams at around age 16. We use education and occupation as a proxy for socioeconomic background. We acknowledge that the social, economic and cultural meaning and capital attached to different education levels and jobs may differ drastically depending on the current location of the respondent.

Contributors with post-secondary education, and in particular, postgraduate education are vastly over-represented. This is almost certainly due to our recruiting strategy targeting students and staff at the University of Edinburgh.

Number of different speakers represented

There are 122 different speakers.

Presence of disordered speech

Question: Do you have a speech or hearing impairment? (If yes, please elaborate)
Question type: *open-ended text box*
Response options: open-ended text box
Description:

Responses One contributor reported a minor hearing impairment.

Comments Contributors with speech or hearing impairments were not discouraged from participating in the data collection process. However, we also did not explicitly attempt to recruit them.

7.4.5 Annotator demographic

All manual orthographic transcriptions of the conversations were made by professional transcribers at an external transcription company. The dataset and recording has no further annotations.

To facilitate data processing some responses were slightly standardised: misspellings were corrected and all language names were capitalised. In two instances the language name was changed ("Castellano" to "Spanish", and "Igbo language" to "Igbo"); where respondents named several languages, they were separated by commas to facilitate data processing. This "data tidying" was done by the third author.

7.4.6 Language characteristics

The dataset comprises 61 conversations and one reading passage per speaker conducted and recorded via the video-call platform Zoom. All conversations were conducted between two friends or acquaintances in English. The contributors were provided with detailed instructions on how to record the conversation. They could talk about any topic but were discouraged from disclosing any personal information about themselves. They were explicitly instructed to avoid disclosing information about anyone else and to ensure no one else was audible or visible in the recording. After recording the reading passage and conversation, contributors shared the recording with us.
We provided some optional prompts for the conversation:

- When you were a kid, what kinds of games did you used to play? Is there one you remember well?

- Did you have a favourite book or film or TV show or cartoon as a child? What was it about?
- Did you play a sport or have another hobby as a kid? Do you still do it?
- If you could go on holiday anywhere, where would you go and why?
- Is there a particular food you used to like a child? Do you know how to make it? Do you still like eating it now?
- What's your favourite thing about Edinburgh? What's your least favourite thing about it?

The reading passage was recorded by each speaker at the beginning of the conversation. It is the "Stella passage" designed and used by the Speech Accent Archive (Weinberger 2015) to elicit a wide range of common phonemes and lexical sets in English varieties. The passage is:

"Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station."

7.4.7 Other

Contributor recruitment

Contributors were initially recruited through (local and global) personal and professional networks of the researchers using emails, social media and personal correspondence. As the data compilation progressed, additional contributors were recruited using micro-work platforms[19].

All contributors were compensated with £10 per 15 minutes of audio.

Funding

The participant and transcription costs associated with this dataset were funded by a Small Research Grant by the Institute for Language, Cognition and Computation at the University of Edinburgh.

Licence

[19] `https://www.fiverr.com/` and `https://www.upwork.com/`

7.5 Evaluating ASR systems on a more "realistic" speech dataset

The results presented in this section were in part published in: Ramon Sanabria, Nikolay Bogoychev, Nina Markl, Andrea Carmantini, Klejch Ondřej, and Peter Bell (2023). "The Edinburgh International Accents of English Corpus: Towards the Democratization of English ASR". in: ICASSP 2023*, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10095057*

7.5.1 Experimental setup

In Sanabria et al. (2023), we evaluate three (types of) ASR systems on the EdAcc dataset. While I direct the interested reader to Sanabria et al. (2023) for technical details on the three models, I briefly present some of the findings here.

Data

For the purposes of evaluating different ASR systems, we decided to standardise speaker-provided accent descriptions somewhat. To do this, I simplified some specific accent descriptors (e.g., "Scottish (Fife)" to "Scottish English"), mapped some broader descriptors to a more commonly used descriptor (e.g., "American accent" to "US English"). I also mapped some generic descriptors (eg., "fluent") to a local descriptor based on information about the participant's location history and linguistic background. These labels are, of course, not definitive or necessarily more accurate than the ones provided by the speakers themselves. Rather, we chose them as they are more "compatible" with the existing labels employed by the ASR developers and in the literature.

As discussed in Section 3.4.1, this reveals a central tension in conducting sociolinguistically-informed audits of ASR technologies (and exploration of algorithmic bias in AI more broadly) where the act of *"proving"* discrimination based on socially constructed (and usually oppressive) categories necessarily involves the reification of those same categories. I return to this tension in Section 7.6.

Models

The three models were Wav2vec2.0 (Baevski et al. 2020), Whisper (Radford et al. 2022), and a commercial engine from an (anonymized) well-established company. As discussed in Chapter 5, Whisper is a traditional encoder-decoder-based system trained on unreleased 680,000 hours of multilingual and multitask data collected from the web. To process long utterances, the system segments the audio in 30 second chunks. We experiment with two Wav2vec2.0 encoders pre-trained on different datasets; Wav2vec2.0 (pre-trained on Libripeech), and Robust Wav2vec2.0 (Hsu et al. 2021) (pre-trained on English Mozilla Common Voice, Libri-light, and Switchboard). We test each encoder fine-tuned on Librispeech, Switchboard, AMI (McCowan et al. 2005), and MGB (Bell et al. 2015), and combine them with language models trained in

Model	EdAcc dev	EdAcc test	LS test-clean	LS test-other
W2V2.0	33.4	36.1	2.9	5.6
Company	17.9	18.7	3.8	7.4
Whisper	16.4	19.7	2.7	5.6

Table 7.2: Results of selected systems on EdAcc test, and development set, and Librispeech (LS) test sets.

these same datasets. The commercial system is provided by a well-known provider. As is conventional in speech technology research, we chose to anonymise the company[20]. The model architecture and training data for this model are undisclosed. Notably, this commercial system automatically selects the "best" model from a range of models for different varieties of English based on the phonetic characteristics of the audio clip. We experimented with manually selecting variety-specific models instead but did not find that this meaningfully improved word error rate.

7.5.2 Results

We start by measuring the general complexity of EdAcc by computing WER on development, and test sets. We only report results on one model for each group. We select them based on their performance on the development set[21]. Table 7.2 shows EdAcc's test and development set results on the selected models. We observe that the commercial model, and Whisper outperform Wav2vec2.0 by a large margin, which might be due to being exposed to more, and more diverse English data. Next, we want to see whether EdAcc reveals biases that Librispeech does not capture. We do this by comparing WER between both datasets on all three models. Table 7.2 shows this comparison on Librispeech's `test-clean`, and `test-other` sets. We observe a considerable drop in performance in Wav2vec2.0 when comparing Librispeech and EdAcc results. This gap indicates a (perhaps expected (Likhomanenko et al. 2021)) lack of robustness of the model when exposed to a real world setting.

Comparing performance between varieties in the EdAcc dataset, we decided to focus report WER on first language varieties represented in the dataset: South African English, Ghanaian English, Irish English, Scottish English, US English, Southern British English, Indian English, Jamaican English and Nigerian English. The L2 speakers of English in the dataset vary in terms of their first languages and how long, how and where they have learned English. As a result it is more difficult to compare within and across L2 speaker groups. We leave this for future work.

[20]This is, of course, different from the approach I take in Chapter 3 and Chapter 6 (and the publications these chapters relate to). It also differs from much of the work being done "from a sociolinguistic perspective" discussed in Chapter 3 which generally names the audited models. As Raji and Buolamwini (2019) points out, these kinds of "audits" or evaluations may be considered a breach of terms of service. While I would argue that there is still value in presenting an evaluation of an unnamed provider, I also acknowledge that this anonymisation likely limits the impact of any "audit" on the actual product.

[21]For Wav2vec2.0 we use an encoder pre-trained on Libri-light, MCV, Switchboard, and Fisher, fine-tuned on Librispeech with a LM trained on MGB. For Whisper, we use the `large` model without conditioning on the previously decoded text.

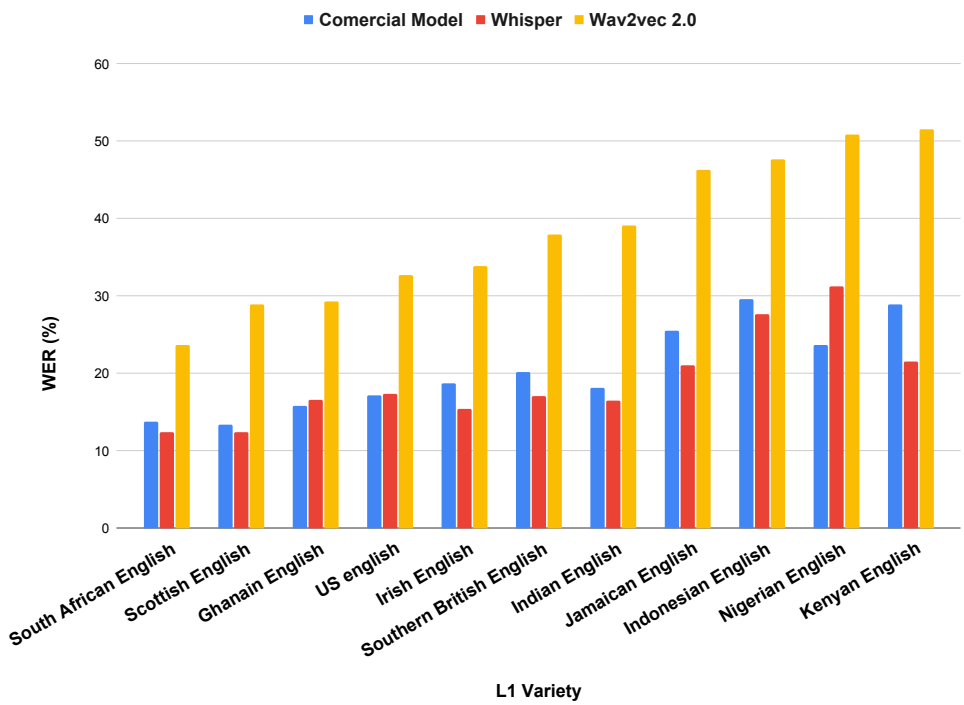


Figure 7.1: Results of selected systems on different English variants from the development and test sets combined.

We calculate word error rates per L1 variety, averaging across the dataset. In Figure 7.1, we see a considerable performance gap between Wav2vec2.0, and other models across all L1 varieties. Consistent with previous work (Meyer et al. 2020; Markl 2022b), we observe considerable performance disparities on specific L1 varieties, such Jamaican, Indonesian, Nigerian, and Kenyan English. These differences are particularly noticeable in Wav2vec2.0 models. [22]

7.5.3 Future Directions

While these initial results indicate predictive bias in some of these models, it is not clear whether these are attributable to phonetic or lexical variation, as we did not control for the lexical content of utterances (but computed error rates across the set). Future work could explore these differences more deeply by considering a range of speaker characteristics (e.g., age, gender, location), and different levels of linguistic variation. We also focussed on first language speakers of English, but as outlined in Section 7.4, this dataset could also be used to evaluate ASR performance for second language speakers. We hope that the dataset provides a starting point for much future work on more useful, "better" ASR systems.

[22] It is interesting to note that overall performance is better for the South African and Scottish English recordings than, for example, the Southern British English ones (on average). Further exploration of the recordings would be required to establish why that is.

7.6 The limits of diversity and inclusion in the face of algorithmic bias

In Hoffmann (2021a), Anna Lauren Hoffmann "seeks to register a discomfort with both emerging ideals of "data ethics" and the critical responses it has garnered" (2021, p. 1). She specifically highlights the "dual imperatives of fairness and inclusion" which "enrol more and more "diverse" persons" (2021, p. 1) into the development and deployment of AI systems, and feminist critiques, such as Data Feminism (D'Ignazio and Klein 2020). Both, she points out, "leave [intact] the dominant logics and structures" of computation, data science, and society more broadly (Hoffmann 2021a, p. 2). In reflecting on the value and cost of "inclusion", I want to echo this discomfort.

In Section 3.4.3, I discussed harms created specifically by *predictive bias* in ASR systems. However, as highlighted in Section 2.4, a narrow focus on "bias" and "fairness" risks overlooking harms inherent to a technology. In the grand scheme of things, many ASR systems do not appear to be as "inherently harmful" as many other AI applications (e.g., the infamous COMPAS "predictive sentencing tool" which became the textbook example of racial bias in algorithmic systems (Angwin et al. 2016)). However, it would be an oversight to only understand ASR tools within their capacity to provide positive new ways of engaging with technologies.

It would also be an oversight to assume that simple "inclusion" of as-yet under-represented groups eliminates all potential harms. In addition to the questions of who gets to design and benefit from datasets raised in Chapter 5, I consider risks inherent to the categorisation of people in datasets, and harms related to "unbiased" ASR tools. In addition to privacy risks, I highlight the risk of considering ASR an "unbiased listener" which is well-positioned to reinforce language ideologies.

7.6.1 Inclusion for what?

Hoffmann (2021b, p. 3546) discusses how "inclusion" is framed as a "discrimination solution" in the form of more diverse datasets and, for example, more inclusive interface designs. Other prominent framings of "inclusion" in the context of sociotechnical systems and algorithmic bias are those of "social commitment" and "affective appeal" – where technology developers are able to use the language of "inclusion" and "data ethics" to not just improve their own image but, ultimately, "[position] data science and technology as [...] the solution to [the violences data technologies produce]" (Hoffmann 2021b, p. 3548). What these discourses foreclose, is the possibility that the "solution" to algorithmic harms might be to refuse to deploy the algorithmic system (Hoffmann 2021b; Cifor et al. 2019; Baumer and Silberman 2011). It is in this context that critiques of "ethics-washing" originate (Washington and Kuo 2020; Siapka 2022; Abdalla and Abdalla 2021; Hampton 2021; Young et al. 2022), which, as I discuss in more detail below, also often point out harms inherent to algorithmic systems, regardless of

"bias".[23]

This narrative of inclusion as a (step towards a) "fix" for predictive bias in ASR was part of the motivation for the Edinburgh International Accents of English Dataset presented above (as we discuss in Sanabria et al. 2023). By creating a dataset more "representative" of the increasingly "diverse" user base of English language ASR, we were hoping to enable better ASR evaluation and speech research. To create a "diverse" or "more inclusive" dataset, (at least) two decisions need to be made: which axes of "difference" matter (i.e., who do we include?) and how do we describe (i.e., construct) that difference in the dataset?

7.6.2 Categories of inclusion

In Section 5.4, I raise the critique that many speech datasets used in ASR development are not just "unrepresentative" but also "under-documented". Without social information about speakers, measuring algorithmic bias is difficult – as Andrus et al. (2021) put it memorably (and perhaps tellingly): "What we can't measure, we can't understand". If we seek to understand adverse impacts of ASR systems (or any other algorithmic system) depending on users' race, ethnicity, gender, age, region, education, disability or social class background, we need to be able to categorise users (see also D'Ignazio and Klein 2020, p. 97). A central tension here is that by focussing on these categories (and differences between them) we risk "[fixing] group identities as stable features of the social landscape" (Benjamin 2019b, p. 147). As Hanna et al. (2020) highlight, the literature on algorithmic fairness tends to treat race as a fixed attribute and uncritically adopt existing "racial schemas" (e.g., from a census). Similarly, the use of binary gender categories in datasets serves to normalise the gender binary (while othering anyone outwith the binary) and the notion of gender as a fixed, perhaps innate attribute (Guyan 2022, p. 54). The same is true for work on "inclusive" or "diverse" datasets. In the context of language research, we also often draw on the category of "native" and "non-native" speakers. As discussed in Chapter 6, this distinction is not only poorly defined but not even necessarily meaningful due to the large amount of variation within each category (Cheng et al. 2021).

In the dataset I present above, we "fill" existing data gaps by including, in principle, any and all people who speak English (and, crucially, are willing and able to record a conversation with a friend or acquaintance for the dataset). While this approach avoids the imposition of "eligibility criteria" commonly used in language research and language technology development (e.g., "only native English speakers", "only speakers without speech or hearing impairments"[24]), the Data Statement (Section 7.4) still reveals significant gaps and imbalances. For example, like many speech datasets, our participants tend to be young (average age: 30) and highly educated (most hold at least an undergraduate degree). To document the dataset, we asked participants to describe themselves and their linguistic experiences. Some of the

[23] As Dunbar-Hester (2020) highlights, open-source and volunteer-driven technology development projects, while not operating within co-operations, have to navigate similar tensions and are often influenced by "corporate" politics of diversity and inclusion.

[24] A criteria often problematically framed as "normal" hearing and speech (Henner and Robinson 2023).

shortcomings of broad categories could be sidestepped by using several questions. For example, in addition to asking participants about their first language, we also asked about the age they started learning English, which language(s) they use most frequently in different domains (home, workplace, with friends) and how they would describe their accent in English. We also tried to capture relevant details about the relationship between each speaker and their interlocutor by asking how long they had known each other and which language(s) they usually use in conversation. These details should enable language (technology) researchers to study language variation and evaluate ASR systems.

7.6.3 Harms beyond bias

It is clear that inclusion of previously excluded groups in technology development (and language research) can be extremely beneficial. As discussed in Section 3.4, predictive bias in ASR alone can cause a lot of different harms to individuals and communities such as a feeling of alienation, loss of economic opportunities, and an entrenchment of existing linguistic hierarchies. As inclusion in datasets can prevent (or mitigate) this predictive bias and its harms, it can improve lives. However, as I have been highlighting throughout this thesis, not all algorithmic harms are related to predictive bias in a narrow sense. Many harms inherent to specific applications of ASR, are, however, still unevenly distributed and affect marginalised groups disproportionately. Here I will discuss concerns around privacy and agency, before turning to the role of language ideologies in ASR systems we might consider "unbiased".

Privacy and agency

A commonly raised concern in the human-computer interaction literature regarding voice technologies is privacy. Privacy risks are, of course, not inherent to all ASR tools but stem from the wider systems they are embedded in and the way they use voice data. This concern is particularly acute for speakers of people who are already marginalised. Rincón et al. (2021), for instance, reports that all 15 trans participants they interviewed voiced concerns about the data privacy in "voice-activated AI" (voice user interfaces). While these misgivings are not unique to trans people, Rincón et al. (2021)'s participants were acutely aware of algorithmic bias and of potential vulnerability, especially with respect to personal information related to their gender identity. They also appear to distrust the developers and any actors they could be compelled to share data with (e.g., police) more than other (non-)users (Lau et al. 2018; Rincón et al. 2021). Just as the (perceived) privacy risks differ between user groups, so does the "cost" of non-use: while for many it is only a trade-off between privacy and convenience, for others, voice user interfaces make their homes more accessible and enable them to live more independently (Lau et al. 2018; Pradhan et al. 2018). In the case of ASR systems embedded in professional settings, including during the hiring process and in the workplace (e.g., to transcribe meetings and presentations), these same concerns about data privacy may matter to individuals, but they

might not have a meaningful way to opt out of having their speech processed automatically. Shelby et al. (2022) use the term "loss of agency" to describe the harm caused to individuals in situations where they are required to use algorithmic systems to access basic services.

Perhaps most controversially, ASR tools can also be used as surveillance tools. In a particularly stark example, Asher-Schapiro and Sherfinski (2021) report that ASR is used by "dozens of county jails and state prisons in seven U.S. states" to monitor phone calls between inmates and friends, family and legal representation. As Asher-Schapiro and Sherfinski (2021) uncover, while the developer LEO technologies and the counties which use it claim that the purpose of the monitoring is to safeguard inmates and staff, conversations about legal representation (in Spanish) and COVID-19 outbreaks have also been flagged. These uses, as well as the stated purpose of these tools to identify mentions of criminal activity (Asher-Schapiro and Sherfinski 2021), have the potential to harm not just incarcerated people but also their associates – whether or not the system produces any errors.[25] In fact, it could be argued that a system with *fewer* errors (and better performance) has the potential to be more, not less, harmful in this application context. Of course, whether surveillance technologies used by the state should exist in the first place and if they do, how well they should work, is a political question, not a technical one (see also Green 2019; Hassein 2017).[26] This leads us back to the critiques of a narrow focus on bias and fairness without consideration of specific harms. It also again highlights the real challenge of assessing "risks" and "harms" of a technology without looking at the wider social and infrastructural context – which raises difficult questions for technologies as versatile as ASR.

As discussed in the Chapter 8, this ambivalence of automatic speech recognition is a cause of significant *discomfort* (for me, at least). Contributing to the building and maintenance of automatic speech recognition which can "understand" an more and more different people in more and more different contexts is not a neutral activity. My colleagues and I have compiled speech datasets with great care for what we consider to be beneficial ends. But it is not lost on me that this kind of data (or indeed, this data) could be used to "improve" surveillance systems employed by a wide range of commercial and institutional actors. These concerns are not unfounded and, as the discussion on privacy above highlights, not unfamiliar to the general population and, in particular, marginalised groups. A full discussion of the moral and ethical complexities of ASR is beyond the scope of the thesis, but I would like to echo Vallor (2016), who calls for the development and nurturing of "technomoral virtues" which can guide us individually and collectively in our use and development of technology.

[25] As Bender and Tatman (2021) were quick to point out, the ASR models (at the time at least) were provided to LEO by Amazon which means that they likely have the same issues with predictive bias regarding African American English Koenecke et al. (2020) highlighted.

[26] It is my view that this use of ASR is fundamentally and irredeemably harmful, as is the entire carceral system it is embedded in.

ASR as a(n unbiased) listener

In Section 3.4, I argued that predictive bias reinforces language ideologies – if an ASR system fails to "understand" a particular variety that signals that the variety is "wrong" to the speakers (and conversely, those it understands are "correct"). In Chapter 4, we explored the potential standardising effect of a system for isiXhosa even in the absence of predictive bias – by choosing to represent the variety using a particular standard and (standard) orthography, existing beliefs about "correct" isiXhosa are reinforced. Since any ASR systems requires developers to make a series of choices about language(s), influence of and impacts on language ideologies are inevitable.

However, some ASR applications are perhaps more likely to have obvious and significant impacts on how speakers perceive themselves and their language(s). In Chapter 6, I identified language teaching and assessment as a popular motivation for researching "accented speech recognition". The potential effect of choosing any one variety as a "standard" to judge speakers' pronunciations against is clear. While this is of course not "new", in the sense that most language teaching involves a particular "target", the fact that students' ability to hit this target is judged by an algorithmic system, is. We see a similar type of linguistic assessment in "automated hiring" or pre-screening, where both content and style of verbal responses to interview questions are automatically analysed and "scored".

In both of these application cases, speakers likely adjust their speech depending on what they think matters to the system[27]. This effect is evident in the proliferation of articles on how to conduct an automatic interview in the popular press, which often include tips like including key words and keeping answers brief and concise (alongside non-verbal cues) – while being "authentic" (e.g., McKeever 2021). As the articles (and the providers of the technology) acknowledge, the prospect of "talking to" an algorithmic system rather than a human recruiter might create new anxieties for applicants (a point also raised by language teachers regarding ASR-based language learning tools (Bashori et al. 2020)). It stands to reason that people who have already had negative experiences with automatic speech recognition are particularly concerned about being understood and might feel particularly pressured to adjust their speech according to what they think the machine "wants to (or can) hear". The limited ability of applicants to repair interactions (or even realise that something was unclear) adds to these anxieties.

Notably, both forms of algorithmic "assessment" (teaching and hiring) are framed as a technical solution to pre-existing discrimination. Bashori et al. (2020) show that second language learners of English in Indonesia are more comfortable practising their pronunciation with an ASR-based system than in the classroom. The students specifically point to worries about teachers and other students mocking them when they make mistakes – and relief that the ASR-

[27] In the context of automated hiring, it is not always clear which aspects of the videos are analysed and how. HireVue for example no longer analyse facial expressions but they do still seem to analyse aspects of the speech such as pauses (Maurer 2021). Other vendors do analyse videos including facial expressions (Drage and Mackereth 2022)

based system does not. Avoiding "unconscious bias" and discrimination in hiring processes is one of the key selling points of algorithmic hiring systems (Drage and Mackereth 2022; Bogen and Rieke 2018; Garr and Jackson 2019). A central assumption of these systems is that it is possible to disentangle candidates' identities (in particular gender and racial identities) from their language use (content and form) in order to use (transcripts or videos) of their verbal responses to "assess" their personalities and skills. From a sociolinguistic perspective, that assumption appears flawed. Scoring applicants based on their language use in terms of attributes like "professional" and perhaps "culture fit"[28], and "authentic" likely produces the same kind of discrimination observed in non-algorithmic hiring, especially since systems are (presumably) trained on successful or "ideal" candidates. Performing the "ideal candidate" is more difficult (and less attractive) for some speakers than others (Drage and Mackereth 2022), and, not all speakers are afforded "authenticity" and "professionalism" at the same time – because notions of "professionalism" are already associated with a particular racial, social class and gender identities (Cushing and Snell 2022; Lippi-Green 2012).

While there is "bias" in these systems, it is much more pernicious and arguably runs deeper than the predictive bias discussed in Chapter 3. Instead it is, as Drage and Mackereth (2022) discuss, a misunderstanding of gender and race, and an encoding of the types of language ideologies and linguistic hierarchies which long pre-date ASR systems. By framing ASR based tools as "neutral" listeners, the role of the wider sociolinguistic context in the formulation of the task, the compilation of appropriate datasets and the development of the system are elided. Zooming out further, beyond this technical or contextual lens, this framing also fails to ask Ehsan et al. (2022)'s "infrastructural question" – why are these systems necessary in the first place? In hiring and teaching, among other contexts, demands of scale and efficiency (e.g., high number of applicants) and existing structural discrimination appear to necessitate algorithmic systems. However, as we can see, these "solutions" fall short – both by (potentially) causing harm to, in particular, marginalised groups and because they, ultimately, do not address these underlying structural concerns.

7.7 Conclusion

On the basis of previous chapters, especially Chapter 5, I explored different ways of building "better" speech technologies. One key approach relates to compiling "better" datasets which can be used to train and test ASR systems. Since predictive bias is rooted in underrepresentation in speech datasets, one seemingly straightforward approach is the compilation of a more representative dataset – covering a larger variety of speakers and speech styles. Coupled with more detailed documentation, these should enable us to spot and mitigate predictive bias and attendant harms in audits and train better, more inclusive technologies. More

[28] A phrase, which, as Drage and Mackereth (2022, p. 89) point out, is "used as a euphemism to justify the gendered and racialized exclusivity of organizations".

thoughtful approaches to compilation of smaller scale datasets (rather than web-scraping of large datasets) further safeguards participants' rights, and allows for a much tighter control on the content in the dataset. Once compiled, these datasets can be used to audit ASR systems, as explored briefly in this chapter. While deeper qualitative error analysis was left for future work, even a simple comparison of error rates between groups highlights the prevalence of predictive bias. Further, the observed differences in ASR performance between conversational speech and read speech emphasises central challenges as systems are continually embedded in new contexts. This approach to dataset compilation and evaluation is illustrated by the Edinburgh Corpus of International Accents of English.

In this chapter, I also wanted to highlight (and return to) key tensions between this drive towards "inclusive" technologies and the way notions of inclusion can be co-opted or distract from the roots of social injustice. As discussed in Chapter 5, data compilation is a complex social process shaped by and embedded in wider power structures, which involves many choices – none of which are "neutral". While we can attempt to justify these decisions, and consider them very carefully, as I do here, good intentions alone may not prevent any and all harms (to data subjects, to users, to language communities). This is particularly true for technologies which aim to "solve" fundamentally social "problems", especially those deeply rooted in oppressive social structures. It is also true for technologies which can be used in a very wide range of ways by a wide range of people and institutions – such as ASR (and, "AI" more broadly).

Chapter 8

Conclusions

8.1 Contexts

My key interest, throughout this thesis, is the *context* of language variation, automatic speech recognition and algorithmic bias. In exploring different aspects of these wider contexts, I have tried to show not just that algorithmic bias exists, but rather *how and why* it affects individuals, communities and language varieties.

In Chapter 3, I draw attention to the *harms* of predictive bias and how they relate to language ideologies and linguistic discrimination in the United Kingdom. These harms relate to the social contexts in which the systems are deployed (e.g., in "high-stakes" contexts like hiring) and used (e.g., as accessibility tools by individuals and institutions). They also relate to the social context of the users (of language and technology) who may already experience discrimination and marginalisation, in part in the form of linguistic discrimination. Predictive bias may both reproduce and further entrench this linguistic discrimination by, one the one hand, causing direct harm to users through poor performance, and (less directly) signalling to them and others that their ways of speaking are less legitimate or "correct".

In Chapter 4, I attend to the *application* of ASR transcription tools in different contexts. I highlight that choosing orthographic representation of speech is always a theoretical and ideological choice. Our work on isiXhosa in Cape Town shows how taking a supposedly "marginal", "low-resource" application context can reveal fundamental flaws in technology development paradigms modelled on an inappropriate "ideal". In the case of ASR, the development pipeline is predicated on monolingual speakers of standard varieties (in standard language cultures). Where speakers draw on multiple languages and feel alienated or uncomfortable with written or spoken standard varieties, these design processes "break down". This should be a reminder that the underlying assumptions of ASR development are not applicable to all, or even most, contexts globally and that close engagement with language communities is key to developing useful technologies.

In Chapter 5, I explore the way *(language) data* is compiled and used in ASR development. Here too, the contexts of the people compiling datasets and their institutions matter. We can

understand some of their choices in the context of language management, which also allows us to understand the decision makers as language policy arbiters. I couple this with a critical analysis of *gaps* in datasets which highlight power structures and alert us to (potential) harms. Importantly, both of these lenses also provide (theoretical) tools to challenge those existing power structures and mitigate the harms. More broadly, I also consider the infrastructural context surrounding speech technology development and suggest that we should perhaps reframe data and models as (public) infrastructure, rather than resources and products.

Building on Chapter 5, I focus on common discourses among researchers and dominant discourses put forward by developers (or companies) about language(s) and linguistic diversity in Chapter 6. The former are not just indicative of language ideologies researchers hold but also impact how and for whom speech technologies are developed. I highlight that the use of imprecise descriptions of speaker groups and varieties are not just reifying those ideologies but also hinder interpretability and reproducibility of speech (technology) research. More broadly, much of the research on "accented" speech also normalises distinctions between "marked" and "unmarked" speech and speakers, and sometimes even explicitly reproduces "deficit discourses" regarding second language speakers. These framings can also be found in some of the discourses deployed in promotion and documentation materials of commercial language technologies, which furthermore rely heavily on notions of "national" languages, linking language varieties to nations, states and territories. Overall, language technologies and the way language and linguistic diversity are conceptualised within them, is an interesting site for the creating and propagation of language ideologies.

Finally, in Chapter 7, I point to ways forward by changing the contexts of data compilation and evaluation through a practical example. I discuss shortcomings of the current approaches of ASR benchmarking and describe the process of compiling a new "diverse" English language ASR evaluation dataset. By focussing on naturalistic conversations between friends from a wide range of linguistic backgrounds, we hope to create a more realistic and ultimately useful benchmark dataset for ASR tasks. Our compilation process prioritises "naturalistic" language use by drawing on insights from sociolinguistic data collection methods and our data documentation process is informed by critical approaches to data science and machine learning datasets and work on relevant documentation. While this dataset aims to mitigate the harms of predictive bias in particular, I also reflect on the real risks and harms associated with compiling and sharing "diverse" speech datasets. Ultimately, it is the contexts in which these systems are deployed which determine who is at risk of harm and how.

8.2 Margins and Centres

Harmful or negative impacts of ASR technologies, like other AI technologies, are disproportionately borne by groups who are already positioned at the margins. Whether it is direct harms of predictive bias as highlighted in Chapter 3, lack of "language resources" pointed out in Chap-

ter 4, non-consensual use of such "language resources" discussed in Chapter 5, the potential hegemonic influence of language technologies explored in Chapter 6 or differential risks of "unbiased" ASR considered in Chapter 7 – it is already marginalised groups who are most likely to be most severely harmed by algorithmic systems when something "goes wrong"[1]. They are further most likely to be impacted by "incidental harm" whether related to the extraction and creation of resources necessary (e.g., data, labour, minerals, energy) or the deployment (e.g., changing work places due to partial automation). I believe that (language) technologies should be designed for and with language communities and minimise harms for marginalised communities in particular. To do this, we (as research communities working on language and/or technology) need to shift our focus towards the margins – or better yet, find ways to make the people, communities and languages at the margins the centre of the work.

In the context of technology design, we can follow and build on the work by the Design Justice Network, whose central principles revolve around centring and empowering affected communities and putting their needs ahead of ours (Costanza-Chock 2020). More broadly, in data science and AI, we can interrogate naturalised categories applied to people (Bowker and Star 2000; D'Ignazio and Klein 2020). In the context of sociolinguistic research, we can draw on critiques of seemingly fundamental categories such as bounded languages (Otheguy et al. 2015; Rosa and Flores 2017; Schneider 2019), and use our work to challenge larger inequities reflected in linguistic discrimination (Craft et al. 2020). In all cases, attending to the wider contexts is key: understanding how, why, for whom, by whom (ASR) technologies are developed and deployed and who interacts with them in what ways is central to glimpsing their "impacts".

This work often entails a certain level of discomfort. It can feel uncomfortable to raise critiques of how things are done, and it is perhaps even more uncomfortable when (well-intentioned) "solutions" are themselves subject of critique. I tried to reflect on this discomfort in Chapter 7 in the context of "inclusion". Hoffmann (2021a) suggests "refusal" as a strategy and value – both in the face of current exclusionary practices in data science and (empty) calls for inclusion. While I am not sure that I fully agree with her on the (lack of) value of notions like "data ethics" and "data feminism", this point should, I think, be highlighted. There are no easy technical or social fixes for social injustice and technologically-facilitated harms, and good intentions, or intentions that seem good to us[2] may not, in fact, lead (only) to good outcomes.

8.3 Future directions

This thesis has picked up many different strands of research. Each strand, prompts, I believe, further inquiry. Firstly, I have laid out a broader theoretical framework drawing on on sociolinguistics and linguistic anthropology which can contribute to our understanding of the

[1] Of course, sometimes this harm is, in fact, intended.

[2] Especially where this "us" consists primarily of researchers, academics, technology developers and other "experts" in already privileged positions.

causes and effects of algorithmic bias in language technologies, as well as the development of language technologies more broadly. In conducting quantitative and qualitative evaluations we can show how existing linguistic discrimination is mirrored in the performance of language technologies. Of particular interest for me is how language technologies and the datasets they rely on interact with processes of standardisation and marginalisation. This, as I have shown, is related to the language ideologies embedded in technology design, which is why thoughtful and critical engagement with the design process is essential.

Regarding predictive bias in commercial ASR systems as discussed in Chapter 3, future work could explore the impacts of it in specific applications (e.g., automatic captioning, interactions with voice user interfaces, "high-stakes" contexts). ASR development with and for speakers of "low-resource" varieties such as isiXhosa could explore different, culturally appropriate ways of representing variation in speech and orthography, building on the findings in Chapter 4. The changing scale of language datasets used in ASR development and their dependence on (and support for) specific infrastructures and development modes which are almost exclusive to large technology companies warrants further attention and critique. Similarly, (commercial) language technologies and the way they are promoted by developers could become a focus for research on language ideologies and language policy, building on Chapter 5 and Chapter 6. Finally, there is (always) a strong need for interdisciplinary and critical perspectives on "better" speech and language technologies and the infrastructures which make them possible.

Bibliography

Abdalla, Mohamed and Moustafa Abdalla (2021). "The Grey Hoodie Project: Big Tobacco, Big Tech, and the Threat on Academic Integrity". In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. ACM. DOI: 10.1145/3461702.3462563.

Abebe, Rediet, Kehinde Aruleba, Abeba Birhane, Sara Kingsley, George Obaido, Sekou L. Remy, and Swathi Sadagopan (2021). "Narratives and Counternarratives on Data Sharing in Africa". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM. DOI: 10.1145/3442188.3445897.

Abid, Abubakar, Maheen Farooqi, and James Zou (2021). "Persistent Anti-Muslim Bias in Large Language Models". In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. ACM. DOI: 10.1145/3461702.3462624.

Adda-Decker, Martine and Lori Lamel (2005). "Do speech recognizers prefer female speakers?" In: *9th European Conference on Speech Communication and Technology* January 2005, pp. 2205–2208.

Agha, Asif (2003). "The social life of cultural value". In: *Language & Communication* 23.3. Words and Beyond: Linguistic and Semiotic Studies of Sociocultural Order, pp. 231–273. DOI: https://doi.org/10.1016/S0271-5309(03)00012-0.

Ahmed, Alex A (2017). "Trans Competent Interaction Design: A Qualitative Study on Voice, Identity, and Technology". In: *Interacting with Computers* 30.1, pp. 53–71. DOI: 10.1093/iwc/iwx018.

Aksënova, Alëna, Daan van Esch, James Flynn, and Pavel Golik (2021). "How Might We Create Better Benchmarks for Speech Recognition?" In: *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*. Association for Computational Linguistics. DOI: 10.18653/v1/2021.bppf-1.4.

Anderson, Michael and Susan Leigh Anderson (2011). *Machine Ethics*. Cambridge University Press.

Anderson, Rindy C, Casey A Klofstad, William J Mayew, and Mohan Venkatachalam (2014). "Vocal fry may undermine the success of young women in the labor market". In: *PloS one* 9.5, e97506–e97506.

Androutsopoulos, Jannis (2016). "Theorizing media, mediation and mediatization". In: *Sociolinguistics*. Cambridge University Press, pp. 282–302. DOI: 10.1017/cbo9781107449787.014.

Andrus, McKane, Elena Spitzer, Jeffrey Brown, and Alice Xiang (2021). "What We Can't Measure, We Can't Understand". In: *Proceedings of the 2021 ACM conference on Fairness, Accountability, and Transparency*. ACM, pp. 249–260. DOI: 10.1145/3442188.3445888.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner (2016). "Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks." In: URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (visited on 05/10/2023).

Ardila, R., 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)*, pp. 4211–4215.

Asher-Schapiro, Avi 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". In: *Thomson Reuters Foundation News*. URL: https://news.trust.org/item/20211115095808-kq7gx/ (visited on 05/10/2023).

Baese-Berk, Melissa M., Drew J. McLaughlin, and Kevin B. McGowan (2020). "Perception of Non-Native Speech". In: *Language and Linguistics Compass* 14.7, e12375. DOI: 10.1111/lnc3.12375. (Visited on 12/01/2021).

Baevski, Alexei, Wei-Ning Hsu, Alexis Conneau, and Michael Auli (2021). "Unsupervised Speech Recognition". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., pp. 27826–27839. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/ea159dc9788ffac311592613b7f71fbb-Paper.pdf.

Baevski, Alexei, Henry Zhou, Abdelrahman Mohamed, and Michael Auli (2020). "Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations". In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS'20. Vancouver, BC, Canada: Curran Associates Inc.

Baker, Carl (2021). "Constituency data: broadband coverage and speeds". In: *House of Commons Library: Data Dashboard*. URL: https://commonslibrary.parliament.uk/constituency-data-broadband-coverage-and-speeds/ (visited on 12/01/2021).

Baratta, Alex (2017). "Accent and linguistic prejudice within british teacher training accent and linguistic prejudice within british teacher training". In: *Journal of Language, Identity & Education* 16.6. Publisher: Routledge, pp. 416–423. DOI: 10.1080/15348458.2017.1359608.

Barker, Jon P., Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe (2017). "The CHiME challenges: Robust speech recognition in everyday environments". In: *New era for robust speech recognition: Exploiting deep learning*. Ed. by Shinji Watanabe, Marc Delcroix, Florian Metze, and John R. Hershey. Cham: Springer International Publishing, pp. 327–344. DOI: 10.1007/978-3-319-64680-0_14.

Barnard, Etienne, Marelie H Davel, Charl van Heerden, Febe De Wet, and Jaco Badenhorst (2014). "The NCHLT speech corpus of the South African languages". In: Workshop Spoken Language Technologies for Under-resourced Languages (SLTU).

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

Barocas, Solon and Andrew D Selbst (2016). "Big Data's Disparate Impact". In: *California Law Review* 104, p. 671. DOI: 10.15779/Z38BG31.

Bashori, Muzakki, Roeland van Hout, Helmer Strik, and Catia Cucchiarini (2020). "Web-based language learning and speaking anxiety". In: *Computer Assisted Language Learning* 35.5-6, pp. 1058–1089. DOI: 10.1080/09588221.2020.1770293.

Baumer, Eric P.S. 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. New York, NY, USA: Association for Computing Machinery, pp. 2271–2274. DOI: 10.1145/1978942.1979275. (Visited on 12/08/2021).

Bell, Peter, Mark JF Gales, Thomas Hain, Jonathan Kilgour, Pierre Lanchantin, Xunying Liu, Andrew McParland, Steve Renals, Oscar Saz, Mirjam Wester, et al. (2015). "The MGB challenge: Evaluating multi-genre broadcast media recognition". In: *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, pp. 687–693.

Bender, Emily M. and Batya Friedman (2018). "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science". In: *Transactions of the Association for Computational Linguistics* 6, pp. 587–604. DOI: 10.1162/tacl_a_00041.

Bender, Emily M., 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. Virtual Event, Canada: Association for Computing Machinery, pp. 610–623. DOI: 10.1145/3442188.3445922.

Bender, Emily M. and Rachael Tatman (2021). *Guest post: AI surveillance in prisons is a terrible idea, both technologically and ethically*. GeekWire. URL: https://www.geekwire.com/2021/guest-post-ai-surveillance-prisons-terrible-idea-technologically-ethically/.

Benjamin, Garfield (2021). "What We Do with Data: A Performative Critique of Data 'Collection'". In: *Internet Policy Review* 10.4. DOI: 10.14763/2021.4.1588. (Visited on 12/08/2021).

Benjamin, Ruha, ed. (2019a). *Captivating Technology*. Duke University Press. DOI: 10.1215/9781478004493.

— (2019b). *Race after technology : abolitionist tools for the New Jim Code*. Newark: Polity Press.

Bennett, Cynthia L. and Os Keyes (2020). “What is the Point of Fairness? Disability, AI and the Complexity of Justice”. In: *SIGACCESS Access. Comput.* 125. DOI: 10.1145/3386296.3386301.

Benzeghiba, M., R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens (2007). “Automatic speech recognition and speech variability: A review”. In: *Speech Communication* 49.10-11, pp. 763–786. DOI: 10.1016/j.specom.2007.02.006.

Bhardwaj, Vivek, Mohamed Tahar Ben Othman, Vinay Kukreja, Youcef Belkhier, Mohit Bajaj, B. Srikanth Goud, Ateeq Ur Rehman, Muhammad Shafiq, and Habib Hamam (2022). “Automatic Speech Recognition (ASR) Systems for Children: A Systematic Literature Review”. In: *Applied Sciences* 12.9, p. 4419. DOI: 10.3390/app12094419.

Bilge, Sirma (2013). “INTERSECTIONALITY UNDONE: Saving Intersectionality from Feminist Intersectionality Studies”. In: *Du Bois Review: Social Science Research on Race* 10.2, pp. 405–424. DOI: 10.1017/S1742058X13000283. (Visited on 02/02/2022).

Bird, Steven (2020). “Decolonising speech and language technology”. In: *Proceedings of the 28th international conference on computational linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 3504–3519. DOI: 10.18653/v1/2020.coling-main.313.

— (2021). “Sparse Transcription”. In: *Computational Linguistics* 46.4, pp. 713–744. DOI: 10.1162/coli_a_00387. (Visited on 03/22/2022).

— (2022). “Local Languages, Third Spaces, and other High-Resource Scenarios”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics. DOI: 10.18653/v1/2022.acl-long.539.

Birhane, Abeba (2020). “Algorithmic Colonization of Africa”. In: *SCRIPTed* 17.2, pp. 389–409. DOI: 10.2966/scrip.170220.389.

— (2021). “Algorithmic injustice: a relational ethics approach”. In: *Patterns* 2.2, p. 100205. DOI: 10.1016/j.patter.2021.100205.

Birhane, Abeba, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao (2022a). “The Values Encoded in Machine Learning Research”. In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM. DOI: 10.1145/3531146.3533083.

Birhane, Abeba, Vinay Uday Prabhu, and Emmanuel Kahembwe (2021). *Multimodal datasets: misogyny, pornography, and malignant stereotypes*. DOI: 10.48550/ARXIV.2110.01963.

Birhane, Abeba, Elayne Ruane, Thomas Laurent, Matthew S. Brown, Johnathan Flowers, Anthony Ventresque, and Christopher L. Dancy (2022b). “The Forgotten Margins of AI Ethics”. In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM. DOI: 10.1145/3531146.3533157.

Blodgett, Su Lin (2021). “Sociolinguistically Driven Approaches for Just Natural Language Processing”. PhD thesis. University of Massachusetts Amherst. DOI: 10.7275/20410631.

Blodgett, Su Lin, 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*. Online: Association for Computational Linguistics, pp. 5454–5476. DOI: 10.18653/v1/2020.acl-main.485.

Blodgett, Su Lin and Brendan O'Connor (2017). *Racial disparity in natural language processing: A case study of social media African-American English*. arXiv: 1707.00061 [cs.CY].

Blommaert, Jan, Helen Kelly-Holmes, Pia Lane, Sirpa Leppänen, Máiréad Moriarty, Sari Pietikäinen, and Arja Piirainen-Marsh (2009). "Media, multilingualism and language policing: an introduction". In: *Language Policy* 8.3, pp. 203–207. DOI: 10.1007/s10993-009-9138-7.

Boersma, Paul (2002). "Praat, a system for doing phonetics by computer". In: *Glot international* 5.

Bogen, Miranda and Aaron Rieke (2018). *Help wanted: An examination of hiring algorithms, equity, and bias.* Upturn. URL: https://www.upturn.org/work/help-wanted/.

Bös, Nadine (2021). *Precire findet einen Käufer*. Frankfurter Allgemeine Zeitung. URL: https://www.faz.net/aktuell/karriere-hochschule/buero-co/umstrittene-technologie-precire-findet-einen-kaeufer-17225356.html.

Bourdieu, Pierre (1977). "The economics of linguistic exchanges". In: *Social Science Information* 16.6, pp. 645–668. DOI: 10.1177/053901847701600601.

Bowker, Geoffrey C. and Susan Leigh Star (2000). *Sorting Things Out: Classification and Its Consequences*. en. The MIT Press. DOI: 10.7551/mitpress/6352.001.0001. (Visited on 02/24/2022).

Bowman, Samuel R. and George Dahl (2021). "What Will it Take to Fix Benchmarking in Natural Language Understanding?" In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 4843–4855. DOI: 10.18653/v1/2021.naacl-main.385.

Bradlow, Ann R., Midam Kim, and Michael Blasingame (2017). "Language-independent talker-specificity in first-language and second-language speech production by bilingual talkers: L1 speaking rate predicts L2 speaking rate". In: *The Journal of the Acoustical Society of America* 141.2, pp. 886–899. DOI: 10.1121/1.4976044.

Branigan, Holly P., Martin J. Pickering, Jamie Pearson, and Janet F. McLean (2010). "Linguistic alignment between people and computers". In: *Journal of Pragmatics* 42.9, pp. 2355–2368. DOI: 10.1016/j.pragma.2009.12.012.

Broussard, Meredith (2019). *Artificial Unintelligence: How Computers Misunderstand the World*. Cambridge, Massachusetts London, England: The MIT Press.

Brunet, Marc-Etienne, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel (2019). "Understanding the Origins of Bias in Word Embeddings". In: *International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, pp. 803–811. URL: http://proceedings.mlr.press/v97/brunet19a.html (visited on 07/07/2021).

Bucholtz, Mary (2000). “The Politics of Transcription”. In: *Journal of Pragmatics* 32.10, pp. 1439–1465. DOI: 10.1016/S0378-2166(99)00094-6. (Visited on 03/22/2022).

— (2007). “Variation in Transcription”. In: *Discourse Studies* 9.6, pp. 784–808. DOI: 10.1177/1461445607082580. (Visited on 03/23/2022).

Bucholtz, Mary and Kira Hall (2005). “Identity and Interaction: A Sociocultural Linguistic Approach”. In: *Discourse Studies* 7.4-5, pp. 585–614. DOI: 10.1177/1461445605054407.

Buolamwini, Joy and Timnit Gebru (2018). “Gender shades: Intersectional accuracy disparities in commercial gender classification”. In: *Proceedings of the 1st conference on fairness, accountability and transparency*. Ed. by Sorelle A. Friedler and Christo Wilson. Vol. 81. Proceedings of machine learning research. New York, NY, USA: PMLR, pp. 77–91. URL: http://proceedings.mlr.press/v81/buolamwini18a.html.

Calder, Jeremy (2019). “From *sissy* to *sickening*: The Indexical Landscape of /s/ in SoMa, San Francisco”. In: *Journal of Linguistic Anthropology* 29.3, pp. 332–358. DOI: 10.1111/jola.12218.

Campbell-Kibler, Kathryn (2009). “The nature of sociolinguistic perception”. In: *Language Variation and Change* 21.1, pp. 135–156. DOI: 10.1017/s0954394509000052.

Canavan, Alexandra, David Graff, and George Zipperlen (1997). *CALLHOME American English Speech*. DOI: 10.35111/EXQ3-X930.

Cardoso, Amanda, Erez Levon, Devyani Sharma, Dominic Watt, and Yang Ye (2019). “Interspeaker variation and the evaluation of British English accents in employment contexts”. In: *Proceedings of the International Congress of Phonetic Sciences*, pp. 1615–1619.

Carmo, Félix do (2020). “‘Time is money’ and the value of translation”. In: *Translation Spaces* 9.1, pp. 35–57. DOI: 10.1075/ts.00020.car.

Cedergren, Henrietta J. and David Sankoff (1974). “Variable Rules: Performance as a Statistical Reflection of Competence”. In: *Language* 50.2, p. 333. DOI: 10.2307/412441.

Chan, May Pik Yu, June Choe, Aini Li, Yiran Chen, Xin Gao, and Nicole Holliday (2022). “Training and typological bias in ASR performance for world Englishes”. In: *Interspeech 2022*. ISCA. DOI: 10.21437/interspeech.2022-10869.

Chao, Monika and Julia R. S. Bursten (2021). “Girl talk: Understanding negative reactions to female vocal fry”. In: *Hypatia-a Journal of Feminist Philosophy* 36.1. Publisher: Cambridge University Press, pp. 42–59. DOI: 10.1017/hyp.2020.55.

Charity, Anne H. (2008). “Linguists as Agents for Social Change”. In: *Language and Linguistics Compass* 2.5, pp. 923–939. DOI: 10.1111/j.1749-818x.2008.00081.x.

Cheng, Lauretta S. P., Danielle Burgess, Natasha Vernooij, Cecilia Solís-Barroso, Ashley McDermott, and Savithry Namboodiripad (2021). “The Problematic Concept of Native Speaker in Psycholinguistics: Replacing Vague and Harmful Terminology With Inclusive and Accurate Measures”. In: *Frontiers in Psychology* 12. DOI: 10.3389/fpsyg.2021.715843.

Cheshire, Jenny. (1982). *Variations in an English dialect : a sociolinguistic study / Jenny Cheshire.* eng. Cambridge studies in linguistics ; 37. Cambridge: Cambridge University Press.

Chiu, C., T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina, N. Jaitly, B. Li, J. Chorowski, and M. Bacchiani (2018). "State-of-the-art speech recognition with sequence-to-sequence models". In: *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4774–4778. DOI: 10.1109/ICASSP.2018.8462105.

Chodroff, Eleanor (2018). *Kaldi Tutorial*. URL: https://eleanorchodroff.com/tutorial/kaldi/index.html.

Choe, June, Yiran Chen, May Pik Yu Chan, Aini Li, Xin Gao, and Nicole Holliday (2022). "Language-specific Effects on Automatic Speech Recognition Errors for World Englishes". In: *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, pp. 7177–7186. URL: https://aclanthology.org/2022.coling-1.628.

Cifor, M., 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". In: URL: https://www.manifestno.com (visited on 12/09/2021).

Clark, Lynn, Helen MacGougan, Jennifer Hay, and Liam Walsh (2016). ""kia ora. this is my earthquake story". Multiple applications of a sociolinguistic corpus". In: *Ampersand* 3, pp. 13–20. DOI: 10.1016/j.amper.2016.01.001.

Coffey, Donavyn (2021). "Māori are trying to save their language from Big Tech". In: *Wired*. URL: https://www.wired.co.uk/article/maori-language-tech (visited on 05/11/2023).

Cooper, Brittney (2016). "Intersectionality". In: *The Oxford Handbook of Feminist Theory*. Ed. by Lisa Disch and Mary Hawkesworth. Vol. 1. Oxford University Press. DOI: 10.1093/oxfordhb/9780199328581.013.20. (Visited on 02/04/2022).

Costanza-Chock, Sasha (2020). *Design Justice*. MIT Press. URL: https://design-justice.pubpub.org/.

Costanza-Chock, Sasha, Inioluwa Deborah Raji, and Joy Buolamwini (2022). "Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem". In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, pp. 1571–1583. DOI: 10.1145/3531146.3533213.

Coto-Solano, Rolando, James N. Stanford, and Sravana K. Reddy (2021). "Advances in Completely Automated Vowel Analysis for Sociophonetics: Using End-to-End Speech Recognition Systems With DARLA". In: *Frontiers in Artificial Intelligence* 4. DOI: 10.3389/frai.2021.662097. (Visited on 04/15/2022).

Coupland, Nikolas and Hywel Bishop (2007). "Ideologised values for British accents". In: *Journal of Sociolinguistics* 11.1, pp. 74–93. DOI: 10.1111/j.1467-9841.2007.00311.x.

Cowan, Benjamin R., Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira (2017). “”What can I help you with?”” In: *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, pp. 1–12. DOI: 10.1145/3098279.3098539.

Cowie, Claire (2007). “The accents of outsourcing: the meanings of ”neutral” in the Indian call centre industry”. In: *World Englishes* 26.3, pp. 316–330. DOI: 10.1111/j.1467-971x.2007.00511.x.

Craft, Justin T., Kelly E. Wright, Rachel Elizabeth Weissler, and Robin M. Queen (2020). “Language and Discrimination: Generating Meaning, Perceiving Identities, and Discriminating Outcomes”. In: *Annual Review of Linguistics* 6.1, pp. 389–407. DOI: 10.1146/annurev-linguistics-011718-011659. (Visited on 11/17/2021).

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

Crawford, Kate and Trevor Paglen (2021). “Excavating AI: the politics of images in machine learning training sets”. In: *AI SOCIETY*. DOI: 10.1007/s00146-021-01162-8.

Crenshaw, Kimberle (1989). “Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics”. In: *University of Chicago Legal Forum* 1989.1. URL: https://chicagounbound.uchicago.edu/uclf/vol1989/iss1/8.

— (1991). “Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color”. In: *Stanford Law Review* 43.6, pp. 1241–1299. DOI: 10.2307/1229039.

Crystal, David (2009). *Dictionary of Linguistics and Phonetics*. Wiley Sons.

Cushing, Ian and Julia Snell (2022). “The (White) Ears of Ofsted: A Raciolinguistic Perspective on the Listening Practices of the Schools Inspectorate”. In: *Language in Society*, pp. 1–24. DOI: 10.1017/S0047404522000094. (Visited on 04/20/2022).

D’Ignazio, Catherine and Lauren F. Klein (2020). *Data Feminism*. The MIT Press. DOI: 10.7551/mitpress/11805.001.0001. (Visited on 06/17/2021).

Dantile, Andiswa Mesatywa (2015). “Language in Public Spaces: Language Choice in Two IsiXhosa Speaking Communities (Langa and Khayelitsha)”. PhD thesis. Stellenbosch, South Africa: University of Stellenbosch.

DARPA (2023). *About DARPA*. URL: https://www.darpa.mil/about-us/about-darpa (visited on 04/12/2023).

David M. Eberhard, Gary F. Simons and Charles D. Fennig, eds. (2021). *Ethnologue: Languages of the World*. 24th Edition. Dallas: SIL International. URL: http://www.ethnologue.com.

Davidson, Lisa (2020). “The versatility of creaky phonation: Segmental, prosodic, and sociolinguistic uses in the world’s languages”. In: *WIREs Cognitive Science* 12.3. DOI: 10.1002/wcs.1547.

Davis, K. H., R. Biddulph, and S. Balashek (1952). "Automatic Recognition of Spoken Digits". In: *The Journal of the Acoustical Society of America* 24.6, pp. 637–642. DOI: 10.1121/1.1906946.

De Russis, Luigi and Fulvio Corno (2019). "On the impact of dysarthric speech on contemporary ASR cloud platforms". In: *Journal of Reliable Intelligent Environments* 5.3, pp. 163–172. DOI: 10.1007/s40860-019-00085-y.

Denton, Emily, Alex Hanna, Razvan Amironesei, Andrew Smart, Hilary Nicole, and Morgan Klaus Scheuerman (2020). "Bringing the People Back In: Contesting Benchmark Machine Learning Datasets". In: *Proceedings of ICML Workshop on Participatory Approaches to Machine Learning, 2020.*

Deumert, Ana (2010). "Imbodela zamakhumsha – Reflections on standardization and destandardization". In: *Multilingua - Journal of Cross-Cultural and Interlanguage Communication* 29.3-4, pp. 243–264. DOI: 10.1515/mult.2010.012.

Deumert, Ana and Nkululeko Mabandla (2017). "Beyond Colonial Linguistics: The Dialectic of Control and Resistance in the Standardization of isiXhosa". In: *Standardizing Minority Languages*. Ed. by Pia Lane, James Costa, and Haley De Korne. Routledge, pp. 200–221. URL: https://library.oapen.org/handle/20.500.12657/24129.

Deumert, Ana and Sibabalwe Oscar Masinyana (2008). "Mobile language choices — The use of English and isiXhosa in text messages (SMS)". In: *English World-Wide* 29.2, pp. 117–147. DOI: 10.1075/eww.29.2.02deu.

Deumert, Ana, Sandrine Mpazayabo, and Miché Thompson (2021). "Cape Town as a multilingual city: Policies, experiences and ideologies". In: *Routledge Handbook of Translation and the City*. Ed. by Tong King Lee. Taylor Francis Group, pp. 448–262.

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

Dev, Sunipa, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang (2021). "Harms of Gender Exclusivity and Challenges in Non-Binary Representation in Language Technologies". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 1968–1994. DOI: 10.18653/v1/2021.emnlp-main.150.

Dias Oliva, Thiago, Dennys Marcelo Antonialli, and Alessandra Gomes (2021). "Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online". In: *Sexuality & Culture* 25.2, pp. 700–732. DOI: 10.1007/s12119-020-09790-w. (Visited on 07/08/2021).

DiChristofano, Alex, Henry Shuster, Shefali Chandra, and Neal Patwari (2022). *Global Performance Disparities Between English-Language Accents in Automatic Speech Recognition*. DOI: 10.48550/ARXIV.2208.01157.

Dixon, John A., Berenice Mahoney, and Roger Cocks (2002). "Accents of Guilt?: Effects of Regional Accent, Race, and Crime Type on Attributions of Guilt". In: *Journal of Language and Social Psychology* 21.2, pp. 162–168. DOI: 10.1177/02627X02021002004.

Dixon, Lucas, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman (2018). "Measuring and Mitigating Unintended Bias in Text Classification". In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '18. New Orleans, LA, USA: Association for Computing Machinery, pp. 67–73. DOI: 10.1145/3278721.3278729.

Dobbe, Roel, Sarah Dean, Thomas Gilbert, and Nitin Kohli (2018). *A Broader View on Bias in Automated Decision-Making: Reflecting on Epistemology and Dynamics*. DOI: 10.48550/ARXIV.1807.00553.

Dodge, Jesse, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner (2021). "Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2021. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 1286–1305. URL: https://aclanthology.org/2021.emnlp-main.98 (visited on 11/26/2021).

Dourish, Paul and Scott D. Mainwaring (2012). "Ubicomp's colonial impulse". In: *Proceedings of the 2012 ACM Conference on Ubiquitous Computing - UbiComp '12*. ACM Press, pp. 133–142. DOI: 10.1145/2370216.2370238.

Drage, Eleanor and Kerry Mackereth (2022). "Does AI Debias Recruitment? Race, Gender, and AI's "Eradication of Difference"". In: *Philosophy Technology* 35.4. DOI: 10.1007/s13347-022-00543-1.

Duffy, Brooke Erin and Colten Meisner (2022). "Platform governance at the margins: Social media creators' experiences with algorithmic (in)visibility". In: *Media, Culture & Society*, p. 016344372211119. DOI: 10.1177/01634437221111923.

Dunbar-Hester, Christina (2020). *Hacking Diversity The Politics of Inclusion in Open Technology Cultures. The Politics of Inclusion in Open Technology Cultures*. Princeton University Press.

Eckert, Penelope (1989). *Jocks and burnouts. social categories and identity in the high school*. Teachers College Press.

— (2008). "Variation and the Indexical Field". In: *Journal of Sociolinguistics* 124, pp. 453–476. DOI: 10.1111/j.1467-9841.2008.00374.x.

— (2012). "Three Waves of Variation Study: The Emergence of Meaning in the Study of Sociolinguistic Variation". In: *Annual Review of Anthropology* (June), pp. 87–100. DOI: 10.1146/annurev-anthro-092611-145828.

Ehsan, Upol, Ranjit Singh, Jacob Metcalf, and Mark Riedl (2022). "The Algorithmic Imprint". In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, pp. 1305–1317. DOI: 10.1145/3531146.3533186.

Eiselen, Roald and Martin Puttkammer (2014). "Developing Text Resources for Ten South African Languages". In: *Proceedings of the Ninth International Conference on Language*

Resources and Evaluation (LREC'14). Reykjavik, Iceland: European Language Resources Association (ELRA), pp. 3698–3703. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/1151_Paper.pdf.

Eubanks, Virginia (2018). *Automating inequality : how high-tech tools profile, police, and punish the poor*. New York, NY: St. Martin's Press.

Fabricius, Anne H. (2018). "Social Change, Linguistic Change and Sociolinguistic Change in Received Pronunciation". In: *Sociolinguistics in England*. Ed. by Natalie Braber and Sandra Jansen. London: Palgrave Macmillan UK, pp. 35–66. DOI: 10.1057/978-1-137-56288-3_3.

Faye, Shon (2021). *The Transgender Issue: An Argument for Justice*. London: Allen Lane. 291 pp.

Field, Anjalie, 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)*. Online: Association for Computational Linguistics, pp. 1905–1925. DOI: 10.18653/v1/2021.acl-long.149.

Flores, Nelson and Jonathan Rosa (2015). "Undoing Appropriateness: Raciolinguistic Ideologies and Language Diversity in Education". In: *Harvard Educational Review* 85.2, pp. 149–171. DOI: 10.17763/0017-8055.85.2.149.

Flores, Nelson L (2019). "Translanguaging into raciolinguistic ideologies: A personal reflection on the legacy of Ofelia García". In: *Journal of Multilingual Education Research* 9.1, pp. 45–59.

Foley, Ben, Josh Arnold, Rolando Coto-Solano, Gautier Durantin, T. Mark Ellison, Daan van Esch, Scott Heath, František Kratochvíl, Zara Maxwell-Smith, David Nash, Ola Olsson, Mark Richards, Nay San, Hywel Stoakes, Nick Thieberger, and Janet Wiles (2018). "Building Speech Recognition Systems for Language Documentation: The CoEDL Endangered Language Pipeline and Inference System". In: *Proceedings of the 6th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU 2018)*. URL: https://www.isca-speech.org/archive_v0/SLTU_2018/pdfs/Ben.pdf.

Foulkes, Paul and Gerald J. Docherty (1999). *Urban Voices: Accent Studies in the British Isles*. English. London, England–New York, NY: Arnold–Oxford UP.

Friedman, Batya and Helen Nissenbaum (1996). "Bias in computer systems". In: *ACM Transactions on Information Systems* 14.3, pp. 330–347. DOI: 10.1145/230538.230561.

Fuller Medina, Nicté (2022). "Data is patrimony: on developing a decolonial model for access and repatriation of sociolinguistic data". In: DOI: 10.7916/ARCHIPELAGOS-N3PB-RX95.

Furui, Sadaoki (2005). "50 Years of Progress in Speech and Speaker Recognition Research". In: *ECTI Transactions on Computer and Information Technology (ECTI-CIT)* 1.2, pp. 64–74. DOI: 10.37936/ecti-cit.200512.51834.

Gabler, Philipp, Bernhard C. Geiger, Barbara Schuppler, and Roman Kern (2023). "Reconsidering Read and Spontaneous Speech: Causal Perspectives on the Generation of Training Data for Automatic Speech Recognition". In: *Information* 14.2, p. 137. DOI: 10.3390/info14020137.

Gal, Susan and Kathryn A. Woolard (1995). "Constructing languages and publics". In: *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA)* 5.2. Publisher: John Benjamins Publishing Company, pp. 129–138. DOI: 10.1075/prag.5.2.01gal.

Gansky, Ben and Sean McDonald (2022). "CounterFAccTual: How FAccT Undermines Its Organizing Principles". In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM. DOI: 10.1145/3531146.3533241.

Garnerin, Mahault, 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*. ACL-GeBNLP-IJCNLP 2021. Online: Association for Computational Linguistics, pp. 86–92. DOI: 10.18653/v1/2021.gebnlp-1.10. (Visited on 11/26/2021).

Garofolo, John S., Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, and Victor Zue (1993). *TIMIT acoustic-phonetic continuous speech corpus LDC93S1*. ISBN: 1-58563-019-5 Place: Philadelphia. DOI: https://doi.org/10.35111/17gk-bn40.

Garr, Stacia Sherman and Careole Jackson (2019). *Diversity Inclusion Technology: The Rise of a Transformative Market*. Tech. rep. Mercer/RedThread Research.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford (2021). "Datasheets for Datasets". In: *Commun. ACM* 64.12, pp. 86–92. DOI: 10.1145/3458723.

Giles, Howard (1970). "Evaluative reactions to accents". In: *Educational Review* 22.3, pp. 211–227. DOI: 10.1080/0013191700220301.

Giles, Howard, Nikolas Coupland, and Justine Coupland (1991). "Accommodation theory: Communication, context, and consequence". In: *Contexts of Accommodation: Developments in Applied Sociolinguistics*. Cambridge University Press.

Gitelman, Lisa, ed. (2013). *"Raw Data" Is an Oxymoron*. The MIT Press. DOI: 10.7551/mitpress/9302.001.0001.

Glass, Ira (2015). "If you don't have anything nice to say, SAY IT IN ALL CAPS: Freedom fries". In: *This American Life* 545. URL: https://www.thisamericanlife.org/545/if-you-dont-have-anything-nice-to-say-say-it-in-all-caps.

Glasser, Abraham (2019). "Automatic speech recognition services: Deaf and hard-of-hearing usability". In: *Extended abstracts of the 2019 CHI conference on human factors in computing systems*. Number of pages: 6 Place: Glasgow, Scotland Uk. New York, NY, USA: Association for Computing Machinery, pp. 1–6. DOI: 10.1145/3290607.3308461.

Godfrey, John J. and Edward Holliman (1993). *Switchboard-1 release 2 LDC97S62*. ISBN: 1-58563-121-3 Place: Philadelphia. DOI: https://doi.org/10.35111/sw3h-rw02.

Goldhahn, Dirk, Thomas Eckart, and Uwe Quasthoff (2012). "Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages". In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA), pp. 759–765. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/327_Paper.pdf.

Goldwater, Sharon, Dan Jurafsky, and Christopher D. Manning (2010). "Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates". In: *Speech Communication* 52.3, pp. 181–200. DOI: https://doi.org/10.1016/j.specom.2009.10.001.

Gonen, Hila and Yoav Goldberg (2019). "Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them". In: *Proceedings of the 2019 Conference of the North*. Association for Computational Linguistics, pp. 609–614. DOI: 10.18653/v1/n19-1061.

Grabe, Esther and Francis Nolan (2002). *The IViE Corpus: English Intonation in the British Isles*. URL: http://www.phon.ox.ac.uk/files/apps/old_IViE/.

Graff, David, Alexandra Canavan, and George Zipperlen (1998). *Switchboard-2 Phase I*. Philadelphia: Linguistic Data Consortium. DOI: 10.35111/c7th-nf28.

Graff, David, Kevin Walker, and Alexandra Canavan (1999). *Switchboard-2 Phase II*. Philadelphia: Linguistic Data Consortium. DOI: 10.35111/5qpg-1r82.

Gray, Mary L. and Siddharth Suri (2019). *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Boston: Houghton Mifflin Harcourt. 1 p.

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

Green, Lisa J. (2002). *African American English: A linguistic introduction*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511800306.

Greene, Daniel (2021). *The Promise of Access: Technology, Inequality, and the Political Economy of Hope*. Cambridge, Massachusetts: The MIT Press. 260 pp.

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

Hall-Lew, Lauren and Zac Boyd (2017). "Phonetic variation and self-recorded data". In: *University of Pennsylvania Working Papers in Linguistics* 23.2, pp. 86–95. URL: https://repository.upenn.edu/pwpl/vol23/iss2/11.

Hall-Lew, Lauren, Claire Cowie, Catherine Lai, Nina Markl, Stephen Joseph McNulty, Shan-Jan Sarah Liu, Clare Llewellyn, Beatrice Alex, Zuzana Elliott, and Anita Klingler (2022). "The Lothian Diary Project: sociolinguistic methods during the COVID-19 lockdown". In: *Linguistics Vanguard* 8.s3, pp. 321–330. DOI: doi:10.1515/lingvan-2021-0053.

Hall-Lew, Lauren and Zuzana Elliott (2015). "Production of FACE and GOAT by Slovak and Czech immigrants in Edinburgh". English. In: *The 18th International Conference of the Phonetic Sciences*. The 18th International Conference of the Phonetic Sciences ; Conference date: 10-08-2015 Through 14-08-2015.

Hampton, Lelia Marie (2021). "Black Feminist Musings on Algorithmic Oppression". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency. Virtual Event Canada: ACM, pp. 1–11. DOI: 10.1145/3442188.3445929. (Visited on 07/07/2021).

Hanna, Alex, 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. Barcelona, Spain: Association for Computing Machinery, pp. 501–512. DOI: 10.1145/3351095.3372826.

Hannun, Awni, 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". In: arXiv: 1412.5567 tex.arxivid: 1412.5567, pp. 1–12. URL: http://arxiv.org/abs/1412.5567.

Haraway, Donna (1988). "Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective". In: *Feminist Studies* 14.3, pp. 575–599. URL: http://www.jstor.org/stable/3178066.

Hargittai, Eszter and Aaron Shaw (2015). "Mind the Skills Gap: The Role of Internet Know-How and Gender in Differentiated Contributions to Wikipedia". In: *Information, Communication & Society* 18.4, pp. 424–442. DOI: 10.1080/1369118X.2014.957711. (Visited on 07/06/2021).

Hassein, Nabil (2017). *Against Black Inclusion in Facial Recognition*. The Digital Talking Drum. URL: https://digitaltalkingdrum.com/2017/08/15/against-black-inclusion-in-facial-recognition/.

Havens, Lucy, 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*. Barcelona, Spain (Online): Association for Computational Linguistics, pp. 107–124. URL: https://www.aclweb.org/anthology/2020.gebnlp-1.10.

Hay, Jennifer and Katie Drager (2010). "Stuffed toys and speech perception". In: *Linguistics* 48.4. DOI: 10.1515/ling.2010.027.

Hazirbas, Caner, Joanna Bitton, Brian Dolhansky, Jacqueline Pan, Albert Gordo, and Cristian Canton Ferrer (2021). "Towards measuring fairness in AI: the Casual Conversations dataset". In: arXiv: http://arxiv.org/abs/2104.02821v1 [cs.CV].

Heller, Monica, ed. (1988). *Codeswitching*. De Gruyter Mouton. DOI: 10.1515/9783110849615.

Heller, Monica and Bonnie S. McElhinny (2017). *Language, Capitalism, Colonialism: Toward a Critical History*. Toronto, Ontario: University of Toronto Press. 310 pp.

Hemamou, Léo, Ghazi Felhi, Vincent Vandenbussche, Jean-Claude Martin, and Chloé Clavel (2019). "HireNet: A hierarchical attention model for the automatic analysis of asynchronous video job interviews". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33.01, pp. 573–581. DOI: 10.1609/aaai.v33i01.3301573.

Hendricks, Lisa Anne, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach (2018). "Women Also Snowboard: Overcoming Bias in Captioning Models". In: *Computer Vision – ECCV 2018*. Springer International Publishing, pp. 793–811. DOI: 10.1007/978-3-030-01219-9_47.

Henner, Jon and Octavian Robinson (2023). "Unsettling Languages, Unruly Bodyminds: A Crip Linguistics Manifesto". In: *Journal of Critical Study of Communication and Disability* 1.1, pp. 7–37. DOI: 10.48516/jcscd_2023vol1iss1.4.

Hill Collins, Patricia (2000 [1990]). *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment*. Second. New York: Routledge.

Himmelmann, Nikolaus P. (2018). "Meeting the transcription challenge". In: *Reflections on Language Documentation 20 Years after Himmelmann 1998. Language Documentation & Conservation Special Publication no. 15.* Ed. by Bradley McDonnell, Andrea L. Berez-Kroeker, and Gary Holton. Honolulu: University of Hawaii Press, pp. 33–40. URL: https://kups.ub.uni-koeln.de/24902/.

Hoffmann, Anna Lauren (2019). "Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse". In: *Information, Communication & Society* 22.7, pp. 900–915. DOI: 10.1080/1369118X.2019.1573912. (Visited on 07/06/2021).

— (2021a). "Even When You Are a Solution You Are a Problem: An Uncomfortable Reflection on Feminist Data Ethics". In: *Global Perspectives* 2.1. DOI: 10.1525/gp.2021.21335.

— (2021b). "Terms of Inclusion: Data, Discourse, Violence". In: *New Media & Society* 23.12, pp. 3539–3556. DOI: 10.1177/1461444820958725. (Visited on 12/09/2021).

Holmquist, Jonathan C. (1985). "Social correlates of a linguistic variable:A study in a Spanish village". In: *Language in Society* 14.2, pp. 191–203. DOI: 10.1017/s004740450001112x.

Hooker, Sara (2021). "Moving beyond "algorithmic bias is a data problem"". In: *Patterns* 2.4. Publisher: Elsevier BV, p. 100241. DOI: 10.1016/j.patter.2021.100241.

Hornberger, Nancy H. and David Cassels Johnson (2007). "Slicing the Onion Ethnographically: Layers and Spaces in Multilingual Language Education Policy and Practice". In: *TESOL Quarterly: A Journal for Teachers of English to Speakers of Other Languages and of Standard English as a Second Dialect* 41.3, pp. 509–532. DOI: 10.1002/j.1545-7249.2007.tb00083.x.

Hosoda, Megumi and Eugene Stone-Romero (2010). "The effects of foreign accents on employment-related decisions". In: *Journal of Managerial Psychology* 25.2. Ed. by Joerg Dietz. Publisher: Emerald, pp. 113–132. DOI: 10.1108/02683941011019339.

Hovy, Dirk, Federico Bianchi, and Tommaso Fornaciari (2020). ""You Sound Just Like Your Father" Commercial Machine Translation Systems Include Stylistic Biases". In: *Proceedings*

of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, pp. 1686–1690. DOI: 10.18653/v1/2020.acl-main.154.

Hovy, Dirk and Anders Søgaard (2015). "Tagging Performance Correlates with Author Age". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics. DOI: 10.3115/v1/p15-2079.

Hsu, Wei-Ning, Anuroop Sriram, Alexei Baevski, Tatiana Likhomanenko, Qiantong Xu, Vineel Pratap, Jacob Kahn, Ann Lee, Ronan Collobert, Gabriel Synnaeve, and Michael Auli (2021). "Robust wav2vec 2.0: Analyzing Domain Shift in Self-Supervised Pre-Training". In: *Proc. Interspeech 2021*, pp. 721–725. DOI: 10.21437/Interspeech.2021-236.

Huang, Xuedong, James Baker, and Raj Reddy (2014). "A historical perspective of speech recognition". In: *Communications of the ACM* 57.1, pp. 94–103. DOI: 10.1145/2500887.

Hudley, Anne H. Charity and Nelson Flores (2022). "Social justice in applied linguistics: Not a conclusion, but a way forward". In: *Annual Review of Applied Linguistics* 42, pp. 144–154. DOI: 10.1017/s0267190522000083.

Hughes, Arthur, Peter Trudgill, and Dominic Watt (2013). *English Accents and Dialects*. Routledge. DOI: 10.4324/9780203784440.

Hult, Francis M and Nancy H Hornberger (2016). "Revisiting Orientations in Language Planning: Problem, Right and Resource as an Analytical Heuristic". In: *The Bilingual Review/La Revista Bilingüe* 33 (3), pp. 30–49. URL: https://repository.upenn.edu/gse_pubs/476.

Hunt, Elle (2021). "Vienna museums open adult-only OnlyFans account to display nudes". In: *The Guardian*. URL: https://www.theguardian.com/artanddesign/2021/oct/16/vienna-museums-open-adult-only-onlyfans-account-to-display-nudes (visited on 05/05/2023).

Hutchinson, Ben, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell (2021). "Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. Virtual Event, Canada: Association for Computing Machinery, pp. 560–575. DOI: 10.1145/3442188.3445918.

Ilbury, Christian (2020). ""Sassy Queens": Stylistic orthographic variation in Twitter and the enregisterment of scpAAVE/scp". In: *Journal of Sociolinguistics* 24.2, pp. 245–264. DOI: 10.1111/josl.12366.

Information Commissioner's Office (2022). *What is special category data?* https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/special-category-data/what-is-special-category-data/.

Inoue, Miyako (2003). "The Listening Subject of Japanese Modernity and His Auditory Double: Citing, Sighting, and Siting the Modern Japanese Woman". In: *Cultural Anthropology* 18.2, pp. 156–193. DOI: 10.1525/can.2003.18.2.156.

Irani, Lilly (2013). "The cultural work of microwork". In: *New Media Society* 17.5, pp. 720–739. DOI: 10.1177/1461444813511926.

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

Jaffe, Alexandra (2000). "Introduction: Non-standard orthography and non-standard speech". In: *Journal of Sociolinguistics* 4.4, pp. 497–513. DOI: 10.1111/1467-9481.00127.

— (2016). "Indexicality, Stance and Fields in Sociolinguistics". In: *Sociolinguistics: Theoretical Debates*. Ed. by Nikolas Coupland. Cambridge: Cambridge University Press, pp. 86–112. DOI: 10.1017/CBO9781107449787.005.

Jiang, May and Christiane Fellbaum (2020). "Interdependencies of gender and race in contextualized word embeddings". In: *Proceedings of the second workshop on gender bias in natural language processing*. Barcelona, Spain (Online): Association for Computational Linguistics, pp. 17–25. URL: https://www.aclweb.org/anthology/2020.gebnlp-1.2.

Joerges, Bernward (1999). "Do Politics Have Artefacts?" In: *Social Studies of Science* 29.3, pp. 411–431. DOI: 10.1177/030631299029003004.

Johnson, David Cassels (2013). *Language policy*. Houndmills, Basingstoke: Palgrave Macmillan.

Johnson, David Cassels and Eric J. Johnson (2014). "Power and agency in language policy appropriation". In: *Language Policy* 14.3, pp. 221–243. DOI: 10.1007/s10993-014-9333-z.

Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury (2020). "The State and Fate of Linguistic Diversity and Inclusion in the NLP World". In: pp. 6282–6293. DOI: 10.18653/v1/2020.acl-main.560.

Juang, B.-H. and L.R. Rabiner (2006). "Speech Recognition, Automatic: History". In: *Encyclopedia of Language & Linguistics*. Elsevier, pp. 806–819. DOI: 10.1016/b0-08-044854-2/00906-8.

Kafle, Sushant and Matt Huenerfauth (2020). "Usability evaluation of captions for people who are deaf or hard of hearing". In: *ACM SIGACCESS Accessibility and Computing* 122, pp. 1–1. DOI: 10.1145/3386410.3386411.

Kasirzadeh, Atoosa (2022). "Algorithmic Fairness and Structural Injustice: Insights from Feminist Political Philosophy". In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, pp. 349–356. DOI: 10.1145/3514094.3534188.

Kendall, Tyler and Charlie Farrington (2021). *The Corpus of Regional African American Language*. DOI: 10.35111/EXQ3-X930.

Keyes, Os, Jevan Hutson, and Meredith Durbin (2019). "A Mulching Proposal". In: *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, pp. 1–11. DOI: 10.1145/3290607.3310433.

Khandelwal, Kartik, Preethi Jyothi, Abhijeet Awasthi, and Sunita Sarawagi (2020). "Black-Box Adaptation of ASR for Accented Speech". In: *Proc. Interspeech 2020*, pp. 1281–1285. DOI: 10.21437/Interspeech.2020-3162.

King, Sharese (2020). "From African American Vernacular English to African American Language: Rethinking the Study of Race and Language in African Americans' Speech". In: *Annual Review of Linguistics* 6.1, pp. 285–300. DOI: 10.1146/annurev-linguistics-011619-030556.

Kirkham, Sam and Emma Moore (2016). "Constructing social meaning in political discourse: Phonetic variation and verb processes in Ed Miliband's speeches". In: *Language in Society* 45.1. Publisher: Cambridge University Press, pp. 87–111. DOI: 10.1017/S0047404515000755.

Koch, Bernard, Emily Denton, Alex Hanna, and Jacob Foster (2021). "Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research". In: *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* 1. URL: https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/3b8a614226a953a8cd9526fca6fe9ba5-Abstract-round2.html (visited on 12/06/2021).

Koenecke, Allison, 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". In: *Proceedings of the National Academy of Sciences* 117.14, pp. 7684–7689. DOI: 10.1073/pnas.1915768117.

Kuhn, Roland, Fineen Davis, Alain Désilets, Eric Joanis, Anna Kazantseva, Rebecca Knowles, Patrick Littell, Delaney Lothian, Aidan Pine, Caroline Running Wolf, Eddie Santos, Darlene Stewart, Gilles Boulianne, Vishwa Gupta, Brian Maracle Owennatékha, Akwiratékha' Martin, Christopher Cox, Marie-Odile Junker, Olivia Sammons, Delasie Torkornoo, Nathan Thanyehténhas Brinklow, Sara Child, Benoît Farley, David Huggins-Daines, Daisy Rosenblum, and Heather Souter (2020). "The Indigenous Languages Technology project at NRC Canada: An empowerment-oriented approach to developing language software". In: *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics. DOI: 10.18653/v1/2020.coling-main.516.

Labov, William (1972). "Some Principles of Linguistic Methodology". In: *Language in Society* 1.1, pp. 97–120. URL: http://www.jstor.org/stable/4166672 (visited on 07/11/2022).

— (1990). "The intersection of sex and social class in the course of linguistic change". In: *Language Variation and Change* 2.2, pp. 205–254. DOI: 10.1017/S0954394500000338.

— (2009). *Social Stratification of English in New York City*. Cambridge University Press.

Larkin, Brian (2013). "The Politics and Poetics of Infrastructure". In: *Annual Review of Anthropology* 42.1, pp. 327–343. DOI: 10.1146/annurev-anthro-092412-155522.

Lau, Josephine, Benjamin Zimmerman, and Florian Schaub (2018). "Alexa, Are You Listening?" In: *Proceedings of the ACM on Human-Computer Interaction* 2.CSCW, pp. 1–31. DOI: 10.1145/3274371.

Laufer, Benjamin, Sameer Jain, A. Feder Cooper, Jon Kleinberg, and Hoda Heidari (2022). "Four Years of FAccT: A Reflexive, Mixed-Methods Analysis of Research Contributions, Shortcomings, and Future Prospects". In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM. DOI: 10.1145/3531146.3533107.

Lawrence, Halcyon M. (2021). "Siri Disciplines". In: *Your Computer Is on Fire*. Ed. by Thomas S. Mullaney, Benjamin Peters, Mar Hicks, and Kavita Philip. The MIT Press, pp. 179–198. DOI: 10.7551/mitpress/10993.003.0013. (Visited on 04/20/2022).

Lee, Dave (2021). "The next Big Tech Battle: Amazon's Bet on Healthcare Begins to Take Shape". In: *Financial Times*. URL: https://www.ft.com/content/fa7ff4c3-4694-4409-9ca6-bfadf3a53a62 (visited on 12/01/2021).

Leemann, Adrian (2016). "Analyzing geospatial variation in articulation rate using crowdsourced speech data". In: *Journal of Linguistic Geography* 4.2, pp. 76–96. DOI: 10.1017/jlg.2016.11.

Leemann, Adrian, Marie-José Kolly, and David Britain (2018). "The English Dialects App: The creation of a crowdsourced dialect corpus". In: *Ampersand* 5, pp. 1–17. DOI: 10.1016/j.amper.2017.11.001.

Lev-Ari, Shiri and Boaz Keysar (2010). "Why don't we believe non-native speakers? The influence of accent on credibility". In: *Journal of Experimental Social Psychology* 46.6, pp. 1093–1096. DOI: https://doi.org/10.1016/j.jesp.2010.05.025.

Levon, Erez, Devyani Sharma, Dominic J. L. Watt, Amanda Cardoso, and Yang Ye (2021). "Accent Bias and Perceptions of Professional Competence in England". In: *Journal of English Linguistics* 49.4, pp. 355–388. DOI: 10.1177/00754242211046316.

Levy, Karen (2022). *Data Driven*. Princeton University Press. DOI: 10.1515/9780691241012.

Li, Tao, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Vivek Srikumar (2020). "UNQOVERing Stereotyping Biases via Underspecified Questions". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. EMNLP-Findings 2020. Online: Association for Computational Linguistics, pp. 3475–3489. DOI: 10.18653/v1/2020.findings-emnlp.311. (Visited on 11/26/2021).

Liberman, Mark Y. (2019). "Corpus Phonetics". In: *Annual Review of Linguistics* 5.1, pp. 91–107. DOI: 10.1146/annurev-linguistics-011516-033830.

Likhomanenko, Tatiana, Qiantong Xu, Vineel Pratap, Paden Tomasello, Jacob Kahn, Gilad Avidov, Ronan Collobert, and Gabriel Synnaeve (2021). "Rethinking Evaluation in ASR: Are Our Models Robust Enough?" In: *Interspeech 2021*. ISCA. DOI: 10.21437/interspeech.2021-1758.

Lin, Guan-Ting, Chan-Jan Hsu, Da-Rong Liu, Hung-Yi Lee, and Yu Tsao (2022). "Analyzing the robustness of unsupervised speech recognition". In: *ICASSP*.

Lippi-Green, Rosina (2012). *English with an accent language, ideology, and discrimination in the United States*. 2nd ed.. London ; New York: Routledge.

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

Loukissas, Yanni Alexander (2019). *All Data Are Local*. The MIT Press. DOI: 10.7551/mitpress/11543.001.0001.

Luccioni, Alexandra and Joseph Viviano (2021). "What's in the Box? An Analysis of Undesirable Content in the Common Crawl Corpus". 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)*. ACL-IJCNLP 2021. Online: Association for Computational Linguistics, pp. 182–189. DOI: 10.18653/v1/2021.acl-short.24. (Visited on 11/26/2021).

Luger, Ewa and Abigail Sellen (2016). ""Like Having a Really Bad PA": The gulf between user expectation and experience of conversational agents". In: *Proceedings of the 2016 CHI conference on human factors in computing systems*. Number of pages: 12. New York, NY, USA: Association for Computing Machinery, pp. 5286–5297. URL: https://doi.org/10.1145/2858036.2858288.

Lukač, Morana (2018). "Grassroots prescriptivism". In: *English Today* 34.4, pp. 5–12. DOI: 10.1017/s0266078418000342.

Macaulay, Ronald K. S. (1977). *Language, social class, and education. a Glasgow study*. Edinburgh University Press, p. 179.

Mack, Sara and Benjamin Munson (2012). "The influence of /s/ quality on ratings of men's sexual orientation: Explicit and implicit measures of the 'gay lisp' stereotype". In: *Journal of Phonetics* 40.1, pp. 198–212. DOI: 10.1016/j.wocn.2011.10.002.

Mackenzie, Laurel and Danielle Turton (2020). "Assessing the Accuracy of Existing Forced Alignment Software on Varieties of British English". In: *Linguistics Vanguard* 6.s1, pp. 1–14. DOI: 10.1515/lingvan-2018-0061.

Mahelona, Keoni, Gianna Leoni, Suzanne Duncan, and Miles Thompson (2023). *OpenAI's Whisper is another case study in Colonisation*. Papa Reo. URL: https://blog.papareo.nz/whisper-is-another-case-study-in-colonisation/ (visited on 03/16/2023).

Markl, Nina (2022a). "(Commercial) Automatic Speech Recognition as a Tool in Sociolinguistic Research". In: *University of Pennsylvania Working Papers in Linguistics* 28.2. URL: https://repository.upenn.edu/pwpl/vol28/iss2/11.

— (2022b). "Language Variation and Algorithmic Bias: Understanding Algorithmic Bias in British English Automatic Speech Recognition". In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. Seoul, Republic of Korea: Association for Computing Machinery, pp. 521–534. DOI: 10.1145/3531146.3533117.

— (2022c). "Mind the data gap(s): Investigating power in speech and language datasets". In: *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Dublin, Ireland: Association for Computational Linguistics, pp. 1–12. URL: https://aclanthology.org/2022.ltedi-1.1.

Markl, Nina and Catherine Lai (2021). "Context-sensitive evaluation of automatic speech recognition: considering user experience & language variation". In: *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*. Online: Association for Computational Linguistics, pp. 34–40. URL: https://aclanthology.org/2021.hcinlp-1.6.

— (2023). "Everyone has an accent". In: *Proc. INTERSPEECH 2023*, pp. 4424–4427. DOI: 10.21437/Interspeech.2023-1847.

Markl, Nina and Stephen Joseph McNulty (2022). "Language technology practitioners as language managers: arbitrating data bias and predictive bias in ASR". In: *Proceedings of the Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 6328–6339. URL: https://aclanthology.org/2022.lrec-1.680.

Markl, Nina, Electra Wallington, Ondrej Klejch, Thomas Reitmaier, Gavin Bailey, Jennifer Pearson, Matt Jones, Simon Robinson, and Peter Bell (2023). "Automatic transcription and (de)standardisation". English. In: *Proceedings - SIGUL 2023, 2nd Annual Meeting of the Special Interest Group on Under-resourced Languages*. URL: https://sigul-2023.ilc.cnr.it/wp-content/uploads/2023/08/9_Paper.pdf.

Martin, Joshua L. (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. Number of pages: 1 Place: Virtual Event, Canada. New York, NY, USA: Association for Computing Machinery, p. 284. DOI: 10.1145/3442188.3445893.

— (2022). "Automatic Speech Recognition Systems, Spoken Corpora, and African American Language: An Examination of Linguistic Bias and Morphsyntactic Features". PhD thesis. University of Florida.

Martin, Joshua L. and Kevin Tang (2020). "Understanding Racial Disparities in Automatic Speech Recognition: The Case of Habitual "be"". In: *Proc. Interspeech 2020*, pp. 626–630. DOI: 10.21437/Interspeech.2020-2893.

Matassi, Mora, Pablo J Boczkowski, and Eugenia Mitchelstein (2019). "Domesticating WhatsApp: Family, friends, work, and study in everyday communication". In: *New Media & Society* 21.10, pp. 2183–2200. DOI: 10.1177/1461444819841890.

Maurer, Roy (2021). *HireVue Discontinues Facial Analysis Screening*. Society for Human Resource Management. URL: https://www.shrm.org/resourcesandtools/hr-topics/talent-acquisition/pages/hirevue-discontinues-facial-analysis-screening.aspx.

McCowan, Iain, Jean Carletta, Wessel Kraaij, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, et al. (2005). "The AMI meeting corpus". In: *International Conference on Methods and Techniques in Behavioral Research*.

McGowan, Kevin B. and Anna M. Babel (2019). "Perceiving isn't believing: Divergence in levels of sociolinguistic awareness". In: *Language in Society* 49.2, pp. 231–256. DOI: 10.1017/s0047404519000782.

McKeever, Vicky (2021). *How to ace a job interview with a robot recruiter*. CNBC. URL: https://www.cnbc.com/2021/04/13/how-to-ace-a-job-interview-with-a-robot-recruiter.html.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan (2021). "A Survey on Bias and Fairness in Machine Learning". In: *ACM Computing Surveys* 54.6, pp. 1–35. DOI: 10.1145/3457607.

Mengesha, Zion, 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". In: *Frontiers in Artificial Intelligence* 4, p. 725911. DOI: 10.3389/frai.2021.725911. (Visited on 01/05/2022).

Metcalf, Jacob, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeleine Clare Elish (2021). "Algorithmic Impact Assessments and Accountability". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, pp. 735–746. DOI: 10.1145/3442188.3445935.

Meyer, Josh, Lindy Rauchenstein, Joshua D. Eisenberg, and Nicholas Howell (2020). "Artie bias corpus: An open dataset for detecting demographic bias in speech applications". In: *Proceedings of the 12th language resources and evaluation conference*. Marseille, France: European Language Resources Association, pp. 6462–6468. URL: https://www.aclweb.org/anthology/2020.lrec-1.796.

Meyerhoff, Miriam and Erik Schleef (2012). "Variation, contact and social indexicality in the acquisition of (ing) by teenage migrants". In: *Journal of Sociolinguistics* 16.3, pp. 398–416. DOI: 10.1111/j.1467-9841.2012.00535.x.

Milroy, James (2001). "Language Ideologies and the Consequences of Standardization". In: *Journal of Sociolinguistics* 5.4, pp. 530–555. DOI: 10.1111/1467-9481.00163. (Visited on 08/16/2021).

Mishra, Taniya, Andrej Ljolje, and Mazin Gilbert (2011). "Predicting human perceived accuracy of ASR systems". In: *Proc. Interspeech 2011*, pp. 1945–1948. DOI: 10.21437/Interspeech.2011-364.

Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru (2019). "Model Cards for Model Reporting". In: *Proceedings of the Conference on Fairness, Accountability, and Trans-*

parency. FAT* '19. Atlanta, GA, USA: Association for Computing Machinery, pp. 220–229. DOI: 10.1145/3287560.3287596.

Mohammad, Saif M. (2017). "Challenges in Sentiment Analysis". In: *A Practical Guide to Sentiment Analysis*. Springer International Publishing, pp. 61–83. DOI: 10.1007/978-3-319-55394-8_4.

Morris, Andrew C, Viktoria Maier, and Phil Green (2004). "From WER and RIL to MER and WIL : improved evaluation measures for connected speech recognition". In: *INTERSPEECH-2004*, pp. 2765–2768.

Morrison, Lennox (2017). *Speech analysis could now land you a promotion*. BBC. URL: https://www.bbc.com/worklife/article/20170108-speech-analysis-could-now-land-you-a-promotion.

Mozilla (2021a). *Mozilla Common Voice Receives $3.4 Million Investment to Democratize and Diversify Voice Tech in East Africa*. Accessed: 24/02/2022. Mozilla. URL: https://foundation.mozilla.org/en/blog/mozilla-common-voice-receives-34-million-investment-to-democratize-and-diversify-voice-tech-in-east-africa/.

— (2021b). *Mozilla partners with NVIDIA to democratize and diversify voice technology*. Accessed: 24/02/2022. Mozilla. URL: https://blog.mozilla.org/en/mozilla/mozilla-partners-with-nvidia-to-democratize-and-diversify-voice-technology/.

Mozilla Common Voice (2022). *How we're making Common Voice even more linguistically inclusive*. Accessed: 24/02/2022. URL: https://foundation.mozilla.org/en/blog/how-we-are-making-common-voice-even-more-linguistically-inclusive/.

Mozilla Common Voice: Community Playbook (n.d.). *Community guidance for languages and variants*. Accessed: 24/02/2022. Mozilla Common Voice: Community Playbook. URL: https://common-voice.github.io/community-playbook/sub_pages/Lang_Variant.html.

Mozilla Common Voice: Discourse (2019). *Feedback needed: Languages and accents strategy*. Accessed: 19/04/2022. URL: https://discourse.mozilla.org/t/feedback-needed-languages-and-accents-strategy/40352.

Muehlmann, Shaylih (2014). "The speech community and beyond". In: *The cambridge handbook of linguistic anthropology*. Ed. by N.J. Enfield, Paul Kockelman, and Jack Sidnell. Cambridge University Press, pp. 577–598. DOI: 10.1017/cbo9781139342872.027.

Mugar, Gabriel and Eric Gordon (2020). "Meaningful inefficiencies: Civic design in an age of digital expediency". In: *Meaningful inefficiencies*. New York: Oxford University Press.

Mugglestone, Lynda (2007). *Talking Proper. The Rise of Accent As Social Symbol*. Oxford University Press, USA.

Nance, Claire, Wilson McLeod, Bernadette O'Rourke, and Stuart Dunmore (2016). "Identity, accent aim, and motivation in second language users: New Scottish Gaelic speakers' use

of phonetic variation". In: *Journal of Sociolinguistics* 20.2, pp. 164–191. DOI: 10.1111/josl.12173.

Nanjo, Hiroaki and Tatsuya Kawahara (2005). "A New ASR Evaluation Measure and Minimum Bayes-Risk Decoding for Open-domain Speech Understanding". In: *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.* IEEE. DOI: 10.1109/icassp.2005.1415298.

Nee, Julia, 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. –, NY, USA: Association for Computing Machinery, pp. 1–9. DOI: 10.1145/3465416.3483301. (Visited on 11/16/2021).

Nekoto, Wilhelmina, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir (2020). "Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. EMNLP-Findings 2020. Online: Association for Computational Linguistics, pp. 2144–2160. DOI: 10.18653/v1/2020.findings-emnlp.195. (Visited on 11/26/2021).

Nguyen, Dong, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong (2016). "Computational Sociolinguistics: A Survey". In: *Computational Linguistics* 42.3, pp. 537–593. DOI: 10.1162/coli_a_00258.

Nguyen, Dong, Laura Rosseel, and Jack Grieve (2021). "On Learning and Representing Social Meaning in NLP: A Sociolinguistic Perspective". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. NAACL-HLT 2021. Online: Association for Computational Linguistics, pp. 603–612. DOI: 10.18653/v1/2021.naacl-main.50. (Visited on 12/06/2021).

Nikulásdóttir, Anna, Jón Guðnason, Anton Karl Ingason, Hrafn Loftsson, Eiríkur Rögnvaldsson, Einar Freyr Sigurðsson, and Steinþór Steingrímsson (2020). "Language Technology Programme for Icelandic 2019-2023". English. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 3414–3422. URL: https://aclanthology.org/2020.lrec-1.418.

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

O'Neil, Cathy (2017). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. London: Penguin Books. 259 pp.

Ochs, Elinor (1979). "Transcription as Theory". In: *Developmental Pragmatics*, pp. 43–72.

Office for National Statistics (2020). *Internet access – households and individuals, Great Britain: 2020*. Tech. rep. Office for National Statistics. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/householdcharacteristics/homeinternetandsocialmediausage/bulletins/internetaccesshouseholdsandindividuals/2020.

Olufemi, Lola (2020). *Feminism, Interrupted: Disrupting Power*. Outspoken. London: Pluto Press. 148 pp.

Onuoha, Mimi (2016). "The Point of Collection". In: *Data & Society*. Accessed: 24/02/2022. URL: https://medium.com/datasociety-points/the-point-of-collection-8ee44ad7c2fa.

Otheguy, Ricardo, Ofelia García, and Wallis Reid (2015). "Clarifying translanguaging and deconstructing named languages: A perspective from linguistics". In: *Applied Linguistics Review* 6.3, pp. 281–307. DOI: 10.1515/applirev-2015-0014.

Pak, Vincent (2021). "(De)coupling race and language: The state listening subject and its rearticulation of antiracism as racism in Singapore". In: *Language in Society*, pp. 1–22. DOI: 10.1017/s0047404521000373.

Panayotov, Vassil, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur (2015). "Librispeech: An ASR corpus based on public domain audio books". In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. DOI: 10.1109/icassp.2015.7178964.

Paullada, Amandalynne (2020). "How does Machine Translation Shift Power?" In: *Resistance AI Workshop, 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.* 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

Paullada, Amandalynne, 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". In: *Patterns* 2.11, p. 100336. DOI: https://doi.org/10.1016/j.patter.2021.100336. (Visited on 11/26/2021).

Pew Research Centre (2021). *Internet/Broadband Fact Sheet*. Tech. rep. Pew Research Centre. URL: https://www.pewresearch.org/internet/fact-sheet/internet-broadband/.

Plank, Barbara, Dirk Hovy, and Anders Søgaard (2014). "Learning part-of-speech taggers with inter-annotator agreement loss". In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics. DOI: 10.3115/v1/e14-1078.

Podesva, Robert J. (2007). “Phonation type as a stylistic variable: The use of falsetto in constructing a persona”. In: *Journal of Sociolinguistics* 11.4, pp. 478–504. DOI: 10.1111/j.1467-9841.2007.00334.x.

Porcheron, Martin, Joel E. Fischer, Stuart Reeves, and Sarah Sharples (2018). “Voice interfaces in everyday life”. In: *Proceedings of the 2018 CHI conference on human factors in computing systems*. CHI ’18. Number of pages: 12. New York, NY, USA: Association for Computing Machinery, pp. 1–12. DOI: 10.1145/3173574.3174214.

Povey, Daniel, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely (2011). “The Kaldi speech recognition toolkit”. In: *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society. URL: https://infoscience.epfl.ch/record/192584?ln=en.

Powles, Julia and Helen Nissenbaum (2018). *The Seductive Diversion of ‘Solving’ Bias in Artificial Intelligence*. URL: https://onezero.medium.com/the-seductive-diversion-of-solving-bias-in-artificial-intelligence-890df5e5ef53.

Pradhan, Alisha, Kanika Mehta, and Leah Findlater (2018). “”Accessibility Came by Accident”: Use of Voice-Controlled Intelligent Personal Assistants by People with Disabilities”. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM. DOI: 10.1145/3173574.3174033.

Pratap, Vineel, Anuroop Sriram, Paden Tomasello, Awni Hannun, Vitaliy Liptchinsky, Gabriel Synnaeve, and Ronan Collobert (2020). *Massively multilingual ASR: 50 languages, 1 model, 1 billion parameters*. arXiv: 2007.03001 [eess.AS].

Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever (2022). *Robust Speech Recognition via Large-Scale Weak Supervision*. DOI: 10.48550/ARXIV.2212.04356.

Raghavan, Manish, Solon Barocas, Jon Kleinberg, and Karen Levy (2020). “Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices”. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* ’20. Barcelona, Spain: Association for Computing Machinery, pp. 469–481. DOI: 10.1145/3351095.3372828.

Raji, Deborah, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada (2021). “AI and the Everything in the Whole Wide World Benchmark”. In: *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* 1. URL: https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/084b6fbb10729ed4da8c3d3f5a3ae7c9-Abstract-round2.html (visited on 12/06/2021).

Raji, Inioluwa Deborah and Joy Buolamwini (2019). “Actionable auditing”. In: *AIES ’19: Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society*. AIES ’19. Number of pages: 7 Place: Honolulu, HI, USA. New York, NY, USA: Association for Computing Machinery, pp. 429–435. DOI: 10.1145/3306618.3314244.

Raji, Inioluwa Deborah, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes (2020). "Closing the AI accountability gap". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, pp. 33–44. DOI: 10.1145/3351095.3372873.

Ramjattan, Vijay A. (2019). "Racializing the problem of and solution to foreign accent in business". In: *Applied Linguistics Review* 13.4, pp. 527–544. DOI: 10.1515/applirev-2019-0058.

— (2022). "Accenting racism in labour migration". In: *Annual Review of Applied Linguistics* 42, pp. 87–92. DOI: 10.1017/s0267190521000143.

Reddy, Sravana and James N. Stanford (2015). "Toward Completely Automated Vowel Extraction: Introducing DARLA". In: *Linguistics Vanguard* 1.1, pp. 15–28. DOI: 10.1515/lingvan-2015-0002.

Reitmaier, Thomas, Electra Wallington, Ondřej Klejch, Nina Markl, Lea-Marie Lam-Yee-Mui, Jennifer Pearson, Matt Jones, Peter Bell, and Simon Robinson (2023). "Situating Automatic Speech Recognition Development within Communities of Under-heard Language Speakers". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23. New York, NY, USA: Association for Computing Machinery. DOI: 10.1145/3544548.3581385.

Reitmaier, Thomas, Electra Wallington, Dani Kalarikalayil Raju, Ondrej Klejch, Jennifer Pearson, Matt Jones, Peter Bell, and Simon Robinson (2022). "Opportunities and Challenges of Automatic Speech Recognition Systems for Low-Resource Language Speakers". In: *CHI Conference on Human Factors in Computing Systems*. ACM, pp. 1–17. DOI: 10.1145/3491102.3517639.

Rhea, Alene, Kelsey Markey, Lauren D'Arinzo, Hilke Schellmann, Mona Sloane, Paul Squires, and Julia Stoyanovich (2022). "Resume Format, LinkedIn URLs and Other Unexpected Influences on AI Personality Prediction in Hiring: Results of an Audit". In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, pp. 572–587. DOI: 10.1145/3514094.3534189.

Ricento, Thomas (2000). "Historical and theoretical perspectives in language policy and planning". In: *Journal of Sociolinguistics* 4.2, pp. 196–213. DOI: 10.1111/1467-9481.00111.

Rickford, John R. (1986). "The need for new approaches to social class analysis in sociolinguistics". In: *Language & Communication* 6.3, pp. 215–221. DOI: 10.1016/0271-5309(86)90024-8.

Rickford, John R. and Sharese King (2016). "Language and Linguistics on Trial: Hearing Rachel Jeantel (and Other Vernacular Speakers) in the Courtroom and Beyond". In: *Language* 92.4, pp. 948–988. DOI: 10.1353/lan.2016.0078. (Visited on 12/09/2021).

Rikap, Cecilia (2023). "Same End By Different Means: Google, Amazon, Microsoft, and Meta's Strategies to Organize Their Frontier AI Innovation Systems". In: *CITYPERC Working Paper*

Series 2023-03. URL: https://researchcentres.city.ac.uk/political-economy#unit=working-papers.

Rincón, Cami, Os Keyes, and Corinne Cath (2021). "Speaking from Experience". In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW1, pp. 1–27. DOI: 10.1145/3449206.

Roessel, Janin, Christiane Schoel, and Dagmar Stahlberg (2020). "Modern Notions of Accentism: Findings, Conceptualizations, and Implications for Interventions and Research on Non-native Accents". In: *Journal of Language and Social Psychology* 39.1, pp. 87–111. DOI: 10.1177/0261927X19884619.

Rosa, Jonathan and Christa Burdick (2016). "Language Ideologies". In: *Oxford Handbook of Language and Society*. Ed. by Ofelia García, Nelson Flores, and Massimiliano Spotti. Oxford University Press, pp. 103–124. DOI: 10.1093/oxfordhb/9780190212896.013.15.

Rosa, Jonathan and Nelson Flores (2017). "Unsettling Race and Language: Toward a Raciolinguistic Perspective". In: *Language in Society* 46.5, pp. 621–647. DOI: 10.1017/S0047404517000562.

Ruiz, Richard (2016 [1984]). "Orientations in Language Planning". In: *Honoring Richard Ruiz and his Work on Language Planning and Bilingual Education*. Multilingual Matters, pp. 13–32. DOI: 10.21832/9781783096701-004.

Sadowski, Jathan (2019). "When data is capital: Datafication, accumulation, and extraction". In: *Big Data & Society* 6.1, p. 205395171882054. DOI: 10.1177/2053951718820549.

Saini, Angela (2019). *Superior: The Return of Race Science*. London: HarperCollins Publishers.

Sakai, T. and S. Doshita (1961). "Phonetic Typewriter". In: *The Journal of the Acoustical Society of America* 33.11, pp. 1664–1664. DOI: 10.1121/1.1936652.

Sanabria, Ramon, Nikolay Bogoychev, Nina Markl, Andrea Carmantini, Klejch Ondřej, and Peter Bell (2023). "The Edinburgh International Accents of English Corpus: Towards the Democratization of English ASR". In: *ICASSP 2023*, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10095057.

Sánchez-Monedero, Javier, Lina Dencik, and Lilian Edwards (2020). "What does it mean to 'solve' the problem of discrimination in hiring?" In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, pp. 458–468. DOI: 10.1145/3351095.3372849.

Sankoff, Gillian and Henrietta Cedergren (1972). "Sociolinguistic research on French in Montréal". In: *Language in Society* 1.1, pp. 173–174. DOI: 10.1017/s004740450000662x.

Sap, Maarten, 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*. Florence, Italy: Association for Computational Linguistics, pp. 1668–1678. DOI: 10.18653/v1/P19-1163.

Savoldi, Beatrice, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi (2021). "Gender Bias in Machine Translation". In: *Transactions of the Association for Computational Linguistics* 9, pp. 845–874. DOI: 10.1162/tacl_a_00401.

Sayers, Dave, Rui Sousa-Silva, Sviatlana Höhn, Lule Ahmedi, Kais Allkivi-Metsoja, Dimitra Anastasiou, Štefan Beňuš, Lynne Bowker, Eliot Bytyçi, Alejandro Catala, Anila Çepani, Rubén Chacón-Beltrán, Sami Dadi, Fisnik Dalipi, Vladimir Despotovic, Agnieszka Doczekalska, Sebastian Drude, Karën Fort, Robert Fuchs, Christian Galinski, Federico Gobbo, Tunga Gungor, Siwen Guo, Klaus Höckner, Petra Lea Láncos, Tomer Libal, Tommi Jantunen, Dewi Jones, Blanka Klimova, Emin Erkan Korkmaz, Mirjam Sepesy Maučec, Miguel Melo, Fanny Meunier, Bettina Migge, Verginica Barbu Mititelu, Aurélie Névéol, Arianna Rossi, Antonio Pareja-Lora, C. Sanchez-Stockhammer, Aysel Şahin, Angela Soltan, Claudia Soria, Sarang Shaikh, Marco Turchi, and Sule Yildirim Yayilgan (2021). *The Dawn of the Human-Machine Era: A forecast of new and emerging language technologies*. Tech. rep. DOI: 10.17011/jyx/reports/20210518/1.

Schiffman, Harold (1996). *Linguistic Culture and Language Policy*. London, United States: Taylor & Francis. URL: http://ebookcentral.proquest.com/lib/ed/detail.action?docID=169268.

Schilling, Natalie (2013). *Sociolinguistic Fieldwork*. Key Topics in Sociolinguistics. Cambridge University Press.

Schlangen, David (2021). "Targeting the Benchmark: On Methodology in Current Natural Language Processing Research". 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)*. Online: Association for Computational Linguistics, pp. 670–674. DOI: 10.18653/v1/2021.acl-short.85.

Schneider, Britta (2019). "Methodological nationalism in Linguistics". In: *Language Sciences* 76, p. 101169. DOI: 10.1016/j.langsci.2018.05.006.

— (2021). "Von Gutenberg zu Alexa". In: *Mensch - Tier - Maschine*. transcript Verlag, pp. 327–346. DOI: 10.1515/9783839453131-014.

Schneider, Britta, Bettina Migge, Doris Dippold, Iker Erdocia, Marie-Theres Fester-Seeger, Sviatlana Höhn, Ledia Kazazi, Mandy Lau, Didem Leblebici, Barbara Lewandowska-Tomaszczyk, Miriam Lind, Morana Lukač, Philipp Meer, Susanne Mohr, Martina Podboj, Agnese Sampietro, Beatrice Savoldi, and Auli Viidalepp (2022). "Changing Language Ideological Concepts in the Human-Machine Era. Questions, Themes and Topics". en. In: DOI: 10.13140/RG.2.2.25867.36649.

Schøning, Signe and Janus Spindler Møller (2009). "Self-recordings as a social activity". In: *Nordic Journal of Linguistics* 32.2, pp. 245–269. DOI: 10.1017/S0332586509990060.

Seaver, Nick (2019). "Knowing Algorithms". In: *digitalSTS*. Princeton University Press, pp. 412–422. DOI: 10.1515/9780691190600-028.

Sebba, Mark (2007). *Spelling and Society*. Cambridge University Press. DOI: 10.1017/cbo9780511486739.

Seymour, William and Max Van Kleek (2021). "Exploring Interactions Between Trust, Anthropomorphism, and Relationship Development in Voice Assistants". In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW2, pp. 1–16. DOI: 10.1145/3479515.

Shah, Chirag and Emily M. Bender (2022). "Situating Search". In: *ACM SIGIR Conference on Human Information Interaction and Retrieval*. ACM. DOI: 10.1145/3498366.3505816.

Shah, Deven Santosh, 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*. Online: Association for Computational Linguistics, pp. 5248–5264. DOI: 10.18653/v1/2020.acl-main.468.

Sharma, Devyani, Erez Levon, and Yang Ye (2022). "50 years of British accent bias: Stability and lifespan change in attitudes to accents". In: *English World-Wide*. DOI: https://doi.org/10.1075/eww.20010.sha.

Shelby, Renee, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk (2022). *Sociotechnical Harms: Scoping a Taxonomy for Harm Reduction*. DOI: 10.48550/ARXIV.2210.05791.

Shohamy, Elana (2006). *Language policy: hidden agendas and new approaches*. London: Routledge.

Siapka, Anastasia (2022). "Towards a Feminist Metaethics of AI". In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, pp. 665–674. DOI: 10.1145/3514094.3534197.

Silverstein, Michael (1996). "Monoglot "Standard" in America: Standardization and Metaphors of Linguistic Hegemony. Contemporary Linguistic Anthropology". In: *The Matrix of Language*. Ed. by Donald Lawrence Brenneis and Ronald K. S. Macaulay. Westview Press, pp. 284–306.

— (2003). "Indexical order and the dialectics of sociolinguistic life". In: *Language & Communication* 23.3-4, pp. 193–229. DOI: 10.1016/s0271-5309(03)00013-2.

Smith, Jennifer and Sophie Holmes-Elliott (2018). "The unstoppable glottal: tracking rapid change in an iconic British variable". In: *English Language and Linguistics* 22.3, pp. 323–355. DOI: 10.1017/S1360674316000459.

Sneller, Betsy (2022). "COVID-era sociolinguistics: introduction to the special issue". In: *Linguistics Vanguard* 8.s3, pp. 303–306. DOI: doi:10.1515/lingvan-2021-0138.

Sneller, Betsy, Suzanne Evans Wagner, and Yongqing Ye (2022). "MI Diaries: ethical and practical challenges". In: *Linguistics Vanguard* 8.s3, pp. 307–319. DOI: doi:10.1515/lingvan-2021-0051.

Spears, Arthur K (1998). "African-American Language Use: Ideology and so-Called Obscenity". In: *African-American English*. Ed. by Guy Bailey, John Baugh, Salikoko S. Mufwene, and John R. Rickford. New York: Routledge, pp. 240–264.

Spolsky, Bernard (2003). *Language policy*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511615245.

Srinivasan, Ramesh (2017). *Whose Global Village?* Kind World Publishing. URL: https://www.ebook.de/de/product/28651032/ramesh_srinivasan_whose_global_village.html.

Stanczak, Karolina and Isabelle Augenstein (2021). "A Survey on Gender Bias in Natural Language Processing". In.

Stanford Linguistics (n.d.). *Voices of California*. http://web.stanford.edu/dept/linguistics/VoCal/.

Stark, Luke and Anna Lauren Hoffmann (2019). "Data Is the New What? Popular Metaphors &amp Professional Ethics in Emerging Data Culture". In: *Journal of Cultural Analytics*. DOI: 10.22148/16.036.

Stolcke, Andreas (2002). "SRILM-an extensible language modeling toolkit". In: *Seventh international conference on spoken language processing*.

Stuart-Smith, Jane (2004). "Scottish english". In: *A handbook of varieties of english*. Ed. by Bernd Kortmann, Kate Burridge, Rajend Mesthrie, Edgar W. Schneider, and Clive Upton. Berlin; Boston: Mouton de Gruyter, pp. 47–67.

Sun, Jiao 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)*. ACL-IJCNLP 2021. Online: Association for Computational Linguistics, pp. 350–360. DOI: 10.18653/v1/2021.acl-short.45. (Visited on 11/26/2021).

Suresh, Harini and John Guttag (2021). "A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle". In: *Equity and Access in Algorithms, Mechanisms, and Optimization*. ACM. DOI: 10.1145/3465416.3483305.

Sutton, Selina Jeanne, Paul Foulkes, David Kirk, and Shaun Lawson (2019). "Voice as a Design Material". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM. DOI: 10.1145/3290605.3300833.

Szymański, Piotr, Piotr Żelasko, Mikolaj Morzy, Adrian Szymczak, Marzena Żyła-Hoppe, Joanna Banaszczak, Lukasz Augustyniak, Jan Mizgajski, and Yishay Carmiel (2020). "WER we are and WER we think we are". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, pp. 3290–3295. DOI: 10.18653/v1/2020.findings-emnlp.295.

Taffel, Sy (2021). "Data and oil: Metaphor, materiality and metabolic rifts". In: *New Media & Society* 0.0, p. 0. DOI: 10.1177/14614448211017887.

Tagliamonte, Sali A. (2011). *Variationist Sociolinguistics Change, Observation, Interpretation. Change, Observation, Interpretation*. Wiley Sons, Incorporated, John, p. 424.

Tagliamonte, Sali A. (2015). *Making Waves*. John Wiley & Sons, Inc. DOI: 10.1002/9781118455494.

Talat, Zeerak, Smarika Lulz, Joachim Bingel, and Isabelle Augenstein (2021). *Disembodied Machine Learning: On the Illusion of Objectivity in NLP*. URL: http://arxiv.org/abs/2101.11974 (visited on 02/14/2022).

Tatman, Rachael (2017). "Gender and dialect bias in YouTube's automatic captions". In: pp. 53–59. DOI: 10.18653/v1/w17-1606.

Tatman, Rachael and Conner Kasten (2017). "Effects of talker dialect, gender & race on accuracy of bing speech and youtube automatic captions". In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 2017-Augus, pp. 934–938. DOI: 10.21437/Interspeech.2017-1746.

Thomas, Rachel L. and David Uminsky (2022). "Reliance on metrics is a fundamental challenge for AI". In: *Patterns* 3.5, p. 100476. DOI: 10.1016/j.patter.2022.100476.

Tiedemann, Jörg (2012). "Parallel Data, Tools and Interfaces in OPUS". In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA), pp. 2214–2218. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf.

Timming, Andrew R (2016). "The effect of foreign accent on employability: a study of the aural dimensions of aesthetic labour in customer-facing and non-customer-facing jobs". In: *Work, Employment and Society* 31.3. Publisher: SAGE Publications, pp. 409–428. DOI: 10.1177/0950017016630260.

Tollefson, James W. (1991). *Planning language, planning inequality : language policy in the community*. London: Longman.

Tomlinson, Barbara (2013). "Colonizing Intersectionality: Replicating Racial Hierarchy in Feminist Academic Arguments". In: *Social Identities* 19.2, pp. 254–272. DOI: 10.1080/13504630.2013.789613. (Visited on 02/02/2022).

Tripodi, Francesca (2021). "Ms. Categorized: Gender, Notability, and Inequality on Wikipedia". In: *New Media & Society*, p. 14614448211023772. DOI: 10.1177/14614448211023772. (Visited on 07/01/2021).

Tripp, Alayo and Benjamin Munson (2021). "Perceiving gender while perceiving language: Integrating psycholinguistics and gender theory". In: *WIREs Cognitive Science* 13.2. DOI: 10.1002/wcs.1583.

Trudgill, Peter (1972). "Sex, covert prestige and linguistic change in the urban British English of Norwich". In: *Language in Society* 1.2, pp. 179–195.

— (2006). "Accent". In: *Encyclopedia of Language Linguistics (Second Edition)*. Ed. by Keith Brown. Second Edition. Oxford: Elsevier, p. 14. DOI: https://doi.org/10.1016/B0-08-044854-2/01506-6.

Tüske, Zoltán, George Saon, Kartik Audhkhasi, and Brian Kingsbury (2020). "Single Headed Attention Based Sequence-to-Sequence Model for State-of-the-Art Results on Switchboard". In: *Proc. Interspeech 2020*, pp. 551–555. DOI: 10.21437/Interspeech.2020-1488.

Tyers, Francis M. and Josh Meyer (2021). "What shall we do with an hour of data? Speech recognition for the un- and under-served languages of Common Voice". In: *CoRR* abs/2105.04674. URL: https://arxiv.org/abs/2105.04674.

Ungless, Eddie L., Björn Ross, and Vaishak Belle (2023). "Potential Pitfalls With Automatic Sentiment Analysis: The Example of Queerphobic Bias". In: *Social Science Computer Review*, p. 089443932311529. DOI: 10.1177/08944393231152946.

Vallor, Shannon (2016). *Technology and the virtues. a philosophical guide to a future worth wanting*. New York, NY: Oxford University Press, p. 309.

van der Westhuizen, Ewald and Thomas Niesler (2018). "A First South African Corpus of Multilingual Code-Switched Soap Opera Speech". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA).

Van Herk, Gerard (2018). *What is sociolinguistics*. Hoboken, NJ: John Wiley & Sons, Inc.

Vergès, Françoise (2021). *A Decolonial Feminism*. Trans. by Ashley J. Bohrer. London: Pluto Press. 110 pp.

Villarreal, Dan, Lynn Clark, Jennifer Hay, and Kevin Watson (2020). "From Categories to Gradience: Auto-coding Sociophonetic Variation with Random Forests". In: *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 11.1. DOI: 10.5334/labphon.216.

Vincent, Nicholas, Hanlin Li, Nicole Tilly, Stevie Chancellor, and Brent Hecht (2021). "Data leverage". In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. FAccT '21. Virtual Event, Canada: ACM, pp. 215–227. DOI: 10.1145/3442188.3445885.

Virpioja, Sami, Peter Smit, Stig-Arne Grönroos, Mikko Kurimo, et al. (2013). *Morfessor 2.0: Python implementation and extensions for Morfessor Baseline*.

Wang, Angelina, Solon Barocas, Kristen Laird, and Hanna Wallach (2022). "Measuring Representational Harms in Image Captioning". In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM. DOI: 10.1145/3531146.3533099.

Washington, Anne L. and Rachel Kuo (2020). "Whose side are ethics codes on?" In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM. DOI: 10.1145/3351095.3372844.

Wassink, Alicia Beckford, Cady Gansen, and Isabel Bartholomew (2022). "Uneven success: automatic speech recognition and ethnicity-related dialects". In: *Speech Communication* 140, pp. 50–70. DOI: 10.1016/j.specom.2022.03.009.

Weidinger, Laura, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha

Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel (2022). "Taxonomy of Risks posed by Language Models". In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM. DOI: 10.1145/3531146.3533088.

Weinberger, Steven (2015). *The speech accent archive*. online. URL: http://accent.gmu.edu.

Weinreich, Uriel, Marvin Herzog, William Labov, and Winfred Lehmann (1968). "Empirical foundations for a theory of language change". In: *Directions for historical linguistics*. Ed. by Yakov Malkiel. University of Texas, pp. 95–188.

Wells, John C. (1982). *Accents of English*. Vol. 2. Cambridge University Press. DOI: 10.1017/CBO9780511611759.

Welsh Language Division (2018). *Welsh language technology action plan*. Tech. rep. Welsh Government. URL: https://gov.wales/welsh-language-technology-and-digital-media-action-plan (visited on 03/31/2021).

— (2020). *Welsh language technology action plan: Progress report 2020*. Tech. rep. Welsh Government. URL: https://gov.wales/welsh-language-technology-action-plan-progress-report-2020 (visited on 03/31/2021).

Wenzel, Kimi, Nitya Devireddy, Cam Davison, and Geoff Kaufman (2023). "Can Voice Assistants Be Microaggressors? Cross-Race Psychological Responses to Failures of Automatic Speech Recognition". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23. Hamburg, Germany: Association for Computing Machinery. DOI: 10.1145/3544548.3581357.

Whittaker, Meredith (2021). "The steep cost of capture". In: *Interactions* 28.6, pp. 50–55. DOI: 10.1145/3488666.

Widder, David Gray, Dawn Nafus, Laura Dabbish, and James Herbsleb (2022). "Limits and Possibilities for "Ethical AI" in Open Source: A Study of Deepfakes". In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM. DOI: 10.1145/3531146.3533779.

Wiley, Terrence (2012). "A Brief History and Assessment of Language Rights in the United States". In: *Language Policies in Education: Critical Issues*. Ed. by W. Tollefson J. 2nd. New York: Routledge, pp. 61–90.

Wilson, Christo, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli (2021). "Building and Auditing Fair Algorithms". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM. DOI: 10.1145/3442188.3445928.

Winner, Langdon (1980). "Do Artifacts Have Politics?" In: *Deadalus* 109.1, pp. 121–136. URL: https://www.jstor.org/stable/20024652.

Woolard, Kathryn (2008). "Language and Identity Choice in Catalonia: The Interplay of Contrasting Ideologies of Linguistic Authority". In: *Lengua, nación e identidad. La regulación*

del plurilingüismo en España y América Latina. Frankfurt am Main: Vervuert/Madrid: Iberoamericana.

Woolard, Kathryn A. and Bambi B. Schieffelin (1994). "Language ideology". In: *Annual Review of Anthropology* 23. Publisher: Annual Reviews, pp. 55–82. URL: http://www.jstor.org/stable/2156006.

Woolgar, Steve and Geoff Cooper (1999). "Do Artefacts Have Ambivalence". In: *Social Studies of Science* 29.3, pp. 433–449. DOI: 10.1177/030631299029003005.

Wright, David and Gavin Brookes (2018). "'This is England, speak English!': a corpus-assisted critical study of language ideologies in the right-leaning British press". In: *Critical Discourse Studies* 16.1, pp. 56–83. DOI: 10.1080/17405904.2018.1511439.

Wu, Yunhan, Daniel Rough, Anna Bleakley, Justin Edwards, Orla Cooney, Philip R. Doyle, Leigh Clark, and Benjamin R. Cowan (2020). "See What I'm Saying? Comparing Intelligent Personal Assistant Use for Native and Non-Native Language Speakers". In: *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM. DOI: 10.1145/3379503.3403563.

Xiong, W., L. Wu, Fil Alleva, Jasha Droppo, Xuedong Huang, and Andreas Stolcke (2017). "The Microsoft 2017 conversational speech recognition system". In: *CoRR* abs/1708.06073. URL: http://arxiv.org/abs/1708.06073.

Young, Meg, Michael Katell, and P.M. Krafft (2022). "Confronting Power and Corporate Capture at the FAccT Conference". In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM. DOI: 10.1145/3531146.3533194.

Young, Victoria and Alex Mihailidis (2010). "Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review". In: *Assistive Technology* 22.2, pp. 99–112. DOI: 10.1080/10400435.2010.483646.

Zhang, Yu, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, Zhong Meng, Ke Hu, Andrew Rosenberg, Rohit Prabhavalkar, Daniel S. Park, Parisa Haghani, Jason Riesa, Ginger Perng, Hagen Soltau, Trevor Strohman, Bhuvana Ramabhadran, Tara Sainath, Pedro Moreno, Chung-Cheng Chiu, Johan Schalkwyk, Françoise Beaufays, and Yonghui Wu (2023). *Google USM: Scaling Automatic Speech Recognition Beyond 100 Languages*. DOI: 10.48550/ARXIV.2303.01037.

Zimman, Lal (2017). "Gender as stylistic bricolage: Transmasculine voices and the relationship between fundamental frequency and /s/". In: *Language in Society* 46.3, pp. 339–370. DOI: 10.1017/s0047404517000070.