April 2021

Sociolinguistically Driven Approaches for Just Natural Language Processing

Su Lin Blodgett
University of Massachusetts Amherst

SOCIOLINGUISTICALLY DRIVEN APPROACHES FOR JUST NATURAL LANGUAGE PROCESSING

A Dissertation Presented

by

SU LIN BLODGETT

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

February 2021

College of Information and Computer Sciences

SOCIOLINGUISTICALLY DRIVEN APPROACHES FOR JUST NATURAL LANGUAGE PROCESSING

A Dissertation Presented

by

SU LIN BLODGETT

Approved as to style and content by:

____

Brendan O'Connor, Chair

____

Mohit Iyyer, Member

____

Hanna Wallach, Member

____

Lisa Green, Member

____

James Allan, Chair of the Faculty
College of Information and Computer Sciences

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Brendan O'Connor, who introduced me to natural language processing and computational sociolinguistics, taught me more than I can say about research, NLP, and writing, and has been endlessly supportive of my many and growing research interests; I consider myself amazingly lucky to have stumbled into his lab. I am also very grateful to my thesis committee members, Mohit Iyyer, Hanna Wallach, and Lisa Green, for their thoughtful feedback and support; in particular, I would like to thank Hanna and Lisa for their invaluable feedback during our collaborations. I would also like to thank Solon Barocas and Hal Daumé III, who along with Hanna were (and continue to be) incredibly generous and intellectually inspiring mentors and collaborators. I am also grateful to Johnny Wei for his thoughtful work during our collaboration.

I have been lucky to pursue this PhD alongside my wonderful labmates Abe Handler and Katie Keith, who over the years have been the source of amazing discussions, inspiration, and support. I've also been fortunate to be part of a fantastic community at UMass, including fellow graduate students Lucas Chaufournier, Kaleigh Clary, Janet Guo, Myungha Jang, Neha Kennard, Kalpesh Krishna, Tiffany Liu, Dirk Ruiken, Rian Shambaugh, Emma Tosch, and Kevin Winner. I'm also grateful to friends outside of UMass who were an important source of research inspiration and support, including Michael Madaio, Ian Stewart, and Zeerak Waseem. Additionally, I'd like to thank the administrative staff at UMass, particularly Leeanne Leclerc, Gwyn Mitchell, Eileen Hamel, and Malaika Ross. I'm also grateful to Sohie Lee, who encouraged me to pursue a PhD. Finally, I am forever indebted to my parents, sister, and Samer for their boundless support, encouragement, and love.

ABSTRACT

SOCIOLINGUISTICALLY DRIVEN APPROACHES FOR JUST NATURAL LANGUAGE PROCESSING

FEBRUARY 2021

SU LIN BLODGETT

B.A., WELLESLEY COLLEGE

M.S., UNIVERSITY OF MASSACHUSETTS AMHERST

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Brendan O'Connor

Natural language processing (NLP) systems are now ubiquitous. Yet the benefits of these language technologies do not accrue evenly to all users, and indeed they can be harmful; NLP systems reproduce stereotypes, prevent speakers of non-standard language varieties from participating fully in public discourse, and re-inscribe historical patterns of linguistic stigmatization and discrimination. How harms arise in NLP systems, and who is harmed by them, can only be understood at the intersection of work on NLP, fairness and justice in machine learning, and the relationships between language and social justice. In this thesis, we propose to address two questions at this intersection: i) How can we conceptualize harms arising from NLP systems?, and ii) How can we quantify such harms?

We propose the following contributions. First, we contribute a model in order to collect the first large dataset of African American Language (AAL)-like social media text. We use the dataset to quantify the performance of two types of NLP systems, identifying disparities in model performance between Mainstream U.S. English (MUSE)- and AAL-like text. Turning to the landscape of bias in NLP more broadly, we then provide a critical survey of the emerging literature on bias in NLP and identify its limitations. Drawing on work across sociology, sociolinguistics, linguistic anthropology, social psychology, and education, we provide an account of the relationships between language and injustice, propose a taxonomy of harms arising from NLP systems grounded in those relationships, and propose a set of guiding research questions for work on bias in NLP. Finally, we adapt the measurement modeling framework from the quantitative social sciences to effectively evaluate approaches for quantifying bias in NLP systems. We conclude with a discussion of recent work on bias through the lens of style in NLP, raising a set of normative questions for future work.

TABLE OF CONTENTS

Page

ACKNOWLEDGMENTS . . . . **iv**

ABSTRACT . . . . **v**

LIST OF TABLES . . . . **xii**

LIST OF FIGURES . . . . **xv**

CHAPTER

1. INTRODUCTION **1**

2. VARIATION ON TWITTER: DEVELOPING A CORPUS OF AAL **4**

2.1 Introduction 4

2.1.1 What is African American Language? 5

2.2 Identifying AAL from demographics 6

2.2.1 AAL on Twitter 6
2.2.2 Twitter and Census data 7
2.2.3 Direct word-demographic analysis 8
2.2.4 Mixed-membership demographic language model 9

2.3 Linguistic validation 12

2.3.1 Lexical-level validation 13
2.3.2 Internet-specific orthography 13
2.3.3 Phonological validation 14
2.3.4 Syntactic validation 14

2.4 A dataset for AAL morphosyntactic analysis 16

2.4.1 Related work 17

2.4.2 Dataset goals . . . 19
2.4.3 Data collection and analysis . . . 22

2.5 Conclusion . . . 26

3. FAIRNESS IN NLP TOOLS: LANGUAGE IDENTIFICATION . . . 27

3.1 Introduction . . . 27
3.2 Related work . . . 28
3.3 Twitter and *langid.py* . . . 28

3.3.1 Adapting language identification for AAL . . . 29

3.4 Commercial systems . . . 31

3.4.1 Experiments . . . 31

3.5 Extended evaluation: World languages . . . 34

3.5.1 Dataset . . . 34
3.5.2 Experiments . . . 36
3.5.3 Results and discussion . . . 37

4. FAIRNESS IN NLP TOOLS: DEPENDENCY PARSING . . . 42

4.1 Introduction . . . 42
4.2 Related Work . . . 42

4.2.1 Parsing for Twitter . . . 42
4.2.2 Parsing for non-standard varieties . . . 43

4.3 Preliminary analysis: Stanford dependencies . . . 44
4.4 Extended analysis: Universal dependencies . . . 45

4.4.1 Dataset . . . 46
4.4.2 AAL annotation . . . 46
4.4.3 Non-AAL Twitter annotation . . . 49
4.4.4 Experiments . . . 53
4.4.5 Results and analysis . . . 57

4.5 Discussion and conclusion . . . 63

5. MEASURING BIAS: A SURVEY OF BIAS IN NLP . . . 64

5.1 Introduction . . . 64
5.2 A critical analysis of bias in NLP . . . 65

5.2.1 Method for gathering papers ... 66
5.2.2 Findings ... 68

5.3 Conclusion ... 74

6. LANGUAGE AND JUSTICE ... 75

6.1 Introduction ... 75
6.2 Social justice ... 75

6.2.1 Social justice and technology ... 77

6.3 Overview of language and justice ... 80

6.3.1 Language *about* ... 81
6.3.2 Language *by* ... 84
6.3.3 Case study: African American Language ... 87
6.3.4 Takeaways ... 91

7. MEASURING BIAS: A TAXONOMY OF HARMS ... 93

7.1 Introduction ... 93
7.2 Towards a taxonomy of representational harms ... 93

7.2.1 Undesirable correlations ... 94
7.2.2 A taxonomy of representational harms ... 96
7.2.3 Other dynamics and effects ... 107
7.2.4 Other taxonomies of harms in NLP ... 110

7.3 Discussion and recommendations ... 114

7.3.1 Language and unjust social arrangements ... 115
7.3.2 Conceptualizations of bias ... 116
7.3.3 Language use in practice ... 118
7.3.4 Case study ... 119

8. MEASURING BIAS: EVALUATING MEASUREMENTS OF BIAS ... 123

8.1 Introduction ... 123
8.2 Measurement modeling ... 123

8.2.1 Evaluating measurement models ... 124
8.2.2 Measurement modeling and bias ... 127

8.3 Measuring bias in embeddings ... 128

8.3.1 Analogy tests 128
8.3.2 Subspace projection 130
8.3.3 WEAT and WEFAT 134
8.3.4 NLI probes 137

8.4 Measuring bias in the world: "Geometry of culture" 142

8.4.1 Embeddings for a language/collective cultural imagination/human semantic memory 143
8.4.2 Embeddings for genres or institutions 146
8.4.3 Evaluating quantitative social science approaches 148

8.5 Measuring bias in NLP systems 152

8.5.1 Machine translation 152
8.5.2 Sentiment analysis 155
8.5.3 Hate speech and toxicity detection 157

8.6 Discussion and recommendations 160

9. STYLE AND BIAS 162

9.1 Introduction 162
9.2 Background 162

9.2.1 Style and attribute transfer 163
9.2.2 Bias through the lens of style 163

9.3 Sociolinguistic conceptions of style and social meaning 163
9.4 Style, identity, and social meaning in NLP 165
9.5 Style and bias 166

10. CONCLUSION 171

10.1 Measuring harms: NLP systems in their sociotechnical context 171
10.2 Challenges and tensions in measurement 173
10.3 Awareness, recourse, participation, and refusal 176

APPENDIX: 179

A.1 Identifying AAL from demographics 179
A.2 A dataset of AAL morphosyntactic analysis 179
A.3 Posterior inference for the ensemble classifier 181
A.4 Preliminary parsing analysis: Stanford dependencies 182
A.5 A survey of bias in NLP 183

A.6 Measuring bias: A taxonomy of harms 186

BIBLIOGRAPHY **189**

LIST OF TABLES

Table **Page**

2.1 Of 31 phonological variant words, top ten by ratio $r_{AA}(w)$. Approximate MUSE equivalents are shown for reference. 15

2.2 AAL syntactic constructions and the ratios of their occurrences in the AA- vs. white-aligned corpora (§2.2.4). 15

2.3 AAL morphosyntactic features and example sentences drawn from Green [2002]. 20

2.4 Statistics for each morphosyntactic feature in our TwitterAAE dataset. Asterisks denote uncertain precision estimates. 23

2.5 Examples of tweets with a word-capitalization pattern indicating stressed *BIN*. 24

2.6 Examples of tweets with *steady* preceding states, rather than activities. 25

3.1 Proportion of tweets in AA- and white-aligned corpora classified as non-English by different classifiers. Twitter-1 excludes Twitter classifications *undefined* and *None*, while Twitter-2 includes them. 28

3.2 Imputed recall of English messages in 2014 messages. For the *General* set these are an approximation; see text. 30

3.3 Percent of the 2,500 tweets in each bin classified as English by each classifier. **Difference** is the difference (disparity on an absolute scale) between the classifier accuracy on the AA-aligned and white-aligned samples. t is the message length for the bin. 33

3.4 Dataset statistics for each language label; the evaluation count refers to the subset used for evaluation. 35

3.5 Dataset statistics for additional labels. 35

3.6 English classification results on not ambiguous, not automatically generated tweets. "+ Demo." indicates an ensemble with the demographics-based English classifier. 38

3.7 Language counts for countries with at least 100 non-ambiguous, non-automatically generated messages (out of 129 countries total), with English recall for the best-performing *langid.py* model and that model in an ensemble classifier. 38

3.8 Sample of tweets which were misclassified as non-English by *langid.py* but correctly classified by the demographic ensemble. @-mentions are shown as @username for display in the table. 39

3.9 Percent of the messages in each bin classified correctly as English or non-English by each classifier; t is the message length for the bin. 41

4.1 Bootstrapped standard errors (from 10,000 message resamplings) are in parentheses; differences are statistically significant ($p < 10^{-6}$ in both cases). 45

4.2 Results from cross-domain training settings. 57

4.3 Results from in-domain training settings (with the *ARK Tagger* setting, see §4.4.4.3). 58

4.4 AA- and white-aligned tweets' labeled attachment scores for UD Treebank-trained models; *Gap* is the WH − AA difference in LAS. 59

4.5 Recall by relation type under UDPipe's *Morpho-Tagger* and *ARK Tagger* settings (+synthetic+embeddings; (3) and (6) from Table 4.4). *Reduction* is the reduction in performance gap (WH - AA) from the *Morpho-Tagger* setting to the *ARK Tagger* setting; bolded numbers indicate a gap reduction of ≥ 10.0. 60

4.6 Examples of AAL syntactic phenomena and occurrence counts in the 250 AA- and 250 white-aligned tweet sets. 61

4.7 Parsing accuracies of syntactic and lexical variations across four UDPipe models. 61

5.1 The NLP tasks covered by the 146 papers. 67

5.2 The categories into which the 146 papers fall. 68

5.3 Examples of the categories into which the papers' motivations and proposed quantitative techniques for measuring or mitigating bias fall. Bold text in the quotes denotes the content that yields our categorizations. 69

5.4 Examples of the categories into which the papers' motivations and proposed quantitative techniques for measuring or mitigating bias fall, continued. 70

A.1 ARK POS tagset [Owoputi et al., 2013]. 180

A.2 Search patterns for each morphosyntactic feature. 181

A.3 The papers of which we are aware on bias in NLP systems. Underlining denotes work not included in the analysis presented in Ch. 5. 186

A.4 Example GPT-3 prompts and generations. 188

LIST OF FIGURES

Figure **Page**

2.1 Mixed-membership model for users (u), messages (m) and tokens (t). Observed variables have a double lined border. 10

2.2 Proportion of tweets containing AAL syntactic constructions by messages' posterior proportion of AA. On the x-axis, 0.1 refers to the decile [0, 0.1). 16

4.1 Example of null copulas and our proposed annotations. 47

4.2 Example of verbal auxiliaries and our proposed annotations. 47

4.3 Example of *bouta* and our proposed annotation. 48

4.4 Example of words separated into multiple tokens and our proposed annotation. 53

7.1 An example prompt provided by OpenAI for using GPT-3. 102

7.2 A tweet describing the languages spoken by Huggingface's team. 103

7.3 A response to Julien's tweet above. 103

7.4 Response from GPT-3. 109

8.1 Tests performed by Caliskan et al. [2017]. Pleasant vs. unpleasant 1, 2, and 3 refer to different sets of pleasant and unpleasant words. 135

8.2 Examples of template sentences, some containing both group-related words (⟨Person⟩) and emotion-related words, and some containing only group-related words [Kiritchenko and Mohammad, 2018]. 155

8.3 Examples of emotional state and emotional situation words [Kiritchenko and Mohammad, 2018]. 156

8.4 Examples of African American- and European American-associated names [Kiritchenko and Mohammad, 2018]. 156

8.5 Examples of female and male-associated noun phrases [Kiritchenko and Mohammad, 2018]. 157

CHAPTER 1

INTRODUCTION

Natural language processing (NLP) systems are now ubiquitous; they translate documents and webpages, fulfill requests to digital assistants, and identify offensive content on social media platforms. Yet the benefits of these language technologies do not accrue evenly to all users, and indeed they can be harmful; NLP systems reproduce stereotypes [Bolukbasi et al., 2016, Caliskan et al., 2017, i.a.], prevent speakers of non-standard language varieties from participating fully in public discourse, and re-inscribe historical patterns of linguistic stigmatization and discrimination [Davidson et al., 2019, Sap et al., 2019].

In this thesis, we argue that how harms arise in NLP systems, and who is harmed by them, can only be understood at the intersection of work on NLP, fairness and justice in machine learning, and the relationships between language and social justice. We propose to address two questions at this intersection:

1. How can we conceptualize harms arising from NLP systems?
2. How can we quantify and mitigate such harms?

We begin with the second question, focusing on the performance of NLP systems on one particular language variety, African American Language (AAL). Although spoken by millions of people across the United States, the variety is stigmatized and little written AAL (or other non-standard varieties) has historically been available; as a consequence, NLP systems have been developed from datasets of largely Mainstream U.S. English (MUSE). The rise of social media, however, both enables and necessitates the collection of large-scale AAL corpora: the former because users are writing in AAL at a large scale for the first time, and the latter because NLP systems are increasingly employed on social media data, and risk amplifying existing injustices of stigmatization, mischaracterization, or erasure of AAL and its speakers if the performance of systems on AAL on social media is not examined.

In Chapter 2, we propose a mixed membership model to identify tweets containing African American Language (AAL)-like language using Census demographics, and show that the language in the resulting corpus follows well-known AAL linguistic phenomena. We further propose a dataset of tweets containing AAL morphosyntactic features in order to enable sociolinguistic analysis of morphosyntactic variation in AAL. In Chapters 3 and 4 we demonstrate performance disparities between tweets with AAL-like language and MUSE-like language for widely used language identification and dependency parsing systems, respectively. In Chapter 3 we additionally contribute an ensemble classifier based on our mixed membership model that reduces these disparities and examine the performance of language identification systems on social media English globally by collecting and annotating a new dataset, finding that our ensemble classifier also aids in correctly classifying English tweets from outside the U.S. In Chapter 4 we further develop Universal Dependencies annotation guidelines for several AAL and Twitter syntactic phenomena, evaluate several strategies for mitigating the performance disparities we identified, and examine parsing performance on particular AAL phonological and syntactic phenomena.

In the remainder of the thesis, we turn to the broader landscape of bias in NLP and develop general frameworks for understanding and evaluating approaches for quantifying harms. In Chapter 5 we provide a critical survey of the emerging literature in the space of bias in NLP. We find that although it has laid vital groundwork for identifying bias in NLP systems, much of it provides vague and inconsistent motivations, lacks any normative reasoning for why system behaviors described as bias are harmful, and fails to engage with relevant literature outside NLP to ground its normative concerns. Moreover, the overuse of the word "bias" obscures important differences between how bias is conceptualized and operationalized between papers.

In Chapter 6, we provide a normative foundation for reasoning about harms arising from NLP systems that we have shown is largely absent from the current literature. Drawing on work across sociology, sociolinguistics, linguistic anthropology, social psychology, and education, we provide this foundation through an account of the relationships between language and injustice. This foundation is critical for a number of reasons: first, it grounds our normative concerns about what NLP practices and system behaviors are harmful in the realities of current unjust social arrangements. Second, the existing pathbreaking literature on language and justice, and social justice more broadly, illuminates the concrete mechanisms by which these social arrangements are produced and maintained, which helps to guide our analyses of NLP systems. Third, an understanding of the space of harms that can arise from NLP systems is necessary in order to effectively quantify harms, and to evaluate our quantification and mitigation approaches.

In Chapter 7, we argue this space of harms is in fact much larger than what has been examined in the literature on bias in NLP, and we propose a taxonomy of *representational harms* grounded in these relationships between language and injustice. We conclude the chapter by proposing a re-orientation of work on bias in NLP towards these relationships and offer guiding research questions focusing on how NLP systems and practices reproduce them. In Chapter 8 we adapt the framework of measurement modeling from the quantitative social sciences to examining bias in NLP systems. This framework disentangles *theoretical constructs*—what it is we wish to measure—from *measurements*—the observable properties, or proxies, proposed to measure them. In this chapter, we apply this framework to evaluate current approaches to quantifying bias in NLP systems by reframing them as *measurement models*. For a range of these approaches, we identify the measurement model and the construct(s) implicitly under measurement, and interrogate the (mis)matches between construct and operationalization. We also examine work in the quantitative social sciences that uses bias-in-embeddings approaches, and analyze how the measurement models implicitly provided by these approaches differ from the superficially similar ones provided by NLP practitioners quantifying bias in embeddings.

Finally, in Chapter 9 we examine recent work on style transfer as well as work that addresses bias in NLP through the lens of style; we draw on the analyses developed in the previous chapter to reframe this work and raise a number of normative questions. We conclude (Chapter 10) with a set of open challenges for the development of more just NLP systems.

CHAPTER 2

VARIATION ON TWITTER: DEVELOPING A CORPUS OF AAL

2.1 Introduction

Though language from non-standard language varieties[1] is increasingly abundant on social media, few resources exist for developing NLP tools to handle such language, or even assessing NLP tools' performance on such language. In this chapter, we conduct a case study of such language in online conversational text by investigating African American Language (AAL) on Twitter. We present a distantly supervised model to identify AAL-like language from demographics associated with geo-located messages, and we verify that this language follows well-known AAL linguistic phenomena.

As many of these non-standard varieties have traditionally existed primarily in oral contexts and treated as illegitimate by formal institutions, they have historically been underrepresented in written sources. for more on the connections between language and power. Consequently, NLP tools have been developed from text which aligns with mainstream language varieties. With the rise of social media, however, non-standard varieties are playing an increasingly prominent role in online conversational text, for which traditional NLP tools may be insufficient. Because NLP systems are increasingly deployed on social media data, for instance for characterizing social movements [Sech et al., 2020] or public health [Santillana et al., 2015], NLP systems risk amplifying existing injustices of stigmatization, mischaracterization, or erasure of non-standard varieties and their speakers if their performance on

[1]Because the terms *language* and *dialect* are generally socio-political designations, throughout this thesis we prefer the term *language variety* instead. In this chapter, we use "non-standard" to describe language varieties, including African American Language, which are (as the name would suggest) not the standard variety, are often associated with particular regions or social groups, and are often socially stigmatized [Craft et al., 2020]. We recognize that what is considered "non-standard" is a function of power; see Ch. 6 for more on the connections between language and power.

social media text is not examined. Since this data is now available, we seek to analyze current NLP challenges and extract non-standard language from online data.

Specifically, we investigate African American Language (AAL), a variety spoken by millions of people across the United States which has been the subject of a rich body of sociolinguistic literature, in publicly available Twitter data. Due to its widespread use, established history in the sociolinguistic literature, and demographic associations, AAL provides an ideal starting point for the development of a statistical model that uncovers non-standard language. In fact, its presence in social media is attracting increasing interest for natural language processing [Jørgensen et al., 2016] and sociolinguistic [Stewart, 2014, Eisenstein, 2016, Jones, 2015] research.

In this chapter, we develop a method to identify *demographically aligned* text and language from geo-located messages (§2.2), based on distant supervision of geographic Census demographics through a statistical model that assumes a soft correlation between demographics and language. We validate our approach by verifying that text aligned with African American demographics follows well-known phonological and syntactic properties of AAL (§2.3). Finally, we provide an additional corpus of tweets containing AAL morphosyntactic features to facilitate sociolinguistic analysis (§2.4).

2.1.1 What is African American Language?

Most simply defined, African American Language (AAL) is "language as spoken by or among African Americans" [Lanehart, 2015], with systematic lexical, phonological, morphological, syntactic, and prosodic patterns [Labov, 1972, Spears, 1998, Rickford, 1999, Green, 2002, Lanehart et al., 2015]. Despite enduring perceptions of AAL as ungrammatical, bad, or lazy English, it is rule-governed with a consistent grammar; we will explore some of its patterns in the validation of the corpus we develop later in this chapter.[2]

The variety has had many different names, but is now generally called "African American English" (AAE), "African American Vernacular English" (AAVE), or "African American

[2]We note that AAL is not homogeneous, but exhibits variation across time, space, and social context [King, 2020]; we discuss this variation later in the chapter.

Language" (AAL) [Green, 2002, Wolfram and Schilling, 2015, Rickford and King, 2016, King, 2020]. The shifts in terminology reflect both changing conceptualizations of the variety and changing socio-political commitments; currently "AAL" is often used as a term that encompasses "all variations of language use in African American communities" [Lanehart et al., 2015] without pre-supposing particular linguistic features [King, 2020]. For Lanehart et al. [2015], the term AAL also reflects a move away from "the problematic implications of 'English' within the socioculture and history of African slave descendants in the United States and the contested connections of their language variety to the motherland and colonization" and toward "encompass[ing] rhetorical and pragmatic strategies that might not be associated with English."[3][4]

Due to its neutrality (with regard to assumed linguistic features) and expansiveness, in this thesis we will use the term "AAL," except where cited work uses different terminology.

2.2 Identifying AAL from demographics

2.2.1 AAL on Twitter

The presence of AAL in social media and the generation of resources of AAL-like text for NLP tasks has attracted recent interest in sociolinguistic and NLP research; Jones [2015] shows that non-standard AAL orthography on Twitter aligns with historical patterns of African American migration in the U.S., while Jørgensen et al. [2015] investigate to what extent Twitter data supports well-known sociolinguistic hypotheses about AAL. Both, however, find AAL-like language on Twitter through keyword searches, which may not yield broad corpora reflective of general AAL use. More recently, Jørgensen et al. [2016] generated a large unlabeled corpus of text from hip-hop lyrics, subtitles from *The Wire* and *The Boondocks*, and tweets from a region of the southeast U.S. While this corpus does in-

[3]See King [2020] for an extended discussion of shifting conceptualizations of the variety and of race and language more broadly, and consequences for sociolinguistic theory and methodology.

[4]For more on Black linguistics, culture, and history globally, see Smitherman [1986], Rickford and Rickford [2000], and Makoni et al. [2003].

deed capture a wide variety of language, we aim to discover AAL-like language by using finer-grained, neighborhood-level demographics from across the country.

Our approach to identifying AAL-like text is to first harvest a set of messages from Twitter, cross-referenced against U.S. Census demographics (§2.2.2), then to analyze words against demographics with two alternative methods, a seedlist approach (§2.2.3) and a mixed-membership probabilistic model (§2.2.4).

2.2.2 Twitter and Census data

In order to create a corpus of demographically associated non-standard language, we turn to Twitter, whose public messages contain large amounts of casual conversation and non-standard speech [Eisenstein, 2016]. It is well-established that Twitter can be used to study both regional language variation—for example, of American English [Doyle, 2014, Huang et al., 2016]—as well as minoritized varieties; for example, Lynn et al. [2015] develop POS corpora and taggers for Irish tweets.

Some methods exist to associate messages with authors' races; one possibility is to use birth record statistics to identify African American-associated names, which has been used in (non-social media) social science studies [Bertrand and Mullainathan, 2004, Sweeney, 2013]. However, metadata about authors is fairly limited on Twitter and most other social media services, and many supplied names are obviously not real; moreover, the automated inference of race using names is an ethically fraught procedure.

Instead, we turn to geo-location to induce a distantly supervised mapping between authors and the demographics of the neighborhoods they live in [O'Connor et al., 2010, Eisenstein et al., 2011, Stewart, 2014]. We draw on a set of geo-located Twitter messages, most of which are sent on mobile phones, by authors in the U.S. in 2013. (These are selected from a general archive of the "Gardenhose/Decahose" sample stream of public Twitter messages [Morstatter et al., 2013].) Geo-located users are a particular sample of the userbase [Pavalanathan and Eisenstein, 2015], but we expect it is reasonable to compare users of different races within this group.

We look up the U.S. Census blockgroup geographic area that the message was sent in; blockgroups are one of the smallest geographic areas defined by the Census, typically containing a population of 600–3000 people. We use race and ethnicity information for each blockgroup from the Census' 2013 American Community Survey, defining four covariates: percentages of the population that are non-Hispanic whites, non-Hispanic blacks, Hispanics (of any race), and Asian. Finally, for each user u, we average the demographic values of all their messages in our dataset into a length-four vector $\pi_u^{(census)}$. Under strong assumptions, this could be interpreted as the probability of which race the user is; we prefer to think of it as a rough proxy for likely demographics of the author and the neighborhood they live in.

Messages were filtered in order to focus on casual conversational text; we exclude tweets whose authors had 1000 or more followers, or that (a) contained 3 or more hashtags, (b) contained the strings "http," "follow," or "mention" (messages designed to generate followers), or (c) were retweeted (either containing the string "rt" or marked by Twitter's metadata as re-tweeted).

Our initial Gardenhose/Decahose stream archive had 16 billion messages in 2013; 90 million were geo-located with coordinates that matched a U.S. Census blockgroup. 59.2 million tweets from 2.8 million users remained after pre-processing; each user is associated with a set of messages and averaged demographics $\pi_u^{(census)}$.

2.2.3 Direct word-demographic analysis

Given a set of messages and demographics associated with their authors, a number of methods could be used to infer statistical associations between language and demographics.

Direct word-demographic analysis methods use the $\pi_u^{(census)}$ quantities to calculate statistics at the word level in a single pass. An intuitive approach is to calculate the *average demographics per word*. For a token in the corpus indexed by t (across the whole corpus), let $u(t)$ be the author of the message containing that token, and w_t be the word token. The average demographics of word type w is:[5]

[5] $\pi_{w,k}$ has the flavor of "soft counts" in multinomial EM. By changing the denominator to $\sum_t \pi_{u(t)}^{(census)}$, it calculates a unigram language model that sums to one across the vocabulary. This hints at a more complete modeling approach (§2.2.4).

$$\pi_w^{(softcount)} \equiv \frac{\sum_t 1\{w_t = w\}\pi_{u(t)}^{(census)}}{\sum_t 1\{w_t = w\}}$$

We find that terms with the highest $\pi_{w,\text{AA}}$ values (denoting high average African American demographics of their authors' locations) are very non-standard, while Stewart [2014] and Eisenstein [2013] find large $\pi_{w,\text{AA}}$ associated with certain AAL linguistic features.

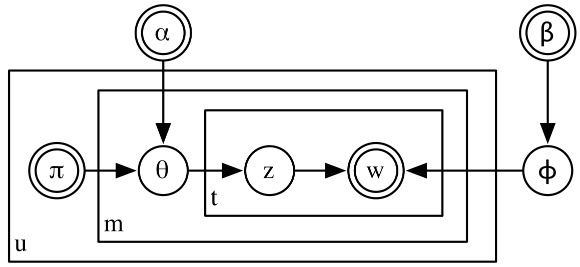
One way to use the $\pi_{w,k}$ values to construct a corpus is through a seedlist approach. In early experiments, we constructed a corpus of 41,774 users (2.3 million messages) by first selecting the $n = 100$ highest-$\pi_{w,\text{AA}}$ terms occurring at least $m = 3000$ times across the data set, then collecting all tweets from frequent authors who have at least 10 tweets and frequently use these terms, defined as the case when at least $p = 20\%$ of their messages contain at least one of the seedlist terms. Unfortunately, the n, m, p thresholds are ad-hoc.

2.2.4 Mixed-membership demographic language model

The direct word-demographics analysis gives useful validation that the demographic information may yield corpora of non-standard language, and the seedlist approach can assemble a set of users with heavy non-standard usage. However, the approach requires a number of ad-hoc thresholds, cannot capture authors who only occasionally use demographically aligned language, and cannot differentiate language use at the message-level. To address these concerns, we develop a mixed-membership model for demographics and language use in social media.

The model directly associates each of the four demographic variables with a topic; i.e. a unigram language model over the vocabulary.[6] The model assumes an author's mixture over the topics tends to be similar to their Census-associated demographic weights, and that every message has its own topic distribution. This allows for a single author to use different types of language in different messages, accommodating authors speaking multiple varieties. The message-level topic probabilities θ_m are drawn from an asymmetric Dirichlet centered on $\pi_u^{(census)}$, whose scalar concentration parameter α controls whether authors' language is very similar to the demographic prior, or can have some deviation. A token t's latent topic

[6] To build the vocabulary, we select all words used by at least 20 different users, resulting in 191,873 unique words; other words are mapped to an out-of-vocabulary symbol.



$$\theta_m \sim Dir(\alpha\pi_u), \quad \phi \sim Dir(\beta/V)$$
$$z_t \sim \theta_m, \quad w_z \sim \phi_{z_t}$$

Figure 2.1: Mixed-membership model for users (u), messages (m) and tokens (t). Observed variables have a double lined border.

z_t is drawn from θ_m, and the word itself is drawn from ϕ_{z_t}, the language model for the topic (Figure 2.1).

Thus the model learns *demographically aligned* language models for each demographic category. The model is much more tightly constrained than a topic model—for example, if $\alpha \to \infty$, θ becomes fixed and the likelihood is concave as a function of ϕ—but it still has more joint learning than a direct calculation approach, since the inference of a messages' topic memberships θ_m is affected not just by the Census priors, but also by the language used. A tweet written by an author in a highly AA neighborhood may be inferred to be non-AA-aligned if it uses non-AAL-associated terms; as inference proceeds, this information is used to learn sharper language models.

We fit the model with collapsed Gibbs sampling [Griffiths and Steyvers, 2004] with repeated sample updates for each token t in the corpus,

$$p(z_t = k \mid w, z_{-t}) \propto \frac{N_{wk} + \beta/V}{N_k + \beta} \frac{N_{mk} + \alpha\pi_{uk}}{N_m + \alpha}$$

where N_{wk} is the number of tokens where word w occurs under topic $z = k$, N_{mk} is the number of tokens in the current message with topic k, etc.; all counts exclude the current t position.

We observed convergence of the log-likelihood within 100 to 200 iterations, and ran for 300 total.[7] We average together count tables from the last 50 Gibbs samples for analysis of posterior topic memberships at the word, message, and user level; for example, the posterior probability a particular user u uses topic k, $P(z = k \mid u)$, can be calculated as the fraction of tokens with topic k within messages authored by u.

We considered α to be a fixed control parameter; setting it higher increases the correlations between $P(z = k \mid u)$ and $\pi_{u,k}^{(census)}$. We view the selection of α as an inherently difficult problem, since the correlation between race and AAL usage is already complicated and imperfect at the author-level, and census demographics allow only for rough associations. We set $\alpha = 10$ which yields posterior user-level correlations of $P(z = AA \mid u)$ against $\pi_{u,AA}$ to be approximately 0.8.

This model has broadly similar goals as non-latent, log-linear generative models of text that condition on document-level covariates [Monroe et al., 2008, Eisenstein et al., 2011, Taddy, 2013]. The formulation here has the advantage of fast inference with large vocabularies (since the partition function never has to be computed), and gives probabilistic admixture semantics at arbitrary levels of the data. This model is also related to topic models where the selection of θ conditions on covariates [Mimno and McCallum, 2008, Ramage et al., 2011, Roberts et al., 2013], though it is much simpler without full latent topic learning.

In early experiments, we used only two classes (AA and not AA), and found Spanish terms being included in the AA topic. Thus we turned to four race categories in order to better draw out non-AAL language. This removed Spanish terms from the AA topic; interestingly, they did not go to the Hispanic topic, but instead to the Asian topic, along with other language varieties.

In fact, the correlation between users' Census-derived proportions of Asian populations, versus this posterior topic's proportions, is only 0.29, while the other three topics correlate to their respective Census priors in the range 0.83 to 0.87. This indicates the "Asian" topic

[7] Our straightforward single core implementation (in Julia) spends 80 seconds for each iteration over 586 million tokens.

actually functions as a background topic (at least in part). Better modeling of demographics and non-English varieties' interactions is interesting potential future work.

By fitting the model to data, we can directly analyze unigram probabilities within the model parameters ϕ, but for other analyses, such as analyzing larger syntactic constructions and testing NLP systems, we require an explicit corpus of AAL- and MUSE-like messages.

To generate a user-based AA-aligned corpus, we collected all tweets from users whose posterior proportion of AA-associated terms used under the model was at least 80%, and generated a corresponding white-aligned corpus as well. In order to remove the effects of non-English language varieties, and given uncertainty about what the model learned in the Hispanic and Asian-aligned demographic topics, we focused only on AA- and white-aligned language by imposing the additional constraint that each user's combined posterior proportion of the Hispanic or Asian topics was less than 5%. Our two resulting user corpora contain 830,000 and 7.3 million tweets. In the rest of this work, we refer to these as the AA- and white-aligned corpora, respectively. We refer to the entire corpus as the TwitterAAE dataset.[8]

2.3 Linguistic validation

Because validation by manual inspection of our AA-aligned text is impractical,[9] we turn to the well-studied phonological and syntactic phenomena that traditionally distinguish AAL from MUSE. We validate our model by reproducing these phenomena, and document a variety of other ways in which our AA-aligned text diverges from MUSE.

---

[8]This name is inconsistent with our use of "AAL" in this thesis; however, since the dataset has been released as the TwitterAAE dataset and has been referred to as such in subsequent work (e.g., Davidson et al. [2019], Sap et al. [2019], Rios [2020]), we keep the name for consistency.

[9]Not only is it impractical to determine whether individual tweets are written "in AAL" or not, but it is also difficult (or perhaps impossible) to develop principled criteria for determining the boundaries of AAL or any other language variety; see the conclusion (§10) for a discussion.

2.3.1 Lexical-level validation

We begin by examining how much AA- and white-aligned lexical items diverge from a standard dictionary. We used SCOWL's largest wordlist with level 1 variants as our dictionary, totaling 627,685 words.[10]

We calculated, for each word w in the model's vocabulary, the ratio

$$r_k(w) = \frac{p(w|z=k)}{p(w|z \neq k)}$$

where the $p(.|.)$ probabilities are posterior inferences, derived from averaged Gibbs samples of the sufficient statistic count tables N_{wk}.

We selected heavily AA- and white-aligned words as those where $r_{AA}(w) \geq 2$ and $r_{white}(w) \geq 2$, respectively. We find that while 58.2% of heavily white-aligned words were not in our dictionary, fully 79.1% of heavily AA-aligned words were not. While a high number of out-of-dictionary lexical items is expected for Twitter data, this disparity suggests that the AA-aligned lexicon diverges from MUSE more strongly than the white-aligned lexicon.

2.3.2 Internet-specific orthography

We performed an "open vocabulary" unigram analysis by ranking all words in the vocabulary by $r_{AA}(w)$ and browsed them and samples of their usage. Among the words with high r_{AA}, we observe a number of Internet-specific orthographic variations, which we separate into three types: abbreviations (e.g. *llh*, *kmsl*), shortenings (e.g. *dwn*, *dnt*), and spelling variations which do not correlate to the word's pronunciation (e.g. *axx*, *bxtch*). These variations do not reflect features attested in the literature; rather, they appear to be purely orthographic variations highly specific to AAL-speaking communities online. They may highlight previously unknown linguistic phenomena; for example, we observe that *thoe* (MUSE *though*) frequently appears in the role of a discourse marker instead of its MUSE usage (e.g., *Girl Madison outfit THOE*). This new use of *though* as a discourse marker, which is difficult to observe using the MUSE spelling amidst many instances of the MUSE

[10] `http://wordlist.aspell.net/`

usage, is readily identifiable in examples containing the *thoe* variant. Thus, non-standard spellings provide valuable windows into a variety of linguistic phenomena.

In the next section, we turn to variations which do appear to arise from known phonological processes.

2.3.3 Phonological validation

Many phonological features are closely associated with AAL [Green, 2002]. While there is not a perfect correlation between orthographic variations and people's pronunciations, Eisenstein [2013] shows that some genuine phonological phenomena, including a number of AAL features, are accurately reflected in orthographic variation on social media. We therefore validate our model by verifying that spellings reflecting known AAL phonological features align closely with the AA topic.

We selected 31 variants of MUSE words from previous studies of AAL phonology on Twitter [Jørgensen et al., 2015, Jones, 2015]. These variations display a range of attested AAL phonological features, such as derhotacization (e.g. *brotha*), deletion of initial g and d (e.g. *iont*), and realization of voiced *th* as d (e.g. *dey*) [Rickford, 1999].

Table 2.1 shows the top ten of these words by their $r_{AA}(w)$ ratio. For 30 of the 31 words, $r \geq 1$, and for 13 words, $r \geq 100$, suggesting that our model strongly identifies words displaying AAL phonological features with the AA topic. The sole exception is the word *brotha*, which appears to have been adopted into general usage as its own lexical item.

2.3.4 Syntactic validation

We further validate our model by verifying that it reproduces well-known AAL syntactic constructions, investigating three well-attested AAL aspectual or preverbal markers: habitual *be*, future *gone*, and completive *done* [Green, 2002]. Table 2.2 shows examples of each construction.

To search for the constructions, we tagged the corpora using the ARK Twitter POS tagger [Gimpel et al., 2011, Owoputi et al., 2013],[11] which Jørgensen et al. [2015] show

[11] Version 0.3.2: `http://www.cs.cmu.edu/~ark/TweetNLP/`

AAL	Ratio	MUSE
sholl	1802.49	sure
iont	930.98	I don't
wea	870.45	where
talmbout	809.79	talking about
sumn	520.96	something
wateva	506.83	whatever
mova	382.93	mother
hea	332.48	here
nun	194.75	nothing
dey	183.39	they

Table 2.1: Of 31 phonological variant words, top ten by ratio $r_{AA}(w)$. Approximate MUSE equivalents are shown for reference.

Construction	Example	Ratio
O-*be*/*b*-V	*I be tripping bruh*	11.94
gone/*gne*/*gon*-V	*Then she gon be single Af*	14.26
done/*dne*-V	*I done laughed so hard that I'm weak*	8.68

Table 2.2: AAL syntactic constructions and the ratios of their occurrences in the AA- vs. white-aligned corpora (§2.2.4).

has similar accuracy rates on both AAL and non-AAL tweets, unlike other POS taggers. We searched for each construction by searching for sequences of unigrams and POS tags characterizing the construction; e.g. for habitual *be* we searched for the sequences O-*be*-V and O-*b*-V. Spelling variants for the unigrams in the patterns were identified from the ranked analysis of §2.3.2.

We examined how a message's likelihood of using each construction varies with the message's posterior proportion of AA. We split all messages into deciles based on the messages' posterior proportion of AA. From each decile, we sampled 200,000 messages and calculated the proportion of messages containing the three syntactic constructions. For all three constructions, we observed the clear pattern that as messages' posterior proportions of AA increase, so does their likelihood of containing the construction. Interestingly, for all three constructions, frequency of usage peaks at approximately the [0.7, 0.8) decile. One possible reason for the decline in higher deciles might be tendency of high-AA messages to be shorter;

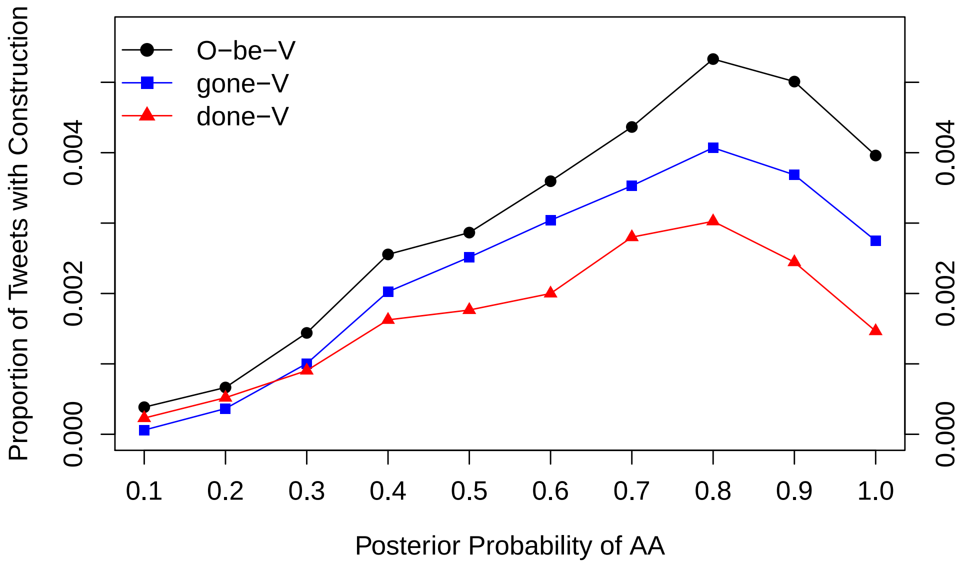


Figure 2.2: Proportion of tweets containing AAL syntactic constructions by messages' posterior proportion of AA. On the x-axis, 0.1 refers to the decile [0, 0.1).

while the mean number of tokens per message across all deciles in our samples is 9.4, the means for the last two deciles are 8.6 and 7.1, respectively.

We next turn to the collection of a dataset for analyzing AAL morphosyntactic features.

2.4 A dataset for AAL morphosyntactic analysis

Since the 1960s, sociolinguists have done a great deal to dispute the idea that African American Language is deficient or linguistically inadequate by providing considerable evidence of its systematicity [Wolfram, 2015]. However, as a result of these early descriptive studies of AAL structural features, there has been what Wolfram [2015] has called an axiomatization of the "structural homogeneity" of the variety, which has been characterized by an "apparent transregional distribution of a shared set of features."

In contrast, more recent work has shown considerable regional and social variation in African American Language; for example, Forrest and Wolfram [2019] compare data from a subset of the Corpus of Regional African American Language (CORAAL) from Washington, D.C. recorded in 1968-1969 with the results of a 1968 study in Detroit, showing important regional differences in usage. Fisher [2018] examines data from a corpus of casual conversations collected in the 1980s from African Americans in Philadelphia, finding that the use of *ain't* followed by a base verb (as opposed to the preterite) is a relatively recent innovation. Farrington [2019] investigates word final /d/, demonstrating both North and South as well as urban and rural differentiation. Such work shows that African American Language is hardly uniform across time, space, or social context, and indeed the assumption of uniformity runs counter to the assumption central to sociolinguistics that language varies; as Lanehart [2015] puts it: "[T]o believe that language used in the African American community is homogeneous across space and place goes against the core of what language variationists and sociolinguists hold as a tenet of their discipline."

Nevertheless, due to the laborious nature of gathering and transcribing sociolinguistic data, work focusing on variation in AAL has largely focused on smaller numbers of speakers in restricted numbers of geographic contexts. One notable exception is the recent work of Jones [2015] and Jones [2020], which gather corpora of tweets containing lexical items whose spellings reflect AAL phonology. They find distinct dialect regions corresponding to patterns of the Great Migration out of the American South, which do not match traditionally described North American dialect regions. Here, we aim to aid this line of work by developing a dataset for examining AAL morphosyntactic variation on Twitter.

In the next section, we motivate this approach by describing some recent work investigating syntactic variation using large corpora, as well as work examining AAL on Twitter, before describing our dataset.

2.4.1 Related work

2.4.1.1 Corpus approaches to geographic syntactic variation

A variety of datasets and computational approaches have been used to examine geographic syntactic variation; for example, Grieve [2012] uses a corpus of letters to the editor from 200 cities across the U.S. to explore regional patterning of adverb phrase positions, finding that three variables are significantly regionally patterned, consistently distinguishing a Northeast dialect region. Grieve [2016] extends this work to wide range of other syntactic features on the same corpus. Haddican and Johnson [2012] examines geographical variation in particle verb alternation by gathering a Twitter corpus containing variations of the strings *turn on the light* and *turn off the light*. Following this work, Grafmiller and Szmrecsanyi [2018] use data from the International Corpus of English (ICE) and the Global Corpus of Web-based English to investigate geographical variation in constraints on the alternation. Brook and Tagliamonte [2016] examine variation, including regional variation, in the use of *try and* vs. *try to* in British and Canadian English using a variety of corpora.

Twitter has proved to be an especially fruitful source of data. Stewart [2014] uses a combination of part-of-speech tagging and string searching to find instances of AAL syntactic features on Twitter, for example using the combination of a non-verb, followed by a pronoun, followed by an adjective, to find instances of copula deletion. Kemp et al. [2016] examine variation in the use of past-tense spreading of the past participle of *go* (*gone* vs. *went*) in the metropolitan areas of San Francisco, Los Angeles, New York City, and Atlanta, finding that Atlanta users are more likely to use *went*, and West Coast cities less likely to do so. Jones [2016] constructs a corpus of tweets containing the item *eem*, showing examples of usage that may represent a syntactic change in progress and (possibly) a step in Jespersen's Cycle [Dahl, 1979]. Stevenson [2016] investigates geographical variation in the use of the ditransitive, which in British English has three variants, by constructing a Twitter corpus of the variants using a small number of string searches, while Bohmann [2017] explores variation in the use of 236 frequency variables across world Englishes using ICE and a new Twitter corpus.

Austen [2017] examines the distribution of the lexicosyntactic variables *put up/put away* and *test over/test on*, finding that while usage of *put up* by white speakers is concentrated in the South, usage by African American speakers patterns like the Great Migration region found by Jones [2015]. Strelluf [2019] studies regional variation in positive *anymore* in American English by constructing a corpus of over 80,000 tweets from eight cities, finding both that positive *anymore* is distinctive to the Midland, and that there is regional variation within the Midland—subtle patterns of variation that would have been harder to gather with sociolinguistic interviews, particularly as positive *anymore* is a sparse feature. Storoshenko

[2020] uses Twitter data to examine the variables governing variation in the use of nine possible English th- reflexive forms.

Finally, Robinson and Duncan [2018] argue for a holistic approach that combines large corpus studies with acceptability judgment tasks; the former (particularly on Twitter) can capture language change in real time and offer insights into broad patterns of usage, including what factors condition usage, while the latter permits examination of a speaker's grammar. They present a case study of the *wh-all* variable (e.g. *Who bought a car?* vs. *Who all bought a car?*), performing corpus analysis in the Corpus of Contemporary American English and in a corpus of Twitter gathered by searching for the strings *who all* and *what all*.

2.4.1.2 AAL on Twitter

Some recent work has begun using Twitter as a source of data for analyzing AAL usage, for instance the work of Stewart [2014], Jones [2015], and Jones [2020] we described above. Meanwhile, Smith [2019] addresses the question of whether the n-word has been reappropriated by examining its usage in Facebook and Twitter data as well as Library of Congress ex-slave recordings. Ilbury [2019] examines not native AAL speakers but rather the adoption of AAL features, specifically spelling variants associated with AAL used by gay British men on Twitter to deploy a "Sassy Queen" persona.

Jørgensen et al. [2015] examine eight hypotheses about AAL, including hypotheses about particular phonological features and about social variables connected to AAL (e.g., income, educational level) using geo-located tweets featuring the spelling variants of interest. Following this work, Jørgensen et al. [2016] propose a better part-of-speech tagger for AAL, trained using a small corpus of annotated tweets [Owoputi et al., 2013] and evaluated using a corpus containing tweets, hip-hop lyrics, and subtitles from *The Wire* and *The Boondocks*.

2.4.2 Dataset goals

As the related work illustrates, large-scale social media datasets offer several important advantages for sociolinguistic analysis. First, they can offer many more examples of linguistic phenomena of interest, particularly if the phenomena are sparse. Second, they can help mitigate the *observer's paradox*, where informants' language use in front of an interviewer

Feature	Examples
Habitual *be*	She be telling people... Your phone bill be high. It be knives in here. I be in my office by 7:30.
Stressed *BIN*	I BIN had this.
Resultant *done*	She done ate. She had/been/had been done ate. I be done ate.
Gone	She gone eat.
Finna	I'm finna leave. They be finna go to bed when I call there.
Steady	Them students be steady trying to make a buck.
Multiple negation	I sure hope it don't be nothing/no leak... I don't never have nothing/no problems. I ain't never seen nobody/nothing.
Negative inversion	Don't/Can't/Ain't nothing/nobody... Don't no game last all night long.
Non-inverted negative concord	Nobody wouldn't ride that bus.

Table 2.3: AAL morphosyntactic features and example sentences drawn from Green [2002].

may be affected by the interview content.[12] Third, they can offer evidence from speakers of a wider range of geographic and socioeconomic backgrounds, which is particularly important for the study of AAL since studies have traditionally focused on younger, urban male speakers [King, 2020].

We gather tweets containing a range of AAL morphosyntactic features, paying special attention to two features, which we describe below. Table 2.3 lists these features and provides examples of each. Through this dataset, we aim to facilitate sociolinguistic analysis of these features in two ways: first, we aim to facilitate large-scale analyses of patterns surrounding feature usage—for instance, what kinds of verbs are likely to follow resultant *done*? Second, we aim to facilitate analyses of regional patterns in usage by providing geo-located data.

[12]This is somewhat complicated, however; though social media datasets do avoid the effect of the observer from a traditional sociolinguistic interview, social media language is nonetheless affected (as is all language) by the audience. See for instance Shoemark et al. [2017] and Stewart et al. [2018] for examinations of audience effects on social media language.

We describe two features of particular interest in more detail: stressed *BIN* and non-inverted negative concord.

2.4.2.1 Stressed *BIN*

In AAL, stressed *BIN*[13] denotes events "within some period in the remote past" [Green, 1998]. It can be used before stative (as in 1a) and 1b)) or non-stative (as in 2)) verbs[14] (all examples drawn from [Green, 1998]):

1. (a) I BIN treating them like that.

   'I have treated them like that for a long time.'

   (b) I BIN had this.

   'I have had this for a long time.'

2. I BIN quit school.

   'I quit school a long time ago.'

We aim to collect instances of stressed *BIN* from Twitter in order to examine its geographic distribution and the factors governing its use. This is especially challenging because this feature is prosodically marked; it is distinguished from unstressed *been* by (as the names suggest) stress on the *been* token, as well as possible other differences in pitch contours across the sentence. This therefore represents a particularly interesting feature for linguistic analysis, since relatively little work has examined how social media users mark prosody, much less on a syntactic feature. Moreover, it has interesting implications for NLP systems, in that it is an example of a feature denoting tense and aspect that is prosodically marked—that is, whether an instance of the token *been* on Twitter indicates a present progressive usage or a remote past usage depends on whether it is stressed. Therefore, accurate computational analysis of the tense and aspect that are intended depends on identifying prosodic features (does the token look like it was intended to be stressed?) and contextual details (does the surrounding context support a remote past reading?).

---

[13]Following Green [1998], we notate stressed *been* as *BIN*.

[14]Stative verbs generally describe a state, as opposed to actions.

2.4.2.2 Non-inverted negative concord

Negative auxiliary inversions are a relatively common feature both in AAL and in the white vernaculars with which AAL often coexists [Green, 2014]:

1. Can't nobody tell you it wasn't meant for you.
2. Wouldn't nobody ride that bus.

These constructions are identified by a negated auxiliary followed by a negative indefinite determiner phrase (e.g., *nobody, no one, nothing*).

Non-inverted negative concord constructions, which feature a negative auxiliary and negative indefinite DP which are not inverted, also occur in AAL:

3. Nobody can't tell you it wasn't meant for you.
4. Nobody wouldn't ride that bus.

We would like to examine the geographic distribution of this feature, as well as factors governing its usage, for two reasons. First, it is unknown whether this non-inverted version occurs in non-AAL varieties of English; Green [2014] notes, for example, that "there have been some reports that these constructions do not occur in West Texas English." Because this feature is relatively sparse, Twitter is far more likely to provide evidence for its usage than typical sociolinguistic interviews are. Second, Green [2014] hypothesizes that the two different constructions give rise to different readings in AAL: specifically, that the inverted structures give rise to an "absolute negation reading"; for example, 2) gives the reading that absolutely no one rode the bus, whereas 4) permits that reading, but also permits a weaker reading where only a few people rode the bus. A large set of tweets containing this feature would therefore make possible an analysis of the pragmatic contexts in which these two constructions occur, to assess whether they do indeed occur in different ones.

2.4.3 Data collection and analysis

We searched for instances of each morphosyntactic feature using a combination of tokens and POS tags in the geo-located TwitterAAE dataset described above in §2.2. Tweets were

Feature	TwitterAAE Count	Sample Precision
Habitual *be*	209,866	85.5
Stressed *BIN*	483	34.5*
Resultant *done*	21,240	85.0
Gone	87,946	96.5
Finna	16,829	99.0
Steady	1,160	93.0
Multiple negation	48,605	94.0*
Negative inversion	4,723	99.5
Non-inverted negative concord	865	85.5

Table 2.4: Statistics for each morphosyntactic feature in our TwitterAAE dataset. Asterisks denote uncertain precision estimates.

tagged using the ARK POS tagger [Owoputi et al., 2013]; the ARK POS tagset is provided in Table A.1 in the appendix. Table A.2 in the appendix provides the full search patterns for each feature we examine. Table 2.4 provides the feature counts in the TwitterAAE dataset, as well as the sample precision for each feature. Sample precision is calculated by randomly sampling 200 tweets from the dataset matching each search pattern and manually counting the number of tweets in the sample correctly containing the given feature.

Habitual be In the annotated samples, false positives from the habitual *be* search pattern primarily consisted of imperatives (e.g., *boo b my prom date*; *you be the judge*), accounting for 8/29 false positives; names (e.g., *mama B*; *Plan B*), for 7 false positives; and cases where the pattern failed to recognize the preceding auxiliary, where the preceding auxiliary was too far away from the *be* token to be identified, or where infinitival *to* was dropped before the *be* token (e.g., *if he does hell be done*; *Why would you have the mute kid be the lookout?*; *Baby want you _ be my saving grace*), for 7 false positives.

Stressed *BIN* Previous work has suggested that Twitter users indicate stress with capitalization [Heath, 2018], and analysis of the clusters in Owoputi et al. [2013] suggests that character reduplication may also be common for indicating stress. We also hypothesized that users might indicate stress using asterisks. Therefore, our search pattern for stressed *BIN* includes capitalized forms such as *BEEN* and *BIN*, forms with reduplicated characters such

Tweet
I BEEN exposed her on here lmao now we have video proof
God BEEN showin me who was fake kickin it & showin fake love but, i was blind asf
Latinos didn't invent lip liner and black people BEEN using lip liner and besides that nicki is South American??
Leonardo decaprio BEEN fine
Bruh Reggie Jackson BEEN nice!...he just ugly af
What you think I don't know, I probably BEEN knew
Dude over here mad bc I'm txtn his girl . Why be mad at me , you see my number BEEN in her phone
Mfs act like they BEEN had it .. bitch you JUST GOT IT
Curry being compared w Durant & LeBron .. Shit , I already thought he BEEN up w the big boys .
negative wale did not make em cool again we BEEN wearin em here in dc n md my whole life

Table 2.5: Examples of tweets with a word-capitalization pattern indicating stressed *BIN*.

as *beennn* and *beeen*, and forms with asterisks such as **been** (and combinations thereof). In Table 2.5 we provide examples of tweets with capitalized *BEEN* indicating stressed *BIN*.

Manual analysis of the tweets found with this search pattern proved to be challenging, as it was often difficult to determine, given individual tweets separated from their conversational context, whether a remote past reading was intended. The precision estimate reported in Table 2.4 therefore counts only those tweets where the remote past reading appeared relatively unambiguous, such as those presented in Table 2.5, reflecting a lower bound on precision. We also identified 24 of the 200 sampled tweets as not intending a remote past reading because they contain a time expression that clearly situates the action as recent or ongoing (e.g., *<u>Beeen</u> bumpin old wiz for the last 3 hours*; *Tf <u>beenn</u> waiting for my friend since sixxxx*; *I <u>beeeen</u> textin' two people ALL daaay*). We identified a further 20 as containing an auxiliary preceding the *BIN* token that the search pattern failed to filter out, or where the *BIN* token was a different word entirely (e.g., a clothing bin). Together, these 44 false positives reflect a precision upper bound of 78%.

Steady We found that the search pattern for *steady* was highly precise; of the 200 manually annotated tweets, only one was irrelevant to the syntactic feature (referencing the song

Tweet
Bitches steady thinking they know me .
ppl steady wanting my Kobe's these hoes stupid old but nobody got em
People steady focused on the past , how you supposed to move forward if you always looking behind ?
She got me up all night steady thinking bout love sing
This nigga steady think cus he lowerin my hours i care .. ion give a fuck i got two new jobs today nd he only fuckn his self
You say you go with somebody but you steady be in every nigga face

Table 2.6: Examples of tweets with *steady* preceding states, rather than activities.

"Rocky Steady" instead). The remaining 13 "false positives" all contained *steady* followed by an *-ing* verb, as expected. However, as *steady* typically "indicates that an activity is conducted in an intense, consistent, and continuous manner," the following verbs must be activity verbs rather than stative verbs; stative verbs are semantically incompatible with *steady* [Green, 2002]. We were therefore surprised to find 13 examples where *steady* was followed by a stative verb such as *think* or *want*, and in one case with the adjective *focused*; Table 2.6 provides examples. In several of these examples, this apparently ungrammatical usage occurs with other well-known linguistic features of AAL. More investigation is needed to understand the prevalence of this pattern.

Negative auxiliary inversion The pattern for this feature searched for instances of a negative indefinite DP (e.g., *nothing*, *nobody*) followed by a negative auxiliary (e.g., *can't, don't*). Here, 25 of the 29 false positives were cases where the negative indefinite DP belonging to a previous clause or sentence was identified as belonging to the same clause as the negative auxiliary, often but not always due to an absence of punctuation (e.g., *A person who trusts no one cant be trusted*; *If I didn't tell you nothing don't believe it from nobody else*; *nothing don't worry about it*).

2.5 Conclusion

In this chapter, we presented the first large dataset of AAL-like social media language, gathered using Census demographics supervision. Given the important linguistic differences we identified between our demographically aligned subcorpora, we hypothesize that current NLP systems may exhibit performance disparities between these subcorpora. We investigate this hypothesis in the following two chapters.

CHAPTER 3

FAIRNESS IN NLP TOOLS: LANGUAGE IDENTIFICATION

3.1 Introduction

Language identification, the task of classifying the major world language variety in which a message is written, is a crucial first step in almost any web or social media text processing pipeline. For example, in order to analyze the opinions of U.S. Twitter users, one might discard all non-English messages before running an English sentiment analyzer. In this chapter, we investigate the performance of language identification systems on AAL-like text.

We take the perspective that since AAL is a variety of English, it ought to be classified as English for the task of major world language identification. We hypothesize that language identification systems trained on Mainstream U.S. English data may exhibit reduced performance on AA-aligned tweets compared to white-aligned ones. To evaluate this hypothesis, we compare the behavior of existing language identifiers on our subcorpora.

We conduct our analysis first for Twitter's internal language identification algorithm as well as *langid.py*, one of the most popular open source language identification tools [Lui and Baldwin, 2012], and provide an ensemble classifier that mitigates performance disparities for *langid.py*. We then extend our analysis to a larger dataset and examine a set of widely used commercial language identification systems. Finally, we gather a new dataset of English worldwide and examine several language identification systems' performance, finding that the ensemble classifier proposed to aid AAL language identification improves performance in this setting as well.

3.2 Related work

Hughes et al. [2006] review language identification methods; social media language identification is challenging since messages are short, and also use non-standard and multiple (often related) varieties [Baldwin et al., 2013]. Researchers have sought to model code-switching in social media language [Rosner and Farrugia, 2007, Solorio and Liu, 2008, Maharjan et al., 2015, Zampieri et al., 2013, King and Abney, 2013], and recent workshops have focused on code-switching [Solorio et al., 2014] and general language identification [Zubiaga et al., 2014]. For Arabic dialect classification, work has developed corpora in both traditional and Romanized script [Cotterell et al., 2014, Malmasi et al., 2015] and tools that use n-gram and morphological analysis to identify code-switching between varieties and with English [Elfardy et al., 2014].

3.3 Twitter and *langid.py*

	AA-aligned	**WH-aligned**
langid.py	13.2%	7.6%
Twitter-1	8.4%	5.9%
Twitter-2	24.4%	17.6%

Table 3.1: Proportion of tweets in AA- and white-aligned corpora classified as non-English by different classifiers. Twitter-1 excludes Twitter classifications *undefined* and *None*, while Twitter-2 includes them.

We begin by testing *langid.py*, a widely used off-the-shelf Naive Bayes language identification system trained on over 97 language varieties and evaluated on both traditional corpora and Twitter messages [Lui and Baldwin, 2012], as well as the output of Twitter's in-house identifier, whose predictions are included in a tweet's metadata (from 2013, the time of data collection); the latter may give a language code or a missing value (*unk* or an empty/null value). We record the proportion of non-English predictions by these systems; *Twitter-1* does not consider missing values to be a non-English prediction, and *Twitter-2* does.

We noticed emojis had seemingly unintended consequences on *langid.py*'s classifications, so removed all emojis by characters from the relevant Unicode ranges. We also removed @-mentions.

User-level analysis We begin by comparing the classifiers' behavior on the AA- and white-aligned corpora. Of the AA-aligned tweets, 13.2% were classified by *langid.py* as non-English; in contrast, 7.6% of white-aligned tweets were classified as such. We observed similar disparities for *Twitter-1* and *Twitter-2*, illustrated in Table 3.1.

It turns out these "non-English" tweets are, for the most part, actually English. We sampled and annotated 50 tweets from the tweets classified as non-English by each run. Of these 300 tweets, only 3 could be unambiguously identified as written in a language variety other than English, verifying our assumption that the vast majority of messages classified as non-English are in fact in English.

Message-level analysis We examine how a message's likelihood of being classified as non-English varies with its posterior probability of AA. As in §2.3.4, we split all messages into deciles based on the messages' posterior probability of AA, and predicted language identifications on 200,000 sampled messages from each decile.

For all three systems, the proportion of messages classified as non-English increases steadily as the messages' posterior probabilities of AA increase. As before, we sampled and annotated from the tweets classified as non-English, sampling 50 tweets from each decile for each of the three systems. Of the 1500 sampled tweets, only 13 ($\sim$0.87%) could be unambiguously identified as being in a variety other than English.

3.3.1 Adapting language identification for AAL

In this section, we contribute a fix to language identification to correctly identify AAL and other non-standard varieties as English.

Message set	***langid.py***	**Ensemble**
AA-aligned	80.1%	99.5%
White-aligned	96.8%	99.9%
General	*88.0%*	*93.4%*

Table 3.2: Imputed recall of English messages in 2014 messages. For the *General* set these are an approximation; see text.

3.3.1.1 Ensemble classifier

We observed that messages where our model infers a high probability of AAL, white-aligned, or "Hispanic"-aligned language almost always are written in English; therefore we construct a simple ensemble classifier by combining it with *langid.py*.

For a new message $\vec{w}$, we predict its demographic-language proportions $\hat{\theta}$ via posterior inference with our trained model, given a symmetric α prior over demographic-topic proportions. The ensemble classifier, given a message, is as follows:

- Calculate langid.py's prediction $\hat{y}$.
- If $\hat{y}$ is English, accept it as English.
- If $\hat{y}$ is non-English, and at least one of the message's tokens are in demographic model's vocabulary: Infer $\hat{\theta}$ and return English only if the combined AA, Hispanic, and white posterior probabilities are at least 0.9. Otherwise return the non-English $\hat{y}$ decision.

Another way to view this method is that we are effectively training a system on an extended Twitter-specific English language corpus softly labeled by our system's posterior inference; in this respect, it is related to efforts to collect new language-specific Twitter corpora [Bergsma et al., 2012] or minority language data from the web [Ghani et al., 2001].

3.3.1.2 Evaluation

Our analysis from above suggests that this method should correct erroneous false negatives for AAL-like messages in the training set for the model. We confirm this by testing the classifier on a sample of 2.2 million geo-located tweets sent in the U.S. in 2014, which are not in the training set.

In addition to performance on the entire sample, we examine our classifier's performance on messages whose posterior probability of using AA- or white-associated terms was greater than 0.8 within the sample, which in this section we will call AA- and white-aligned messages, respectively. Our classifier's precision is high across the board, at 100% across manually annotated samples of 200 messages from each sample.[1] Since we are concerned about the system's overall recall, we impute recall (Table 3.2) by assuming that all high AA and high white messages are indeed English. Recall for *langid.py* alone is calculated by $\frac{n}{N}$, where n is the number of messages predicted to be English by *langid.py*, and N is the total number of messages in the set. We estimate the ensemble's recall as $\frac{n+m}{N}$, where $m = (n_{flip})P(\text{English} \mid flip)$ is the expected number of correctly changed classifications (from non-English to English) by the ensemble and the second term is the precision (estimated as 1.0). We observe the baseline system has considerable difference in recall between the groups which is mitigated by the ensemble.

We also apply the same calculation to the general set of all 2.2 million messages; the baseline classifies 88% as English. This is a less accurate approximation of recall since we have observed a substantial presence of non-English messages. The ensemble classifies an additional 5.4% of the messages as English; since these are all (or nearly all) correct, this reflects at least a 5.4% gain to recall.

3.4 Commercial systems

In this section, we extend our previous analysis from 200 to 20,000 tweets, evaluating the disparity for several black-box commercial services.

3.4.1 Experiments

We wish to assess the *racial disparity accuracy difference*:

[1] We annotated 600 messages as English, not English, or not applicable, from 200 sampled each from general, AA-aligned, and white-aligned messages. Ambiguous tweets which were too short (e.g. "Gm") or contained only named entities (e.g. "Tennessee") were excluded from the final calculations. The resulting samples have 197/197, 198/198, and 200/200 correct English classifications, respectively.

$$p(\text{correct} \mid \text{Wh}) - p(\text{correct} \mid \text{AA}) \tag{3.1}$$

A disparity of 0 indicates a language identifier that is fair across these classes.[2]

We conduct an evaluation of four different off-the-shelf language identifiers. Our work described above examined only *langid.py* and Twitter, and without length breakdowns.

- ***langid.py* (software):** one of the most popular open source language identification tools, *langid.py* was originally trained on over 97 language varieties and evaluated on both traditional corpora and Twitter messages [Lui and Baldwin, 2012].
- **IBM Watson (API):** the Watson Developer Cloud's Language Translator service supports language identification of 62 language varieties.[3]
- **Microsoft Azure (API):** Microsoft Azure's Cognitive Services supports language identification of 120 language varieties.[4]
- **Twitter (metadata):** the output of Twitter's in-house identifier, whose predictions are included in a tweet's metadata.
- **Google (API, excluded):** We attempted to test Google's language detection service,[5] but it returned a server error for every message we gave it to classify.

We queried the remote API systems in May 2017.

From manual inspection, we observed that longer tweets are significantly more likely to be correctly classified, which is a potential confound for a race disparity analysis, since the length distribution is different for each demographic group. To minimize this effect in our comparisons, we group messages into four bins (shown in Table 3.3) according to the number of words in the message. For each bin, we sampled 2,500 AA-aligned and 2,500

[2]In the taxonomy of harms we propose in Ch. 7, this operationalizes a *quality of service* harm.

[3]`https://www.ibm.com/watson/developercloud/doc/language-translator/index.html`

[4]`https://docs.microsoft.com/en-us/azure/cognitive-services/text-analytics/overview#language-detection`

[5]`https://cloud.google.com/translate/docs/detecting-language`

		AA Accuracy	WH Accuracy	Difference
langid.py	$t \leq 5$	68.0	70.8	2.8
	$5 < t \leq 10$	84.6	91.6	7.0
	$10 < t \leq 15$	93.0	98.0	5.0
	$t > 15$	96.2	99.8	3.6
IBM Watson	$t \leq 5$	62.8	77.9	15.1
	$5 < t \leq 10$	91.9	95.7	3.8
	$10 < t \leq 15$	96.4	99.0	2.6
	$t > 15$	98.0	99.6	1.6
Microsoft Azure	$t \leq 5$	87.6	94.2	6.6
	$5 < t \leq 10$	98.5	99.6	1.1
	$10 < t \leq 15$	99.6	99.9	0.3
	$t > 15$	99.5	99.9	0.4
Twitter	$t \leq 5$	54.0	73.7	19.7
	$5 < t \leq 10$	87.5	91.5	4.0
	$10 < t \leq 15$	95.7	96.0	0.3
	$t > 15$	98.5	95.1	-3.0

Table 3.3: Percent of the 2,500 tweets in each bin classified as English by each classifier. **Difference** is the difference (disparity on an absolute scale) between the classifier accuracy on the AA-aligned and white-aligned samples. t is the message length for the bin.

white-aligned tweets, yielding a total of 20,000 messages across the two categories and four bins.[6] We limited pre-processing of the messages to fixing of HTML escape characters and removal of URLs, keeping "noisy" features of social media text such as @-mentions, emojis, and hashtags. We then calculated, for each bin in each category, the number of messages predicted to be in English by each classifier. Accuracy results are shown in Table 3.3.[7]

As predicted, classifier accuracy does increase as message lengths increase; classifier accuracy is generally excellent for all messages containing at least 10 tokens. However, the classifier results display a disparity in performance among messages of similar length; for all but one length bin under one classifier, accuracy on the white-aligned sample is higher than on the AA-aligned sample. The disparity in performance between AA- and white-

[6]Due to a data processing error, there are 5 duplicates (19,995 unique tweets); we report on all 20,000 messages for simplicity.

[7]We have made the 20,000 messages publicly available at `http://slanglab.cs.umass.edu/TwitterAAE/`.

aligned messages is greatest when messages are short; the gaps in performance for extremely short messages ranges across classifiers from 6.6% to 19.7%. This gap in performance is particularly critical as 41.7% of all AA-aligned messages in the corpus as a whole have 5 or fewer tokens.[8]

3.5 Extended evaluation: World languages

Here, to support work on English language identification, we contribute a new dataset of tweets annotated for English versus non-English, with attention to ambiguity, code-switching, and automatic generation issues. It is randomly sampled from all public messages, avoiding biases towards pre-existing language classifiers. We find that our ensemble classifier can be used to improve English language identification performance when combined with a traditional supervised language identifier. It increases recall with almost no loss of precision, including, surprisingly, for English messages written by non-U.S. authors.

We evaluate as fairly and completely as possible; we first annotate a new dataset of uniformly sampled tweets for whether they are English versus non-English (§3.5.1). In §3.5.2, we apply our model to infer U.S. demographic language proportions in new tweets, finding that when added as an ensemble to a pre-existing identifier, performance improves—including when paired with feature-based, neural network, and proprietary identifiers. Such ensembles perform better than in-domain training with the largest available annotated Twitter dataset, and also better than a self-training domain adaptation approach on the same dataset used to construct the mixed membership model—and the accuracy increases for English messages from many different countries around the world.

3.5.1 Dataset

We sampled 10,502 messages from January 1, 2013 to September 11, 2016 from an archive of publicly available geotagged tweets. We annotated the tweets with three mutually

[8] For most (system,length) combinations, the accuracy difference is significant under a two-sided t-test ($p < .01$) except for two rows ($t \leq 5$, *langid.py*, $p = .03$) and ($10 < t \leq 15$, Twitter, $p = 0.5$). Accuracy rate standard errors range from 0.04% to 0.9% ($\approx \sqrt{acc(1-acc)/2500}$).

exclusive binary labels: *English*, *Not English*, and *Ambiguous*. These tweets were further annotated with descriptive labels:

- *Code-switched*: Tweets containing both text in English and text in another language variety.
- *Ambiguous due to named entities*: Tweets containing only named entities, such as *Vegas!*, and therefore whose language could not be unambiguously determined.
- *Automatically generated*: Tweets whose content appeared to be automatically generated, such as *I just finished running 15.21 km in 1h:17m:32s with #Endomondo #endorphins https://t.co/bugbJOvJ31*.

Label	**Full Count**	**Evaluation Count**
English	5086	3758
Not English	4646	4608
Ambiguous	770	0
Total	10502	8366

Table 3.4: Dataset statistics for each language label; the evaluation count refers to the subset used for evaluation.

Label	**Count**
Code-Switched	162
Ambiguous due to Named Entities	132
Automatically Generated	1371

Table 3.5: Dataset statistics for additional labels.

We excluded any usernames and URLs in a tweet from the judgment of the tweet's language, but included hashtags. Tables 3.4 and 3.4 contain the statistics for these labels in our annotated dataset. For all our experiments, we evaluate only on the subset of messages in the dataset not labeled as ambiguous or automatically generated, which we call the evaluation dataset.

3.5.2 Experiments

3.5.2.1 Training datasets

We investigate the effect of in-domain and extra out-of-domain training data with two datasets. The first is a dataset released by Twitter of 120,575 tweets uniformly sampled from all Twitter data. which were first labeled by three different classifiers (Twitter's internal algorithm, Google's Compact Language Detector 2, and *langid.py*), then annotated by humans where classifiers disagreed.[9] We reserve our own dataset for evaluation, but use this dataset for in-domain training. This dataset is only made available by tweet ID, and many of its messages are now missing; we were able to retrieve 74,259 tweets (61.6%). For the rest of this work, we call this the Twitter70 dataset (since it originally covered about 70 language varieties). In addition, following Jaech et al. [2016], we supplemented Twitter70 with out-of-domain Wikipedia data for 41 language varieties,[10] sampling 10,000 sentences from each variety.

3.5.2.2 Classifiers

We tested a number of classifiers on our annotated dataset trained on a variety of domains, and in some cases retrained.

- CLD2: a Naive Bayes classifier with a pretrained model from a proprietary corpus; it offers no support for re-training.
- Twitter: as before, the output of Twitter's in-house identifier.
- *langid.py*
- Neural model: a hierarchical neural classifier that learns both character and word representations. It provides a training dataset with 41,250 Wikipedia sentence fragments in 33 language varieties [Jaech et al., 2016].

---

[9] `https://blog.twitter.com/2015/evaluating-language-identification-performance`

[10] `https://sites.google.com/site/rmyeid/projects/polyglot`

Self-training We experimented with one simple approach to unsupervised domain adaptation: self-training with an unlabeled target domain corpus [Plank, 2009] by using *langid.py* to label the TwitterAAE corpus of tweets, then collecting those tweets classified with posterior probability greater than or equal to 0.98. We downsampled tweets classified as English to 1 million, yielding a total corpus of 2.2 million tweets. Since we did not have access to *langid.py*'s original training data, we trained a new model on this data, then combined it as an ensemble with the original model, where a tweet was classified as English if either component classified it as English.

Demographic ensemble classifier We applied our mixed membership model as an ensemble classifier much as we did above, where tweets were first classified by an off-the-shelf classifier.

3.5.2.3 Length-normalized analysis

From manual inspection, we observed that longer tweets are significantly more likely to be correctly classified; we investigate this length effect by grouping messages into five bins (shown in Table 3.9) according to the number of words in the message. We pre-processed messages by fixing HTML escape characters and removing URLs, @-mentions, emojis, and the "RT" token. For each bin, we calculate recall of the *langid.py* and the demographic ensemble classifier with *langid.py*.

3.5.3 Results and discussion

We evaluated on the 8,366 tweets in our dataset that were not annotated as ambiguous or automatically generated. Table 3.6 shows the precision and recall for each experiment. We focus on recall, as our analysis indicates that while precision is largely consistent across experiments, there is a significant gap in recall performance across different varieties of English.

Unsurprisingly, we found that training on Twitter data improved classifiers' English recall, compared to their pre-trained models. In our experiments, we found that recall was best when training on the subset of the Twitter70 dataset containing only language varieties with at least 1,000 tweets present in the dataset. We also found that the additional

Model	Training	Precision	Recall
CLD2	(1) Pre-trained	0.948	0.863
	(2) + Demo.	0.946	0.924 (+ 6.1%)
Twitter internal	(3) Pre-trained	0.979	0.866
	(4) + Demo.	0.974	0.925 (+ 5.9%)
langid.py	(5) Pre-trained	0.923	0.886
	(6) + Vocab.	0.472	0.993
	(7) Self-trained	0.924	0.894
	(8) + Demo.	0.923	0.930 (+ 3.6%)
	(9) Re-trained on Twitter70	0.927	0.940
	(10) + Demo.	0.923	0.957 (+ 1.7%)
	(11) Twitter70 and Wiki.	0.946	0.903
	(12) + Demo.	0.943	0.946 (+ 4.3%)
Neural	(13) Pre-trained	0.973	0.415
	(14) + Demo.	0.976	0.773 (+ 35.8%)
	(15) Re-trained on Twitter70	0.949	0.840
	(16) + Demo.	0.946	0.892 (+ 5.2%)

Table 3.6: English classification results on not ambiguous, not automatically generated tweets. "+ Demo." indicates an ensemble with the demographics-based English classifier.

Country	En	∼En	*langid.py* Recall	Ens. Recall
USA	2368	80	0.968	0.982
Brazil	42	945	0.833	0.833
Indonesia	161	707	0.764	0.767
Turkey	13	304	0.769	0.846
Japan	14	340	0.929	1.0
United Kingdom	401	18	0.962	0.980
Malaysia	90	174	0.833	0.833
Spain	28	263	0.75	0.821
Argentina	10	291	0.7	0.7
France	26	206	0.846	0.846
Mexico	25	162	0.76	0.76
Philippines	91	86	0.934	0.945
Thailand	14	111	0.643	0.786
Russia	9	129	0.667	0.778
Canada	96	7	0.979	0.990

Table 3.7: Language counts for countries with at least 100 non-ambiguous, non-automatically generated messages (out of 129 countries total), with English recall for the best-performing *langid.py* model and that model in an ensemble classifier.

Tweet
@username good afternoon and Happy Birthdayyyyyyyyyy *Turns on music* Time to partyyyyy
I miss you! #vivasantotomas #goUST #igers #igdaily #igersasia #igersmanila #instagood...
Sooo fucked yuuuuppp bouuutta start a figgght
catch mines you catch yours we both happy...
Go follow me on Instagram @username and like 5 pics for a goodmorningg post
Think me & my baddies getting rooms dis weekend!
@username HML if u do B
@username @username FR LIKE I CANT EVEN DEAL WITH PEOPLE LIKE THIS
I k you dont like me lowkey but hey
@username I DORN WVEN WTCH GIRL MEETS WORLDBUT IM WATCHINF THAT EPISODE

Table 3.8: Sample of tweets which were misclassified as non-English by *langid.py* but correctly classified by the demographic ensemble. @-mentions are shown as @username for display in the table.

information provided by the demographic model's predictions still adds to the increased performance from training on Twitter data. Notably, precision decreased by no more than 0.4% when the demographic model is added.

We also noted that pre-processing improved recall by 1 to 5%.

Proprietary algorithms We found that neither CLD2 nor Twitter's internal algorithm was competitive with *langid.py* out of the box, in line with previous findings, but combining their predictions with demographic predictions did increase recall.[11]

langid.py Self-training *langid.py* produced little change compared to the original pre-trained model, (rows (5) vs. (7)), despite its use of 2.2 million new tweets from self-training step. We observed that even tweets that *langid.py* classified as non-English with more than 0.98 posterior probability were, in fact, generally English. This suggests that tweets are sufficiently different from standard training data that it is difficult for self-training to be effective. In contrast, simple in-domain training was effective; retraining with the Twitter70

[11] We tried several times to run the Google Translate API's language identifier, but it returned an internal server error for approximately 75% of the tweets.

dataset achieved substantially better recall with a 5.4% raw increase compared to its out-of-domain original pretrained model. (rows (5) vs. (9)).

In all cases, regardless of the data used to train the model, *langid.py*'s recall was improved with the addition of demographic predictions; for example, the demographic predictions added to the pre-trained model brought recall close to the model trained on Twitter70 alone, indicating that in the absence of in-domain training data, the demographic model's predictions can make a model competitive with a model that does have in-domain training data. (rows (8) vs. (9)). Of course, in-domain labeled data only helps more (10).

Neural model Finally, the neural model performed worse than *langid.py* when trained on the same Twitter70 dataset, (rows (9) vs. (15)), and its performance lagged when trained on its provided dataset of Wikipedia sentence fragments. As with the other models, demographic predictions again improve performance.

Table 3.8 shows a sample of ten tweets misclassified as non-English by *langid.py* and correctly classified by the demographic ensemble as English. Several sources of potential error are evident; many non-conventional spellings, such as *partyyyyy* and *watchinf*, do not challenge a human reader but might reasonably challenge character n-gram models. Similarly, common social abbreviations such as *hml* and *fr* are challenging.

Improving English recall worldwide We further analyzed our English recall results according to messages' country of origin, limiting our analysis to countries with at least 100 non-ambiguous, non-automatically generated messages in our dataset. For each country's messages, we compared the recall from best standalone *langid.py* model (trained on Twitter70) and the recall from same model combined with demographic predictions, as shown in Table 3.7. Surprisingly, for ten of the fifteen countries we found that using demographic predictions improved recall performance, suggesting that the additional soft signal of "Englishness" provided by the demographic model aids performance across tweets labeled as English globally.

Improving recall for short tweets Our results from the length-normalized analysis, shown in Table 3.9, demonstrate that recall on short tweets, particularly short English

	Message Length	*langid.py* Recall	Ensemble Recall
English	$t \leq 5$	80.7	91.9
	$5 < t \leq 10$	88.8	92.4
	$10 < t \leq 15$	91.9	93.0
	$15 < t \leq 20$	96.1	96.7
	$t \geq 20$	97.2	97.5
Non-English	$t \leq 5$	90.0	99.9
	$5 < t \leq 10$	95.2	99.5
	$10 < t \leq 15$	95.6	99.9
	$15 < t \leq 20$	95.2	1.0
	$t \geq 20$	95.2	1.0

Table 3.9: Percent of the messages in each bin classified correctly as English or non-English by each classifier; t is the message length for the bin.

tweets, is challenging; unsurprisingly, recall increases as tweet length increases. More importantly, for short tweets the demographic ensemble classifier greatly reduces this gap; while the difference in *langid.py*'s recall performance between the shortest and longest English tweets is 16.5%, the difference is only 5.6% for the ensemble classifier. The gap is similarly decreased for non-English tweets. We note also that precision is consistently high across all bins for both *langid.py* and the ensemble classifier. The experiment indicates that the demographic model's signal of "Englishness" may aid performance not only for global varieties of English, but also for short messages of any kind.

In the next chapter, we turn to another task early in the NLP pipeline: dependency parsing.

CHAPTER 4

FAIRNESS IN NLP TOOLS: DEPENDENCY PARSING

4.1 Introduction

In this chapter, we investigate the performance of dependency parsing systems on AAL-like text. Like language identification, parsing is an early step in many NLP pipelines; the output dependency relations are frequently used in downstream tasks such as event and relation extraction.

Here, much as in the previous chapter, we conduct an initial analysis of parsing performance on AAL-like text using a small set of partially annotated sentences drawn from the AA-aligned corpus developed in Ch. 2. We then propose new annotation guidelines for the Universal Dependencies formalism to handle features found in AAL and on Twitter, and evaluate performance under a range of training paradigms.

4.2 Related Work

4.2.1 Parsing for Twitter

Parsing for social media data presents interesting and significant challenges. Foster et al. [2011] develop a dataset of 519 constituency-annotated English tweets, which were converted to Stanford dependencies. Their analysis finds a substantial drop in performance of an off-the-shelf dependency parser on the new dataset compared to a WSJ test set. Sanguinetti et al. [2017] annotate a dataset of 6,738 Italian tweets according to UD 2.0 and examined the performance of two parsers on the dataset, finding that they lagged considerably relative to performance on the Italian UD Treebank.

Kong et al. [2014] develop an English dependency parser designed for Twitter, annotating a dataset of 929 tweets (Tweebank v1) according to the unlabeled FUDG dependency formalism [Schneider et al., 2013]. It has substantially different structure than UD (for example, prepositions head PPs, and auxiliaries govern main verbs).

More recently, Liu et al. [2018] develop Tweebank v2, fully annotating Tweebank v1 according to UD 2.0 and annotating additionally sampled tweets, for a total of 3,550 tweets. They find that creating consistent annotations is challenging, due to frequent ambiguities in interpreting tweets; nevertheless, they were able to train a pipeline for tokenizing, tagging, and parsing the tweets, and develop ensemble and distillation models to improve parsing accuracy. Our work encounters similar challenges; in our approach, we intentionally oversample AAL-heavy messages for annotation, detail specific annotation decisions for AAL-specific phenomena, and analyze parser performance between varieties and for particular constructions.

One line of work for parsing noisy social media data, including Khan et al. [2013] and Nasr et al. [2016], examines the effects of domain mismatches between traditional sources of training data and social media data, finding that matching the data as closely as possible aids performance. Other work focuses on normalization, including Daiber and van der Goot [2016] and van der Goot and van Noord [2017], which develop a dataset of 500 manually normalized and annotated tweets and uses normalization within a parser. Separately, Zhang et al. [2013] create a domain-adaptable, parser-focused system by directly linking parser performance to normalization performance.

4.2.2 Parsing for non-standard varieties

For Arabic dialects, Chiang et al. [2006] parse Levantine Arabic by projecting parses from Modern Standard Arabic translations, while Green and Manning [2010] conduct extensive error analysis of Arabic constituency parsers and the Penn Arabic Treebank. Scherrer [2011] parse Swiss German dialects by transforming Standard German phrase structures. We continue in this line of work in our examination of AAL-specific syntactic structures and generation of synthetic data with such structures.

Less work has examined parsing non-standard language on social media. Recently, Wang et al. [2017] annotate 1,200 Singlish (Singaporean English) sentences from a Singaporean talk forum, selecting sentences containing uniquely Singaporean vocabulary items. Like other work, they observe a drop in performance on dialectal Singlish text, but increase performance through a stacking-based domain adaptation method.

4.3 Preliminary analysis: Stanford dependencies

Given the many syntactic features present in AAL not present in MUSE [Green, 2002], we hypothesize that tools for syntactic analysis, such as dependency parsing tools, are likely to exhibit reduced performance on AAL-like tweets compared to MUSE-like tweets.

We assess a publicly available syntactic dependency parser on our AA- and white-aligned corpora. Syntactic parsing for tweets has received some research attention; Foster et al. [2011] create a corpus of constituent trees for English tweets, and Kong et al.'s [2014] *Tweeboparser* is trained on a Twitter corpus annotated with a customized unlabeled dependency formalism; since its data was uniformly sampled from tweets, we expect it may have low disparity between demographic groups.

We focus on widely used syntactic representations, testing the *SyntaxNet* neural network-based dependency parser [Andor et al., 2016],[1] which reports state-of-the-art results, including for web corpora. We evaluate it against a new manual annotation of 200 messages, 100 randomly sampled from each of the AA- and white-aligned corpora described in §2.2.4.

SyntaxNet outputs grammatical relations conforming to the Stanford Dependencies (SD) system [de Marneffe and Manning, 2008], which we used to annotate messages using *Brat*,[2] comparing to predicted parses for reference. Message order was randomized and demographic inferences were hidden from the annotator. To increase statistical power relative to annotation effort, we developed a partial annotation approach to only annotate edges for the root word of the first major sentence in a message. Generally, we found that that SD worked well as a descriptive formalism for tweets' syntax.

We evaluate labeled recall of the annotated edges for each message set:

---

[1] Using the publicly available *mcparseface* model: `https://github.com/tensorflow/models/tree/master/syntaxnet`

[2] `http://brat.nlplab.org/`

Parser	AA	WH	Difference
SyntaxNet	64.0 (2.5)	80.4 (2.2)	16.3 (3.4)
CoreNLP	50.0 (2.7)	71.0 (2.5)	21.0 (3.7)

Table 4.1: Bootstrapped standard errors (from 10,000 message resamplings) are in parentheses; differences are statistically significant ($p < 10^{-6}$ in both cases).

The white-aligned accuracy rate of 80.4% is broadly in line with previous work (compare to the parser's unlabeled accuracy of 89% on English Web Treebank full annotations), but parse quality is much worse on AA-aligned tweets at 64.0%. We test the Stanford CoreNLP neural network dependency parser [Chen and Manning, 2014] using the *english_ SD* model that outputs this formalism;[3] its disparity is worse.

4.4 Extended analysis: Universal dependencies

We broaden Universal Dependencies parsing[4] to handle social media English, particularly social media AAL, by developing and annotating a new dataset of 500 tweets, 250 of which are drawn from our AA-aligned corpus (Ch. 2), within the Universal Dependencies 2.0 framework. Here, we describe our standards for handling Twitter- and AAL-specific features and evaluate several state-of-the-art dependency parsers, finding that, unsurprisingly, they perform poorly on our dataset relative to the UD English Treebank. We additionally evaluate a variety of cross-domain strategies for improving parsing with no, or very little, in-domain labeled data, including a new data synthesis approach. We analyze these methods' impact on performance disparities between tweets drawn from the AA- and white-aligned corpora, and assess parsing accuracy for specific AAL lexical and syntactic features.

Specifically, we compare training in a rich out-of-domain setting—with fine-grained POS tags and morphological information—to training in a more impoverished in-domain setting—with coarse POS tags and no morphological information. We find that the in-domain but impoverished tagger leads to better parsing performance, and narrows the performance dis-

[3] *pos,depparse* options in version 2015-04-20, using tokenizations output by SyntaxNet.

[4] `http://universaldependencies.org/`

parity between MUSE- and AAL-like tweets. We confirm the importance of POS tagging, finding that coarser, but Twitter-specific and better-accuracy POS tags greatly help performance. Additionally, we evaluate a new heuristic data synthesis method to add AAL syntax and Twitter-specific features to the UD Treebank, as well as the use of in-domain, unsupervised word embeddings, and investigate their effects on the performance disparity between MUSE- and AAL-like tweets. Finally, we provide an error analysis of the parsers' performance on AAL lexical and syntactic constructions in our dataset.

4.4.1 Dataset

Our dataset contains 500 tweets, with a total of 5,951 non-punctuation edges, sampled from the corpus developed in Ch. 2. We create a balanced sample to get a range of language, sampling 250 tweets from those where the African American topic has at least 80% probability, and 250 from those where the white topic has at least 80% probability. We refer to these two subcorpora as the AA- and white-aligned tweets.

The 250 AA-aligned tweets include many spellings of common words that correspond to well-known phonological phenomena—including *da, tha* (the), *dat, dhat* (that), *dis, dhis* (this), *ion, iont* (I don't), *ova* (over), *yo* (your), *dere, der* (there), *den, dhen* (then), *ova* (over), and *nall, null* (no, nah)—where each of the mentioned italicized AAL variants appears in the AA-aligned tweets, but never in the white-aligned tweets. We examine these lexical variants more closely in §4.4.5. Across the AA-aligned tweets, 18.0% of tokens were not in a standard English dictionary, while the WH tweets' OOV rate was 10.7%.[5] We further observe a variety of AAL syntactic phenomena in our AA tweets, several of which are described in §4.4.2 and §4.4.5.

4.4.2 AAL annotation

To effectively measure parsing quality and develop better future models, we first focus on developing high-quality annotations for our dataset, for which we faced a variety of

[5]The dictionary of 123,377 words with American spellings was generated using `http://wordlist.aspell.net/`.

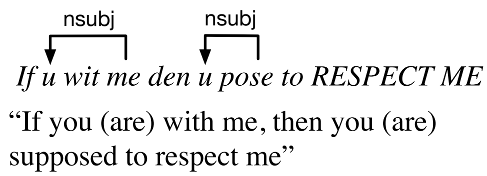


Figure 4.1: Example of null copulas and our proposed annotations.

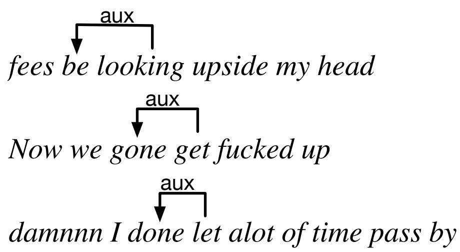


Figure 4.2: Example of verbal auxiliaries and our proposed annotations.

challenges. We detail our annotation principles for null copulas, AAL verbal auxiliaries, and shortened verbs using Universal Dependency 2.0 relations [Nivre et al., 2016].

Our annotation principles are in alignment with those proposed by Liu et al. [2018], with the exception of contraction handling, which we discuss briefly in §4.4.2.2.

4.4.2.1 Null copulas

AAL is prominently characterized by the drop of copulas, which can occur when the copula is present tense, not first person, not accented, not negative, and expressing neither the habitual nor the remote present perfect tenses [Green, 2002].

The first dropped *are* in Figure 4.1 is a null copula; UD2.0 would analyze the MUSE version as $you \xleftarrow{nsubj} me \xrightarrow{cop} are$, which we naturally extend to analyze the null copula by simply omitting *cop* (which is now over a null element, so cannot exist in a dependency graph). The second *are* is a null auxiliary (in MUSE, $you \xleftarrow{nsubj} supposed \xrightarrow{aux} are$), a tightly related phenomenon, which we analyze similarly by simply omitting the *aux* edge.

4.4.2.2 AAL verbal auxiliaries

We observed AAL verbal auxiliaries, e.g., including habitual *be* (“Continually, over and xcomp

Lol he bouta piss me off "He is about to piss me off"

Figure 4.3: Example of *bouta* and our proposed annotation.

over, fees are looking at me...”), future *gone* (“we are going to get...”), and completive *done* (“I did let time pass by,” emphasizing the speaker completed a time-wasting action) (Figure 4.2).

We attach the auxiliary to the main verb with the *aux* relation, as UD2.0 analyzes other English auxiliaries (e.g. *would* or *will*).

4.4.2.3 Verbs: Auxiliaries vs. main verbs

We observed many instances of quasi-auxiliary, “-to” shortened verbs such as *wanna, gotta, finna, bouta, tryna, gonna*, which can be glossed as *want to*, *got to*, *fixing to*, *about to*, etc. They control modality, mood and tense; for example, *finna* and *bouta* denote an immediate future tense [Green, 2002]. From UD’s perspective, it is difficult to decide if they should be subordinate auxiliaries or main verbs. In accordance with the UD Treebank’s handling of MUSE *want to X* and *going to X* as main verbs ($want \xrightarrow{xcomp} X$), we analyzed them similarly, e.g. as in Figure 4.3. This is an instance of a general principle that, if there is a shortening of an MUSE multiword phrase into a single word, the annotations on that word should mirror the edges in and out of the original phrase’s subgraph (as in Schneider et al.’s [2013] fudge expressions).

However, in contrast to the UD Treebank, we did not attempt to split up these words into their component words (e.g. *wanna* → *want to*), since to do this well, it would require a more involved segmentation model over the dozens or even hundreds of alternate spellings each of the above can take;[6] we instead rely on O’Connor et al. [2010] and Owoputi et al.’s [2013] rule-based tokenizer that never attempts to segment within such shortenings. This annotation principle contrasts with that of Liu et al. [2018], which follows UD tokenization for contractions.

[6] For example, Owoputi et al.’s [2013] Twitter word cluster 0011000 has 36 forms of *gonna* alone: `http://www.cs.cmu.edu/~ark/TweetNLP/cluster_viewer.html`

4.4.3 Non-AAL Twitter annotation

We also encountered many parsing challenges general to Twitter but not AAL, whose handling we describe here. When possible, we adapted Kong et al.'s [2014] annotation conventions into the Universal Dependencies context, which are the only published conventions we know of for Twitter dependencies (for the FUDG dependency formalism).

4.4.3.1 @-mentions: Address vs. Subject/Object

Since @-mentions do not contain linguistic content, we replaced all instances in our dataset with the token *ATMENTION*.

Twitter users employed @-mentions in a variety of ways. The most frequent way was as a means of addressing that user, in which case we annotated using the *vocative* relation to the head of the closest preceding (or following, if no preceding) clause, e.g.

ATMENTION yep look to yo left when you pass by blanton kml

@-mentions were also occasionally used as appositives, in which case we used the natural *appos* relation, e.g.

My nigga ATMENTION turnt up eith us to at the beach

Finally, we observed cases where @-mentions played important syntactic roles, as in the following example:

I rather enjoy listening to ATMENTION yelling at the staff lol

4.4.3.2 Quoting

Users frequently quote one another in retweets, as in the following example:

" ATMENTION : Cough drops are my best friend right now . " same EMOJI

In this type of message, we annotated as the root the head of the outermost main clause; in this case, the word *same*; we attached *ATMENTION* to *same* via the *parataxis* relation, and attached *friend* to *ATMENTION* again via the *parataxis* relation.

4.4.3.3 Multiple utterances and parataxis

Many messages contained multiple utterances. Since we do not attempt sentence segmentation, we follow UD's convention of the *parataxis* relation for what are described as "side-by-side run-on sentences." For each utterance after the first, we attached the head of that utterance to the head of the first utterance via the *parataxis* relation, e.g. with

> *Dan Uggla didn't make the postseason roster . About time they made a good decision*

we attached *time* via the *parataxis* relation to *make*.

In many cases, the breaks between utterances are not marked by punctuation; for example, the first of the following examples contains three separate utterances (*So you play the battle monsters*, *are they fun*, *will I be able to keep up?*). In the second example, the breaks between utterances are even less clear.

> *So you play the battle monsters are they fun will I be able to keep up ?*

> *I Aint Got A Girl Married 2 Tha Streets Or That Be In Them She Sumthin New #800*

Where sentence boundaries were not marked, we used our best judgment in determining them.

In messages with multiple utterances, words attached with the *discourse* relation were attached to the head of the utterance on the left; for example, in

> *I'm still sooooooo tired ! I just want to go back to sleep ! EMOJI*

EMOJI is attached to *want*, the head of the second utterance.

4.4.3.4 Discourse markers versus multiple utterances

We observed many cases where a message began with an utterance that was discourse marker-like. We chose to treat them as full separate utterances (and not as discourse markers) if they were separated from the rest of the utterance by sentence-ending punctuation. For example, in the case of

> *ATMENTION oh shit ! Didnt do my homework ! Thought it was day 3 !*

we annotated *oh* to be the root of the sentence. Where there was no such separating punctuation, we treated them as discourse markers; for example, with

> *Ohhhh Hell Naw Dis Bitch Shay Got My Last Name *Johnson**

we annotated *Got* as the root of the message, with *Ohhhh*, *Hell*, and *Naw* attached via the *discourse* relation.

4.4.3.5 Hashtags

We largely treated hashtags as full separate utterances; for example, in the following tweet, we annotated *#tweetliketheoppositegender* to be the root of the sentence, and attached *#bigoleboner* to *#tweetliketheoppositegender* with the parataxis relation.

> *#tweetliketheoppositegender Oh damn . I love it when all these bitches wear yoga pants . #bigoleboner*

In contrast, with

> *My boss fired somebody on site n front of everybody #She's A Savage*

there appear to be two utterances, *My boss...everybody* and *#She's A Savage*. In the latter, while *#She's* is marked with a hashtag, it functions as the subject of the second utterance. This is in contrast to hashtags such as *#tweetliketheoppositegender* above; while this hashtag does appear to have an internal syntactic structure, it has not been separated out into its constituent words by the author.

4.4.3.6 Emoticons and emoji

We use *emoticon* to mean an ascii-art-style depiction of an ideogrammic emotive symbol like [[:)]], and *emoji* to mean a specialized character that is rendered as a small image (the same terminology as Owoputi et al. [2013]). In terms of syntactic annotations we treat them the same; in general, we attached them with the *discourse* relation to the head of the nearest preceding (or following, if no preceding) clause, in line with UD's treatment of interjections.

Since emojis can interact poorly with the annotation software we used, each sequence of emojis was converted to an EMOJI token, e.g.

“ ATMENTION : “ ATMENTION : Lil booty girls >>>> y’all winning 😍😍😍😍💦💦💦💦💦👅 *yes lawwwd* 🙌🏼🙌🏼🙌🏼 *””*

was converted to

> *“ ATMENTION : “ ATMENTION : Lil booty girls >>>> y’all winning EMOJI yes lawwwd EMOJI ””*

4.4.3.7 Repeated Words

When words were repeated, we annotated them with the *conj* relation; for example, with

> *Hate hate hate hate hate hate hate having to work there . EMOJI*

we annotated the first instance of *Hate* as the root of the utterance, with each subsequent *hate* attached to the first with the *conj* relation, in accordance with UD’s standard branching structure for conjunctions.

4.4.3.8 Collapsed Phrases

Frequently, phrases were abbreviated, as in the following, where *on my way* is abbreviated to *omw*:

> *Now omw to get my hair done for coronation*

For each of these cases, we used our best judgment; in this case, we annotated *omw* as the root of the sentence (see main text).

A common and challenging type of collapsed phrase are *{the, as} fuck* abbreviations.

> *“ ATMENTION : Two faced af #leavemealone ” ill leave who ever I want alone ! Haha*

> *Ik dis nig saw my stats/tweets why tf he woke me out my sleep*

where *af* and *tf* stand for *as fuck* and *the fuck*, respectively. We generally annotated them with the *advmod* relation.

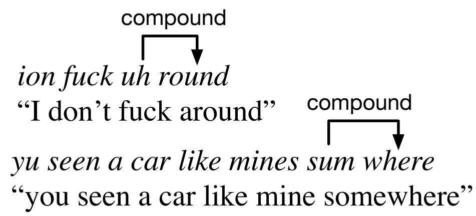


Figure 4.4: Example of words separated into multiple tokens and our proposed annotation.

4.4.3.9 Separated words

We observed multiple cases of words separated into multiple tokens, such as *uh round* and *sum where* in the examples in Figure 4.4. We decided to use UD 2.0's *compound* relation, since they do not appear to be disfluencies or poor editing (which would argue for *goeswith*), and *fixed* is intended for grammaticized fixed expressions (e.g. *because of*), which does not apply. *flat* is an arguable alternative for these cases; based on the UD documentation, and inspecting cases in the UD Treebank, we were unable to make a determination which was more appropriate for these cases.

4.4.3.10 Punctuation

For all punctuation tokens, including quotes, commas, periods, exclamation marks, and question marks, we attach them to the root via the *punct* relation. Emoticons and emojis do not count as punctuation.

4.4.4 Experiments

Models

Our experiments use the following two parsers.

UDPipe [Straka et al., 2016] is a neural pipeline containing a tokenizer, morphological analyzer, tagger, and transition-based parser and is intended to be easily retrainable. The parser attains 80.2% LAS (labeled attachment score) on the UD English treebank with automatically generated POS tags, and was a baseline system used in the CoNLL 2017 Shared Task [Zeman et al., 2017].[7]

Deep Biaffine [Dozat et al., 2017, Dozat and Manning, 2017] is a graph-based parser incorporating neural attention and biaffine classifiers for arcs and labels. We used the version of the parser in the Stanford CoNLL 2017 Shared Task submission, which attained 82.2% LAS on the UD English treebank with automatically generated tags, and achieved the best performance in the task. The model requires pre-trained word embeddings.[8]

4.4.4.1 Experimental setup

We considered a series of experiments within both a cross-domain scenario (§4.4.4.2), where we trained only on UD Treebank data, and an in-domain scenario (§4.4.4.3) using small amounts of our labeled data. We use the parsing systems' default hyperparameters (e.g., minibatch size and learning rate) and the default training/development split of the treebank (both systems perform early stopping based on development set performance).

4.4.4.2 Cross-domain settings

Morpho-Tagger vs. ARK POS tags The UD Treebank contains extensive fine-grained POS and morphological information, on which UDPipe's morphological analyzer and tagging system is originally trained. This rich information should be useful for parsing, but the analyzers may be highly error-prone on out-of-domain, non-standard Twitter data, which may contribute to poor parsing performance. We hypothesize that higher quality, even if coarser, POS information should improve parsing.

To test this, we retrain UDPipe in two different settings. We first retrain the parser component with fine-grained PTB-style POS tags and morphological information provided by the tagger component;[9] we call this the *Morpho-Tagger* setting. Second, we retrain the

[7] `https://github.com/ufal/udpipe`

[8] `https://github.com/tdozat/UnstableParser/`

[9] We also retrained this component, to maintain consistency of training and development split. We also remove the universal (coarse) POS tags it produces, replacing them with the same PTB tags.

parser with morphological information stripped and its tags predicted from the ARK Twitter POS tagger [Owoputi et al., 2013], which is both tailored for Twitter and displays a smaller AAL vs MUSE performance gap than traditional taggers [Jørgensen et al., 2016]; we call this the *ARK Tagger* setting.[10] The ARK Tagger's linguistic representation is impoverished compared to Morpho-Tagger: its coarse-grained POS tag system does not include tense or number information, for example.

Synthetic Data Given our knowledge of Twitter- and AAL-specific phenomena that do not occur in the UD Treebank, we implemented a rule-based method to help teach the machine-learned parser these phenomena; we generated synthetic data for three Internet-specific conventions and one set of AAL syntactic features. This is inspired by Scherrer's [2011] rule transforms between Standard and Swiss German. We performed each of the following transformations separately on a copy of the UD Treebank data and concatenated the transformed files together for the final training and development files, so that each final file contained several transformed copies of the original UD Treebank data.

1. *@-mentions, emojis, emoticons, expressions, and hashtags:* For each sentence in the UD Treebank we inserted at least one @-mention, emoji, emoticon, expression (Internet-specific words and abbreviations such as *lol*, *kmsl*, and *xoxo*), or hashtag, annotated with the correct relation, at the beginning of the sentence. An item of the same type was repeated with 50% probability, and a second item was inserted with 50% probability. @-mentions were inserted using the *ATMENTION* token and emojis using the *EMOJI* token. Emoticons were inserted from a list of 20 common emoticons, expressions were inserted from a list of 16 common expressions, and hashtags were sampled for insertion according to their frequency in a list of all hashtags observed in the TwitterAAE corpus.

2. *Syntactically participating @-mentions:* To replicate occurrences of syntactically participating @-mentions, for each sentence in the UD Treebank with at least one token annotated with an *nsubj* or *obj* relation and an *NNP* POS tag, we replaced one at random with the *ATMENTION* token.

[10]We strip lemmas from training and development files for both settings.

3. *Multiple utterances:* To replicate occurrences of multiple utterances, we randomly collapsed pairs of two short sentences ($<$ 15 tokens) together, attaching the root of the second to the root of the first with the *parataxis* relation.

4. *AAL preverbal markers and auxiliaries:* We introduced instances of verbal constructions present in AAL that are infrequent or non-existent in the UD Treebank data. First, constructions such as *going to*, *about to*, and *want to* are frequently collapsed to *gonna*, *bouta*, and *wanna*, respectively; for each sentence with at least one of these constructions, we randomly chose one to collapse. Second, we randomly replaced instances of *going to* with *finna*, a preverbal marker occurring in AAL and in the American South [Green, 2002]. Third, we introduced the auxiliaries *gone* and *done*, which denote future tense and past tense, respectively; for the former, for each sentence containing at least one auxiliary *will*, we replace it with *gone*, and for the latter, for each sentence containing at least one non-auxiliary, non-passive, past-tense verb, we choose one and insert *done* before it. Finally, for each sentence containing at least one copula, we delete one at random.

Word Embeddings Finally, since a tremendous variety of Twitter lexical items are not present in the UD Treebank, we use 200-dimensional word embeddings trained with *word2vec*[11] [Mikolov et al., 2013] on our TwitterAAE dataset from Ch. 2. Before training, we processed the corpus by replacing @-mentions with ATMENTION, replacing emojis with EMOJI, and replacing sequences of more than two repeated letters with two repeated letters (e.g. *partyyyyy* $\rightarrow$ *partyy*). This resulted in embeddings for 487,450 words.

We retrain and compare UDPipe on each of the *Morpho-Tagger* and *ARK Tagger* settings with synthetic data and pre-trained embeddings, and without. We additionally retrain Deep Biaffine with and without synthetic data and embeddings.[12]

[11] `https://github.com/dav/word2vec`

[12] As the existing implementation of Deep Biaffine requires pre-trained word embeddings, for the Deep Biaffine baseline experiments we use the CoNLL 2017 Shared Task 100-dimensional embeddings that were pretrained on the English UD Treebank.

4.4.4.3 In-domain training

We additionally investigate the effects of small amounts of in-domain training data from our dataset. We perform 2-fold cross-validation, randomly partitioning our dataset into two sets of 250 tweets. We compare two different settings (all using the UDPipe *ARK Tagger* setting):

Twitter-only To explore the effect of training with Twitter data alone, for each set of 250 we trained on that set alone, along with our Twitter embeddings, and tested on the remaining 250.

UDT+Twitter To explore the additional signal provided by the UD Treebank, for each set of 250 we trained on the UD Treebank concatenated with that set (with the tweets upweighted to approximately match the size of the UD Treebank, in order to use similar hyperparameters) and tested on the remaining 250.

4.4.5 Results and analysis

In our evaluation, we ignored punctuation tokens (labeled with *punct*) in our LAS calculation.

Effects of cross-domain settings

Model	LAS
(1) UDPipe, Morpho-Tagger, UDT	50.5
(2) + Twitter embeddings	53.9
(3) + synthetic, Twitter embeddings	58.9
(4) UDPipe, ARK Tagger, UDT	53.3
(5) + Twitter embeddings	58.6
(6) + synthetic, Twitter embeddings	64.3
Deep Biaffine, UDT	
(7) + CoNLL MUSE embeddings	62.3
(8) + Twitter embeddings	63.7
(9) + synthetic, Twitter embeddings	65.0

Table 4.2: Results from cross-domain training settings.

Model	LAS
(10) UDPipe, Twitter embeddings	62.2
(11) + UDT	70.3
(12) + UDT, synthetic	68.7

Table 4.3: Results from in-domain training settings (with the *ARK Tagger* setting, see §4.4.4.3).

***Morpho-Tagger* vs. ARK Tagger** As hypothesized, UDPipe's *ARK Tagger* setting outperformed the *Morpho-Tagger* across all settings, ranging from a 2.8% LAS improvement when trained only on the UD Treebank with no pre-trained word embeddings, to 4.7% and 5.4% improvements when trained with Twitter embeddings and both Twitter embeddings and synthetic data, respectively. The latter improvements suggest that the *ARK Tagger* setup is able to take better advantage of Twitter-specific lexical information from the embeddings and syntactic patterns from the synthetic data. Table 4.2 shows the LAS for our various settings.

After observing the better performance of the *ARK Tagger* setting, we opted not to retrain the Deep Biaffine parser in any *Morpho-Tagger* settings due to the model's significantly longer training time; all our Deep Biaffine results are reported for models trained with an *ARK Tagger* setting.

Synthetic data and embeddings We observed that synthetic data and Twitter-trained embeddings were independently helpful; embeddings provided a 1.4–5.3% boost across the UDPipe and Deep Biaffine models, while synthetic data provided a 1.3–5.7% additional boost (Table 4.2).

UDPipe vs. Deep Biaffine While the baseline models for UDPipe and Deep Biaffine are not directly comparable (since the latter required pre-trained embeddings), in the Twitter embeddings setting Deep Biaffine outperformed UDPipe by 5.1%. However, given access to both synthetic data and Twitter embeddings, UDPipe's performance approached that of Deep Biaffine.

Model	AA LAS	WH LAS	Gap
(1) UDPipe, Morpho-Tagger	43.0	57.0	14.0
(2) + Twitter embeddings	45.5	61.2	15.7
(3) + synthetic, Twitter embeddings	50.7	66.2	15.5
(4) UDPipe, ARK Tagger	50.2	56.1	5.9
(5) + Twitter embeddings	54.1	62.5	8.4
(6) + synthetic, Twitter embeddings	59.9	68.1	8.2
Deep Biaffine, ARK Tagger			
(7) + CoNLL MUSE embeddings	56.1	67.7	11.6
(8) + Twitter embeddings	58.7	66.7	8.0
(9) + synthetic, Twitter embeddings	59.9	70.8	10.9

Table 4.4: AA- and white-aligned tweets' labeled attachment scores for UD Treebank-trained models; *Gap* is the WH − AA difference in LAS.

Perhaps surprisingly, training with even limited amounts of in-domain training data aided in parsing performance; training with just in-domain data produced an LAS comparable to that of the baseline Deep Biaffine model, and adding UD Treebank data further increased LAS by 8.1%, indicating that they independently provide critical signal.

AAL/MUSE performance disparity

For each model in each of the cross-domain settings, we calculated the LAS on the 250 AA- and 250 white-aligned tweets. We observed clear disparities in performance between the two sets of tweets, ranging from 5.9% to 15.7% (Table 4.4). Additionally, across settings, we observed several patterns.

First, the UDPipe *ARK Tagger* settings produced significantly smaller gaps (5.9–8.4%) than the corresponding *Morpho-Tagger* settings (14.0–15.7%). Indeed, most of the performance improvement of the *ARK Tagger* setting comes from the AA-aligned tweets; the LAS on the AA-aligned tweets jumps 7.2–9.2% from each *Morpho-Tagger* setting to the corresponding *ARK Tagger* setting, compared to differences of −0.9–1.9% for the WH tweets.

Second, the Deep Biaffine *ARK Tagger* settings produced larger gaps (8.0–11.6%) than the UDPipe *ARK Tagger* settings, with the exception of the embeddings-only setting.

We also observed the surprising result that adding Twitter-trained embeddings and synthetic data, which contains both Twitter-specific and AAL-specific features, increases the

	Morpho-Tagger			ARK Tagger			
Relation	**AA Recall**	**WH Recall**	**Gap**	**AA Recall**	**WH Recall**	**Gap**	**Reduction**
compound	36.4	71.2	34.8	42.4	72.9	30.5	4.4
obl:tmod	25.0	51.7	26.7	43.8	55.2	11.4	**15.3**
nmod	28.6	54.4	25.8	45.7	51.5	5.8	**20.1**
cop	56.5	82.1	25.6	65.2	79.1	13.9	**11.7**
obl	41.4	65.4	24.0	56.8	62.5	5.7	**18.3**
cc	56.9	79.0	22.1	78.5	82.7	4.3	**17.8**
ccomp	33.3	54.2	20.8	40.5	54.2	13.7	7.1
obj	61.3	81.5	20.2	72.8	83.5	10.7	9.5
case	60.5	79.8	19.3	75.2	83.4	8.2	**11.1**
det	73.1	90.7	17.5	83.4	92.2	8.8	8.7
advmod	53.8	71.2	17.3	62.9	72.1	9.1	8.2
advcl	31.5	46.8	15.3	25.9	46.8	20.9	-5.6
root	56.4	71.6	15.2	62.8	74.0	11.2	4.0
xcomp	40.0	54.9	14.9	51.2	50.0	1.2	**13.7**
discourse	30.7	44.9	14.2	46.0	51.4	5.4	8.8

Table 4.5: Recall by relation type under UDPipe's *Morpho-Tagger* and *ARK Tagger* settings (+synthetic+embeddings; (3) and (6) from Table 4.4). *Reduction* is the reduction in performance gap (WH - AA) from the *Morpho-Tagger* setting to the *ARK Tagger* setting; bolded numbers indicate a gap reduction of ≥ 10.0.

performance gap across both UDPipe settings. We hypothesize that while UDPipe is able to effectively make use of both Twitter-specific lexical items and annotation conventions within MUSE-like syntactic structures, it continues to be stymied by AAL-like syntactic structures, and is therefore unable to make use of the additional information.

We further calculated recall for each relation type across the AA tweets and WH tweets, and the resulting performance gap, under the UDPipe *Morpho-Tagger* and *ARK Tagger* models trained with synthetic data and embeddings. Table 4.5 shows these calculations for the 15 relation types for which the performance gap was highest and which had at least 15 instances in each of the AA and WH tweet sets, along with the corresponding calculation under the *ARK Tagger* model. The amount by which the performance gap is reduced from the first setting to the second setting is also reported. Of the 15 relations shown, the gap was reduced for 14, and 7 saw a reduction of at least 10%.

Feature	AA Count	WH Count	Example
Dropped copula	44	0	*MY **bestfrienddd mad** at me tho*
Habitual *be*, describing repeated actions	10	0	*fees **be** looking upside my head likee ion kno wat **be** goingg on .* *I kno that clown, u don't **be** around tho*
Dropped possessive marker	5	0	*ATMENTION on Tv...tawkn bout dat **man** gf* *Twink rude lol can't be calling ppl ugly that's **somebody** child lol...*
Dropped 3rd person singular	5	0	*When a female **owe** you sex you don't even wanna have a conversation with her*
Future *gone*	4	0	*she **gone** dance without da bands lol*
Expletive *it*	2	1	***It was** too much goin on in dat mofo .*
Completive *done*	1	0	*damnnn I **done** let alot of time pass by . .*

Table 4.6: Examples of AAL syntactic phenomena and occurrence counts in the 250 AA- and 250 white-aligned tweet sets.

AAL Feature	Morpho-Tagger Baseline	ARK Tagger Baseline	ARK Tagger with Embeddings	ARK Tagger with Synthetic, Embeddings
Lexical Variants	16.3 (13/80)	61.3 (49/80)	63.8 (51/80)	57.5 (46/80)
Dropped copula	54.5 (24/44)	70.5 (31/44)	61.4 (27/44)	68.2 (30/44)
Habitual *be*	50.0 (5/10)	80.0 (8/10)	90.0 (9/10)	90.0 (9/10)

Table 4.7: Parsing accuracies of syntactic and lexical variations across four UDPipe models.

Finally, we examine the performance of the *ARK Tagger* and *Morpho-Tagger* settings on AAL lexical and syntactic phenomena in our dataset, finding that while the *ARK Tagger* settings outperformed the *Morpho-Tagger* settings, syntactic features such as dropped copulas presented significant challenges to both.

4.4.5.1 Lexical and syntactic analysis of AAL

In this section, we discuss AAL lexical and syntactic variations observed in our dataset, with the aim of providing insight into reduced parsing accuracy on AA-aligned tweets, and the impact of various parser settings on their parsing accuracy.

AAL contains a variety of phonological features which present themselves on Twitter through a number of lexical variations [Green, 2002, Jones, 2015], instances of which occur a total of 80 times in the AA-aligned tweets; notably, none occur in the white-aligned tweets.

We investigated the accuracy of various cross-domain parser settings on these lexical variants; for each of the baseline *Morpho-Tagger*, baseline *ARK Tagger*, *ARK Tagger* with embeddings, and *ARK Tagger* with synthetic data and embeddings models, we counted the number of instances of lexical variants from §4.4.1 for which the model gave the correct head with the correct label.

While the lexical variants challenged all four models, switching from the *Morpho-Tagger* setting to the *ARK Tagger* settings produced significant accuracy increases (Table 4.7). We observed that the greatest improvement came from using the *ARK Tagger* setting with Twitter-trained embeddings; the Twitter-specific lexical information provided by the embeddings was critical to recognizing the variants. Surprisingly, adding synthetic data decreased the model's ability to parse the variants.

We next investigated the presence of AAL syntactic phenomena in our dataset. Table 4.6 shows examples of seven well-documented AAL morphological and syntactic features and counts of their occurrences in our AA and WH tweet sets; again, while several of the phenomena, such as dropped copulas and habitual *be*, occur frequently in our AA tweets, there is only one instance of any of these features occurring in the WH tweet set.

We measured the parsing accuracy for the two most frequent syntactic features, dropped copulas and habitual *be*, across the four models; accuracies are given in Table 4.7. For dropped copulas, we measured parsing correctness by checking if the parser correctly attached the subject to the correct predicate word via the *nsubj* relation; for the first example in Table 4.6, for example, we considered the parser correct if it attached *bestfrienddd* to *mad* via the *nsubj* relation. For habitual *be*, we checked for correct attachment via the *aux* or *cop* relations as in the first and second examples in Table 4.6, respectively.

As before, we observed significant increases in accuracy moving from the *Morpho-Tagger* to the *ARK Tagger* settings. However, neither adding embeddings nor synthetic data appeared to significantly increase accuracy for these features. From manual inspection, most of the dropped copulas errors appear to arise either from challenging questions (e.g. *AT-*

MENTION what yo number ?) or from mis-identification of the word to which to attach the subject (e.g. *He claim he in love llh*, where *he* was attached to *llh* rather than to *love*).

4.5 Discussion and conclusion

While current neural dependency parsers are highly accurate on MUSE-like text, our analyses suggest that AAL-like text presents considerable challenges due to lexical and syntactic features which diverge systematically from MUSE. While the cross-domain strategies we presented can greatly increase accurate parsing of these features, narrowing the performance gap between AAL- and MUSE-like tweets, much work remains to be done for accurate parsing of even linguistically well-documented features. For applications like opinion analysis and information retrieval, which require equal performance across social groups so that concepts or opinions inferred from groups of authors (e.g., AAL speakers) are not under-counted or under-represented in results returned to a user or analyst, accurate parsing may be critical.

In this and the previous two chapters, we proposed a method for developing a corpus of social media text exhibiting features of AAL, which we used to quantify disparities in performance between AAL-like and MUSE-like text for two types of NLP systems. This work addresses one of the questions we posed in the introduction: How do we concretely quantify harms arising from NLP systems? We emphasize that this work was driven by sociolinguistic insights about AAL, which enabled our use of Census data as distant supervision for our mixed membership model, linguistic validation of the resulting AA-aligned corpus, and analyses of the particular linguistic challenges faced by language identification and dependency parsing tools.

CHAPTER 5

MEASURING BIAS: A SURVEY OF BIAS IN NLP

5.1 Introduction

In the next four chapters, we zoom out from the question of how to concretely quantify harms, and turn to the second, much broader set of questions posed in the introduction: What kinds of undesirable behaviors arise from NLP systems? About which should we be concerned, and why? How do we understand these in connection to already existing injustices? How do we evaluate our approaches to quantifying harms?

A wealth of work examining bias in NLP systems has appeared in recent years, including work examining bias in embedding spaces [Bolukbasi et al., 2016, Caliskan et al., 2017, Gonen and Goldberg, 2019, May et al., 2019, i.a.] as well as in systems developed for a breadth of tasks including language modeling [Lu et al., 2018, Bordia and Bowman, 2019], coreference resolution [Rudinger et al., 2018, Zhao et al., 2018], machine translation [Vanmassenhove et al., 2018, Stanovsky et al., 2019], sentiment analysis [Kiritchenko and Mohammad, 2018], and hate speech/toxicity detection [Park et al., 2018, Dixon et al., 2018], among others. In this chapter, we provide a critical survey of the 146 papers that have emerged in the space of bias in NLP systems before July 2020; to our knowledge, this is the fullest analysis of the existing landscape on bias in NLP.

Throughout this chapter and the remainder of this thesis, we depart from existing work on bias in NLP[1] to take a *critical* perspective. That is, we recognize that NLP systems are developed in particular socio-cultural contexts, and we seek to surface and interrogate the assumptions and values—particularly about language and speakers—embedded in the development, deployment, and analyses (of bias or otherwise) of NLP systems. As we will

[1] With few exceptions; see for example Cao and Daumé [2019] and McBain-Ashfield and Millar [2020].

explore in much greater length in Chapters 6 and 7, NLP systems (as with all technologies) are developed in contexts of profoundly unjust social arrangements—unjust distributions of economic resources and political power—and we aim to re-orient the study of bias in NLP towards exploring how NLP systems might reproduce or challenge these arrangements.

5.2 A critical analysis of bias in NLP

Although the papers described above have laid vital groundwork by illustrating some of the ways that NLP systems can be harmful, the majority of them fail to engage critically with what constitutes bias in the first place. Despite the fact that analyzing bias is an inherently normative process—in which some system behaviors are deemed good and others harmful—papers on bias in NLP systems are rife with unstated assumptions about what kinds of system behaviors are harmful, in what ways, to whom, and why. Indeed, the term bias (or "gender bias" or "racial bias") is used to describe a wide range of system behaviors, even though they may be harmful in different ways, to different groups, or for different reasons. Even papers analyzing bias in NLP systems developed for the same task often conceptualize it differently.

For example, the following system behaviors are all understood to be self-evident statements of "racial bias": (a) embedding spaces in which embeddings for names associated with African Americans are closer (compared to names associated with European Americans) to unpleasant words than pleasant words [Caliskan et al., 2017]; (b) sentiment analysis systems yielding different intensity scores for sentences containing names associated with African Americans and sentences containing names associated with European Americans [Kiritchenko and Mohammad, 2018]; and (c) toxicity detection systems scoring tweets containing features associated with African American Language as more offensive than tweets without these features [Davidson et al., 2019, Sap et al., 2019]. Moreover, some of these papers focus on "racial bias" expressed in written text, while others focus on "racial bias" against authors. This use of imprecise terminology obscures these important differences.

Here, we perform a critical analysis of the papers described above, finding that their motivations are often vague and inconsistent. Many lack any normative reasoning for why the system behaviors that are described as bias are harmful, in what ways, and to whom. Moreover, the vast majority of these papers do not engage with the relevant literature outside of NLP to ground normative concerns when proposing quantitative techniques for measuring or mitigating bias. As a result, we find that many of these techniques are poorly matched to their motivations, and are not comparable to one another.

5.2.1 Method for gathering papers

Our survey includes all papers known to us analyzing bias in NLP systems—146 papers in total. We omit papers about speech, for which there is a growing body of work [Vergyri et al., 2010, Lehr et al., 2014, Tatman, 2017, Garnerin et al., 2019, Koenecke et al., 2020, i.a.] restricting our survey to papers about written text only. To identify the 146 papers, we first searched the ACL Anthology[2] for all papers with the keywords "bias" or "fairness" that were made available prior to July 2020. We retained all papers about social bias, and discarded all papers about other definitions of the keywords (e.g., hypothesis-only bias, inductive bias, media bias). We also discarded all papers using bias in NLP systems to measure social bias in text or the real world [e.g., Garg et al., 2018].

To ensure that we did not exclude any relevant papers without the keywords "bias" or "fairness," we also traversed the citation graph of our initial set of papers, retaining any papers analyzing bias in NLP systems that are cited by or cite the papers in our initial set. Finally, we manually inspected any papers analyzing bias in NLP systems from leading machine learning, human-computer interaction, and web conferences and workshops, such as ICML, NeurIPS, AIES, FAccT, CHI, and WWW, along with any relevant papers that were made available in the "Computation and Language" and "Computers and Society" categories on arXiv prior to July 2020, but found that they had already been identified via our traversal of the citation graph. In Table 5.1, we provide a breakdown of the NLP tasks covered by the papers; a full list of papers and the tasks on which they focus is provided in Table A.3 in the appendix. We note that counts do not sum to 146, because some papers cover multiple

[2] https://www.aclweb.org/anthology/

NLP task	Papers
Embeddings (type-level or contextualized)	54
Coreference resolution	20
Language modeling or dialogue generation	17
Hate-speech detection	17
Sentiment analysis	15
Machine translation	8
Tagging or parsing	5
Surveys, frameworks, and meta-analyses	20
Other	22

Table 5.1: The NLP tasks covered by the 146 papers.

tasks. For example, a paper might test the efficacy of a technique for mitigating bias in embedding spaces in the context of sentiment analysis.

We read each of the papers with the goal of categorizing their motivations and their proposed quantitative techniques for measuring or mitigating bias. We use a previously developed taxonomy of harms for this categorization, which differentiates between so-called *allocational* and *representational* harms [Barocas et al., 2017, Crawford, 2017]. Allocational harms arise when an automated system allocates resources (e.g., credit) or opportunities (e.g., jobs) unfairly to different social groups; representational harms arise when a system (e.g., a search engine) represents some social groups in a less favorable light than others, demeans them, or fails to recognize their existence altogether.[3] Adapting and extending this taxonomy, we categorize the 146 papers' motivations and techniques into the following categories:

- *Allocational harms.*
- *Representational harms:*[4]

[3]We will motivate and expand this taxonomy in Ch. 7.

[4]We grouped several types of representational harms into two categories to reflect that the main point of differentiation between the 146 papers' motivations and proposed quantitative techniques for measuring or mitigating bias is whether or not they focus on stereotyping. Among the papers that do not focus on stereotyping, we found that most lack sufficiently clear motivations and techniques to reliably categorize them further.

Category	Papers: Motivation	Papers: Technique
Allocational harms	30	4
Stereotyping	50	58
Other representational harms	52	43
Questionable correlations	47	42
Vague/unstated	23	0
Surveys, frameworks, and meta-analyses	20	20

Table 5.2: The categories into which the 146 papers fall.

  - *Stereotyping* that propagates negative generalizations about particular social groups.
  - Differences in *system performance* for different social groups, language that *misrepresents* the distribution of different social groups in the population, or language that is *denigrating* to particular social groups.

- *Questionable correlations* between system behavior and features of language that are typically associated with particular social groups.
- *Vague descriptions* of bias (or gender bias or racial bias) or *no description* at all.
- *Surveys, frameworks, and meta-analyses.*

In Table 5.2 we provide counts for each of the six categories listed above. Again, we note that the counts do not sum to 146, because some papers state multiple motivations, propose multiple techniques, or propose a single technique for measuring or mitigating multiple harms.

5.2.2 Findings

Categorizing the 146 papers' motivations and proposed quantitative techniques for measuring or mitigating bias into the six categories listed above enabled us to identify several commonalities, which we present below, along with illustrative quotes. Tables 5.3 and 5.4 contain examples of the papers' motivations and techniques across a range of different NLP tasks.

NLP task	Stated motivation	Categories Motivations	 Techniques
Language modeling [Bordia and Bowman, 2019]	*"Existing biases in data can be amplified by models and the resulting output consumed by the public* ***can influence them****,* ***encourage and reinforce harmful stereotypes****, or* ***distort the truth****. Automated systems that depend on these models can* ***take problematic actions*** *based on biased profiling of individuals."*	Allocational harms, stereotyping	Questionable correlations
Sentiment analysis [Kiritchenko and Mohammad, 2018]	*"Other biases can be inappropriate and result in* ***negative experiences*** *for some groups of people. Examples include,* ***loan eligibility and crime recidivism prediction systems****...and* ***resumé sorting systems*** *that believe that men are more qualified to be programmers than women (Bolukbasi et al., 2016). Similarly, sentiment and emotion analysis systems can also perpetuate and accentuate inappropriate human biases, e.g., systems that consider* ***utterances from one race or gender to be less positive*** *simply because of their race or gender, or customer support systems that* ***prioritize a call from an angry male over a call from the equally angry female****."*	Allocational harms, other representational harms (system performance differences w.r.t. text written *by* different social groups)	Questionable correlations (differences in sentiment intensity scores w.r.t. text *about* different social groups)
Machine translation [Cho et al., 2019]	*"[MT training] may incur an association of gender-specified pronouns (in the target) and gender-neutral ones (in the source) for lexicon pairs that frequently collocate in the corpora. We claim that this kind of phenomenon seriously threatens the fairness of a translation system, in the sense that it* ***lacks generality*** *and* ***inserts social bias*** *to the inference. Moreover, the input is* ***not fully correct*** *(considering gender-neutrality) and* ***might offend the users*** *who expect fairer representations."*	Questionable correlations, other representational harms	Questionable correlations
Machine translation [Stanovsky et al., 2019]	*"Learned models exhibit social bias when their training data* ***encode stereotypes not relevant for the task, but the correlations are picked up anyway****."*	Stereotyping, questionable correlations	Stereotyping, other representational harms (system performance differences), questionable correlations

Table 5.3: Examples of the categories into which the papers' motivations and proposed quantitative techniques for measuring or mitigating bias fall. Bold text in the quotes denotes the content that yields our categorizations.

NLP task	Stated motivation	Categories: Motivations	Categories: Techniques
Type-level embeddings [Zhao et al., 2018]	*"However, embeddings trained on human-generated corpora have been demonstrated to inherit* ***strong gender stereotypes*** *that reflect social constructs....Such a bias substantially affects* ***downstream applications****....This concerns the practitioners who use the embedding model to build* ***gender-sensitive applications such as a resume filtering system or a job recommendation system*** *as the automated system may discriminate candidates based on their gender, as reflected by their name. Besides, biased embeddings may implicitly affect downstream applications used in our daily lives. For example, when searching for 'computer scientist' using a search engine...a search algorithm using an embedding model in the backbone tends to* ***rank male scientists higher than females'*** *[sic],* ***hindering women from being recognized*** *and further exacerbating the gender inequality in the community."*	Allocational harms, stereotyping, other representational harms	Stereotyping
Type-level and contextualized embeddings [May et al., 2019]	*"[P]rominent word embeddings such as word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) encode systematic biases against women and black people (Bolukbasi et al., 2016; Garg et al., 2018), implicating many NLP systems in scaling up social injustice."*	Vague	Stereotyping
Dialogue generation [Liu et al., 2019]	*"Since the goal of dialogue systems is to talk with users...if the systems show discriminatory behaviors in the interactions, the user experience will be adversely affected. Moreover, public commercial chatbots can get resisted for their improper speech."*	Vague/unstated	Stereotyping, other representational harms, questionable correlations

Table 5.4: Examples of the categories into which the papers' motivations and proposed quantitative techniques for measuring or mitigating bias fall, continued.

5.2.2.1 Motivations

Papers state a wide range of motivations, multiple motivations, vague motivations, and sometimes no motivations at all. We find that the papers' motivations span all six categories, with several papers falling into each one. Appropriately, papers that provide surveys or frameworks for analyzing bias in NLP systems often state multiple motivations [Hovy and Spruit, 2016, Bender, 2019, Sun et al., 2019, Rozado, 2020, Shah et al., 2020, i.a.]. However, as the examples in Tables 5.3 and 5.4 illustrate, many other papers (33%) do so as well. Some papers (16%) state only vague motivations or no motivations at all. For example, as

> [N]o human should be discriminated on the basis of demographic attributes by an NLP system. [Kaneko and Bollegala, 2019]

> [P]rominent word embeddings [...] encode systematic biases against women and black people [...] implicating many NLP systems in scaling up social injustice. [May et al., 2019]

These examples leave unstated what it might mean for an NLP system to "discriminate," what constitutes "systematic biases," or how NLP systems contribute to "social injustice" (itself undefined).

Papers' motivations sometimes include no normative reasoning. We find that some papers (32%) are not motivated by any apparent normative concerns, often focusing instead on concerns about system performance. For example, the first quote below includes normative reasoning—namely that models should not use demographic information to make predictions—while the other focuses on learned correlations impairing system performance.

> In [text classification], models are expected to make predictions with the semantic information rather than with the demographic group identity information (*e.g.,* 'gay', 'black') contained in the sentences. [Zhang et al., 2020]

> An over-prevalence of some gendered forms in the training data leads to translations with identifiable errors. Translations are better for sentences involving men and for sentences containing stereotypical gender roles. [Saunders and Byrne, 2020]

Even when papers do state clear motivations, they are often unclear about why the system behaviors that are described as bias are harmful, in what ways, and to whom. We find that even papers with clear motivations often fail to explain what kinds of system behaviors are harmful, in what ways, to whom, and why. For example,

> Deploying these word embedding algorithms in practice, for example in automated translation systems or as hiring aids, runs the serious risk of perpetuating problematic biases in important societal contexts. [Brunet et al., 2019]

> [I]f the systems show discriminatory behaviors in the interactions, the user experience will be adversely affected. [Liu et al., 2019][5]

These examples leave unstated what "problematic biases" or non-ideal user experiences might look like, how the system behaviors might result in these things, and who the relevant stakeholders or users might be. In contrast, we find that papers that provide surveys or frameworks for analyzing bias in NLP systems often name who is harmed, acknowledging that different social groups may experience these systems differently due to their different relationships with NLP systems or different social positions. For example, Ruane et al. [2019] argue for a "deep understanding of the user groups [sic] characteristics, contexts, and interests" when designing conversational agents.

Papers about NLP systems developed for the same task often conceptualize bias differently. Even papers that cover the same NLP task often conceptualize bias in ways that differ substantially and are sometimes inconsistent. Rows 3 and 4 of Table 5.3 contain machine translation papers with different conceptualizations of "bias," leading to different proposed techniques, while rows 1 and 2 of Table 5.4 contain papers on bias in embedding spaces that state different motivations, but propose techniques for quantifying stereotyping.

Papers' motivations conflate allocational and representational harms. We find that the papers' motivations sometimes (16%) name immediate representational harms, such as stereotyping, alongside more distant allocational harms, which, in the case of stereotyping,

[5] Since we performed our analysis, this paper has been published at COLING; we include the reference to the arXiv version we analyzed in the citation.

are usually imagined as downstream effects of stereotypes on résumé filtering. Many of these papers use the imagined downstream effects to justify focusing on particular system behaviors, even when the downstream effects are not measured. Papers on bias in embedding spaces are especially likely to do this because embeddings are often used as input to other systems:

> However, none of these papers [on embeddings] have recognized how blatantly sexist the embeddings are and hence risk introducing biases of various types into real-world systems. [Bolukbasi et al., 2016]

> It is essential to quantify and mitigate gender bias in these embeddings to avoid them from affecting downstream applications. [Zhou et al., 2019]

In contrast, papers that provide surveys or frameworks for analyzing bias in NLP systems treat representational harms as harmful in their own right. For example, Mayfield et al. [2019] and Ruane et al. [2019] cite the harmful reproduction of dominant linguistic norms by NLP systems (a point to which we return in Chapter 7), while Bender [2019] outlines a range of harms, including seeing stereotypes in search results and being made invisible to search engines due to language practices.

5.2.2.2 Techniques

Papers' techniques are not well grounded in the relevant literature outside of NLP. Perhaps unsurprisingly given that the papers' motivations are often vague, inconsistent, and lacking in normative reasoning, we also find that the papers' proposed quantitative techniques for measuring or mitigating bias do not effectively engage with the relevant literature outside of NLP. Papers on stereotyping are a notable exception: the Word Embedding Association Test [Caliskan et al., 2017] draws on the Implicit Association Test [Greenwald et al., 1998] from the social psychology literature, while several techniques operationalize the well-studied "Angry Black Woman" stereotype [Kiritchenko and Mohammad, 2018, May et al., 2019, Tan and Celis, 2019] and the "double bind" faced by women [May et al., 2019, Tan and Celis, 2019], in which women who succeed at stereotypically male tasks are perceived to be less likable than similarly successful men [Heilman et al., 2004]. Tan and Celis [2019] also examine the compounding effects of race and gender, drawing on Black feminist scholarship on intersectionality [Crenshaw, 1989].

Papers' techniques are poorly matched to their motivations. We find that although 21% of the papers include allocational harms in their motivations, only four papers actually propose techniques for measuring or mitigating allocational harms.

Papers focus on a narrow range of potential sources of "bias." We find that nearly all of the papers focus on system predictions as the potential sources of "bias," with many additionally focusing on bias in datasets (e.g., differences in the number of gendered pronouns in the training data [Zhao et al., 2019]). Most papers do not interrogate the normative implications of other decisions made during the development and deployment lifecycle—perhaps unsurprising given that their motivations sometimes include no normative reasoning. A few papers are exceptions, illustrating the impacts of task definitions, annotation guidelines, and evaluation metrics: Cao and Daumé [2019] study how folk conceptions of gender [Keyes, 2018] are reproduced in coreference resolution systems that assume a strict gender dichotomy, thereby maintaining cisnormativity; Sap et al. [2019] focus on the effect of priming annotators with information about possible dialectal differences when asking them to apply toxicity labels to sample tweets, finding that annotators who are primed are significantly less likely to label tweets containing features associated with African American Language as offensive.

5.3 Conclusion

In this chapter, we demonstrated that the existing literature on bias provides many different, often inconsistent conceptualizations of bias. As such, we argue that a normative foundation for conceptualizing harms arising from NLP systems is needed; moreover, we argue that such a foundation must be grounded in literature outside NLP that examines how language is implicated in harm more broadly. In the next chapter, we draw on literature across a number of disciplines to provide such a foundation.

CHAPTER 6

LANGUAGE AND JUSTICE

6.1 Introduction

In this chapter, we provide a normative foundation for reasoning about harms arising from NLP systems that we have shown is largely absent from the current literature by examining the relationships between language and justice across literature from sociolinguistics, linguistic anthropology, sociology, education, and more.

6.2 Social justice

Outside of NLP, there is a vibrant, growing body of work on bias in automated systems that understands biases not as arising from automated systems alone, but as products of larger patterns of injustice. That is, automated systems do not invent injustices out of whole cloth, but transform existing ones arising from systemic and institutionalized racism, sexism, cis heteronormativity, ableism, and so forth. Here, we briefly sketch some foundational work on these social processes as well as the emerging literature at the intersection of technology and social (in)justice.

Most fundamentally, the social processes mentioned above—racism, sexism, and so forth—create and maintain profoundly unjust distributions of economic resources and political power. These social processes are understood across a number of disciplines not merely as instances of individual ill will, but as systemic and institutionalized [Feagin and Ducey, 2000, McCann and Kim, 2013, Delgado and Stefancic, 2017]. In her foundational work, Collins [2000] theorizes four domains of power through which some social groups' subordination is maintained; the *structural*, *disciplinary*, *hegemonic*, and *interpersonal* domains. Where the interpersonal domain of power is perhaps the most familiar, constituted by individuals' daily interpersonal interactions and routines, the structural and disciplinary are constituted by institutional decisions, policies, and practices, for example through laws preventing Black people from exercising their right to vote (structural) and bureaucratic practices that maintain power relations, for example through surveillance (disciplinary). Finally, the hegemonic domain legitimizes these social arrangements through cultural systems of ideas, which are reproduced by a range of social institutions and normalized as common sense. Such an idea is the stereotype of Black women as lazy and immoral "welfare mothers" (one of many *controlling images*) which "shifts the angle of vision away from structural sources of poverty and blames the victims themselves" [Collins, 2000, Ch. 4].

Importantly, these social processes do not function independently but as interlocking systems experienced simultaneously, an insight that has been recognized across Black feminist scholarship. In its influential statement, the Combahee River Collective wrote: "We also often found it difficult to separate race from class from sex oppression because in our lives they are most often experienced simultaneously" [Combahee River Collective Statement, 1977]. Throughout, the document articulates the unique struggles experienced by Black women precisely because of their existence at the intersection of multiple social categories, and the inability of either the civil rights movement (against racism) or the feminist movement (against sexism) for addressing these struggles, dominated as they were by Black men and white women, respectively. Later, Kimberlé Crenshaw introduced the term *intersectionality*; her pioneering analyses illustrate the experiences of Black women as the "product of intersecting patterns of racism and sexism" and demonstrate the limitations of single-axis analyses in anti-discrimination law [Crenshaw, 1989, 1991]. Collins [2000] provides a related conceptual model, the *matrix of domination*, which "helps us think about how power, oppression, resistance, privilege, penalties, benefits, and harms are systematically distributed" [Costanza-Chock, 2018].

Abnormal justice In a different academic tradition, Fraser [2008] examines the social assumptions underpinning justice claims. Such assumptions, which are contested, include not only who can make such claims but also what kinds of claims can be admitted. Historically, these latter have been "(economic) claims for redistribution", which Fraser terms *maldistribution*. According to Fraser, however, two other types of justice claims have arisen more recently: claims for "legal or cultural recognition" and claims for political representation. The former addresses the concern that many justice claims center not only on injustices of distribution of economic goods, but also on injustices of social standing; due to historically unequal social hierarchies, many people are denied equal recognition and respect. Fraser terms this type of injustice *misrecognition*. Finally, political representation addresses concerns about the "fair terms of political representation and equal voice", recognizing that injustice claims may additionally center on a lack of democratic processes or decision procedures, which may render some people voiceless. Such injustices, originating in the political organization and procedures of a society as opposed to entrenched hierarchies of status, are termed *misrepresentation*. In order to allow claims of different types of injustices to be evaluated according to a single principle, Fraser additionally proposes a normative principle: the principle of parity of participation. "According to this principle, justice requires social arrangements that permit all to participate as peers in social life. On the view of justice as participatory parity, overcoming injustice means dismantling institutionalized obstacles that prevent some people from participating on a par with others, as full partners in social interactions" [Fraser, 2008].

6.2.1 Social justice and technology

Research in science and technology studies has long explored the social processes, practices, discourses, and institutions through which technologies are developed, and how technologies shape the world in turn, for instance Winner's [1980] argument that technologies embody social relations, Agre's [1997] analysis of AI researchers' self-understanding and discursive practices, and Bowker and Star's [2000] analysis of the processes through which classification systems come to be, showing that they are neither neutral nor inevitable, but the product of social processes. More recently, work at the intersection of social justice and technology that acknowledges and draws upon this and the scholarship we described above is beginning to emerge. This work understands automated systems' bad behaviors not as aberrations or glitches, but as glimpses of the unjust social arrangements conditioning the systems' development [Benjamin, 2019]. Such work has produced a number of critiques of existing algorithmic fairness approaches, which we sketch briefly.

First, algorithmic fairness approaches often fail to account for *historical* bias [Suresh and Guttag, 2019], where data is generated via unjust processes—for instance, in the case of recidivism prediction, by the over-policing of Black communities [Richardson et al., 2019]—which cannot be addressed by technical solutions aiming to gather more data or model it more effectively. Algorithms that satisfy group fairness (e.g., demographic parity) often disregard the reality that groups are not all treated the same [Hanna et al., 2020] as well as significant within-group differences [Kasy and Abebe, 2020] long recognized by Black feminist scholarship. Moreover, their inattention to larger legitimating ideologies and relations of power leaves these algorithms unable to address the assumptions and values (e.g., beliefs about criminality, creditworthiness, and hiring potential) that affect system development and deployment, power relations between technologists and affected communities, and questions of who gets to be at the table in the first place. As a result, existing research disproportionately focuses on individual bad features, models, and datasets rather than the larger ideologies, institutions, policies, and practices that produce them [Hoffmann, 2019]. Even efforts aimed at inclusive technologies maintain the assumption that technology ought to be a solution in the first place [Hoffmann, 2020], failing to ask whether some systems ought to be built at all, and keeping development and deployment decisions in the hands of technologists [Bennett and Keyes, 2019, Green, 2019, Katell et al., 2020].[1]

In the fairness community, a turn towards adopting structural injustice perspectives and examining systems in their broader sociotechnical context can be seen in the case of facial recognition, as attention has shifted from improving (in pathbreaking work) systems' ability to recognize faces of women and people of color, particularly Black women [Buolamwini and Gebru, 2018], to recognizing the disproportionate burden of surveillance systems on minori-

[1] In recent work, Kasy and Abebe [2020] address power by asking who gets to choose objective functions.

tized communities and asking whether such systems should be built at all.[2][3][4] Other work at this intersection of social justice and technology includes Browne's [2015] analysis of modern surveillance tools as emerging from histories of anti-Black surveillance; Keyes's [2018] analysis of the erasure of transgender people by automated gender recognition systems; Noble's [2018] finding that search engines reproduce anti-Black stereotypes, challenging views of search engines as value-neutral; Costanza-Chock [2018] and Costanza-Chock's [2020] introduction of design justice principles grounded in the concepts of intersectionality and the matrix of domination; Benjamin's [2019] examination of how racial hierarchies impact technology (which in turn, reproduce racial hierarchies); Cifor et al.'s [2019] *Data Manifest-No* "refus[ing] harmful data regimes and commit[ting] to new data futures"; D'Ignazio and Klein's [2020] analysis of how data science sustains (and can be turned to challenge) relations of power; Ogbonnaya-Ogburu et al.'s [2020] introduction of critical race theory for HCI and counterstories as a tool for challenging dominant narratives; Birhane's [2020] analysis of technology companies' exploitation of Africa as *algorithmic colonialism*, drawing parallels to traditional colonialism; Mohamed et al.'s [2020] application of decolonial theory to digital technologies; Birhane and Guest's [2020] call for computational scientists to reckon with the field's harmful histories and dynamics; and the Indigenous Protocol and Artificial Intelligence Position Paper [Lewis, 2020].[5]

Together, this scholarship offers several key insights for understanding the harms arising from automated systems: first, the essential insight that resources and power are indeed distributed unjustly, with profound consequences for oppressed social groups across all facets of life, including criminal justice, education, employment, housing, and health. As does much of the work we have described, we take the position that the normative foundation of

[2] `https://onezero.medium.com/we-must-fight-face-surveillance-to-protect-black-lives-5ffcd0b4c28a`

[3] `https://lpeproject.org/blog/the-second-wave-of-algorithmic-accountability/`

[4] `https://venturebeat.com/2020/08/23/the-term-ethical-ai-is-finally-starting-to-mean-something/`

[5] The Critical Race Digital Studies syllabus provides a reading list at the intersection of critical race and technology studies: `https://criticalracedigitalstudies.com/syllabus/`

efforts to build more just automated systems—the principles that help us reason about what systems are harmful and why, and what systems we ought to be building—must be rooted in a) recognizing and mitigating systems' ability to reproduce these social arrangements, and b) dismantling these social arrangements, in part by developing systems that explicitly challenge these arrangements. Second, automated systems might not only help maintain unjust social arrangements by helping to distribute resources and opportunities unjustly, but might also legitimize and sustain these arrangements by reproducing certain systems of ideas (through the hegemonic domain of power); this, therefore, requires us to be attentive to the ideologies and discourses that systems (re)produce. Third, we should be attentive not only to the output decisions of automated systems, but also to the institutional policies and practices through which systems are produced. Fourth, the work we have sketched insists on centering the lived experiences of those at the margins, particularly those who are multiply marginalized. Finally, this work acknowledges multiple interlocking systems of oppression and recognizes the limitations of single-axis analyses of harm.

Having examined some of the foundational work on unjust social arrangements and emerging analyses of automated systems with respect to those social arrangements, we return our focus to NLP systems and examine the many relationships between language and social arrangements.

6.3 Overview of language and justice

Language is essential to the human experience, participating in so many of our social processes.We therefore argue that in order to understand what harms NLP systems can give rise to, we must first understand the relationships between language and the unjust social arrangements we illustrated above. That is, we need to understand how uses of language, and beliefs about language and speakers, participate in producing, maintaining, and contesting unjust social arrangements. In this section, we aim to illustrate the breadth and depth of literature examining these relationships. We argue that this work provides the necessary social and historical context for understanding the environments in which NLP systems are developed, and it is to this literature that work on bias in NLP should be re-oriented.

Through understanding it, we are better placed to recognize and evaluate behaviors rooted in these mechanisms as they arise from technologies of language, and can identify concretely who may be harmed by such technologies.

In our illustration of these relationships, we draw a distinction between language *by* and language *about* people—that is, we draw on literature examining both language *about* different groups of people, as well as language *by* different groups of people, in order to understand how NLP systems might give rise to harms in both kinds of situations. Following *This Ain't Another Statement! This is a DEMAND for Black Linguistic Justice* [CCCC, 2020], throughout this overview we aim to center the work of scholars who, "informed by their lived experiences," have been the first to identify many connections between language and social arrangements and champion linguistic justice.

6.3.1 Language *about*

Language participates in the construction of social categories in at least three ways. First, language maintains social categories by naming them: "[T]he label content functions to identify a given category of people, and thereby conveys category boundaries and a position in a hierarchical taxonomy" [Beukeboom and Burgers, 2019]. Second, language transmits beliefs about social categories, both through explicit stereotypes and through subtle asymmetries in language use [Macrae et al., 1996, Maass, 1999, Ellemers, 2018, Beukeboom and Burgers, 2019]. Finally, language features themselves become associated with different groups of people and are thereby ascribed social meaning, as they index membership in and "invok[e] ways of belonging to, or characteristics or stances associated with" those groups [Eckert, 2012].

Language performs an essentializing function. By naming groups and transmitting stereotypes, language reinforces the idea that the groups thus named, the distinctions between them, and the properties or behaviors described by the stereotypes are meaningful and describe some underlying reality. For instance, conceptualizations of gender as binary are reinforced linguistically in English through pronoun conventions, where until recently *he* and *she*, but not *they*, were acceptable as singular pronouns, through phrases such as *opposite* or *both genders*, and through administrative processes that deny individuals the choice of a (linguistic) label other than *man* or *woman* (an instance of what Spade [2011] terms *administrative violence*). Stereotypes both create expectations and normalize differential outcomes; for instance, they create gendered expectations about academic achievement, affecting students' performance, and justify gendered differences in outcomes as natural [Ellemers, 2018]. Similarly, "controlling images," such as stereotypes of Black women, which are linguistically and visually transmitted through literature, news media, television, and so forth, provide "ideological justification" for their continued oppression [Collins, 2000, Ch. 4]. Thus language maintains unjust social arrangements by naturalizing them.

Many groups have sought to bring about social changes through changes in language use, disrupting patterns of oppression and marginalization via so-called "gender-fair" language (e.g., by shifting from gendered to gender-neutral defaults) across a range of languages [Sczesny et al., 2016, Menegatti and Rubini, 2017, Kotek et al.], trans-inclusive language [Zimman, 2017], and language that is less dehumanizing towards immigrants (e.g., abandoning the use of the term "illegal" in the U.S. [Rosa, 2019]).

Language performs demeaning and disciplinary functions. Waldron [2012] illustrates how demeaning or denigrating language, such as hate speech, creates an "environmental threat" to targeted groups by reproducing stereotypes and "intimating discrimination and violence," thereby "compromis[ing] the dignity of those at whom it is targeted" and undermining individuals' sense of security in their daily lives [Waldron, 2012, Ch. 1]. Thus, language maintains unjust social arrangements when it is used to undermine targeted group members' sense of dignity, discourage participation in public life, and distribute feelings of security and respect unequally.

The ability to control language use is a function of power. Struggles over language use often arise from dominant social groups' desire to "control both material and symbolic resources"—i.e., "the right to decide what words will mean and to control those meanings"—as was the case with some white speakers' insistence on using offensive place names against the objections of Indigenous speakers [Hill, 2008, Ch. 3].

Language choices shape beliefs and discourses. This effect is most intuitive in the language used to describe social groups; for instance the construction of the social category of (illegal) “alien” has shaped discourse on immigration, “rationaliz[ing] the harsh treatment of persons from other countries” and turning public debate towards immigration enforcement [Johnson, 1996]. But language choices elsewhere can also enable or foreclose different discourses. For instance, Cohn [1987] describes the “technostrategic” language used by defense analysts to describe and develop nuclear strategy, finding that the abstract, weapons-oriented language distanced analysts from the victims of their decisions and made it difficult to discuss peace.

Discursive processes legitimize relations of power. For instance, Bucholtz [2019] identifies five discourse strategies employed in “white public discourse”; as overtly racist language has become unacceptable in public discourse, Bucholtz shows how these strategies, which position whiteness as “beleaguered,” function to maintain existing relations of power. Discourses of colorblindness also maintain power relations by denying the reality of their existence [Bonilla-Silva, 2014]. More recently, Amber Hamilton shows that corporate statements on racial justice released by U.S. tech companies are “reluctant to even use the word ‘race”’ and “rarely nam[e] whiteness,” thereby “obscuring the central role that whiteness and racism play in the injustices Black people endure” [Hamilton, 2020]. As Hamilton observes, these discursive processes refuse to acknowledge the reality of anti-Black injustice and cannot possibly produce meaningful solutions, as they do not even name the structures and mechanisms by which injustice is mantained.

Even the definitions and discourses in language-related research can maintain unjust power relations; for instance, Davis [2017] explores several harmful discursive processes in academic and public language surrounding language endangerment and reclamation, including those that erase “colonial agency” by “minimiz[ing] historical and ongoing causes of language endangerment and dormancy, sometimes to the extent of misattributing agency onto Indigenous communities themselves,” as well as those that frame Indigenous populations as "vanishing" and thereby erase Indigenous resistance and reclamation efforts.[6] Braithwaite [2020] shows how researchers working on sign languages define communities and language with labels that are "exoticizing and objectifying," leading to research which is "insufficiently grounded in the realities and concerns of community members."

Language ideologies are partly produced through metalinguistic discourse. One mechanism by which language ideologies (which we address in more detail below) are (re)produced is through explicit discussion of language and speakers. For instance, tourist guidebooks for destinations in the Caribbean and the Indian Ocean produce "stereotypical and exoticizing views" of both French creole languages and their speakers through descriptions of creoles as picturesque and musical but grammatically simple, and reproduce colonial views of African languages and cultures [Krämer and von Sicard, 2020].

6.3.2 Language *by*

Language is a crucial resource through which people construct identity, which includes large macro social categories such as gender and race as well as "local, ethnographically specific cultural positions" and "temporary and interactionally specific stances and participant roles" [Bucholtz and Hall, 2010]. This continuous co-construction of language and identity—in which speakers draw on linguistic resources to construct identity, thereby ascribing social meaning to particular language features—drives variation in language [Eckert, 2012]. This variation in language, the social meaning that such variation takes on, what language reveals about social relations and structures, and how social relations and structures are produced and contested through language, have been the central concern of sociolinguistics and linguistic anthropology [Duranti, 2004, Meyerhoff, 2019, Craft et al., 2020].

Social categories are highly salient to listeners and shape their inferences. Social information and linguistic perception are tightly connected; listeners infer a great deal of social information from what is heard, and listeners' social expectations even affect what

[6] See Leonard [2011] and Messner [2018] for more on "extinction" or "trauma" narratives and their impacts on Indigenous language reclamation efforts, and De Korne and Leonard [2017] and Leonard [2017] for additional critical examinations of language revitalization discourses and practices.

is heard [Cargile et al., 1994, Gluszek and Dovidio, 2010, Rubin, 2012, Craft et al., 2020]. For instance, listeners use inferred speaker gender to categorize ambiguous sounds [Johnson et al., 1999], and judge statements to be less credible when spoken by non-native speakers than by native speakers [Levi-Ari and Keysar, 2010].

People carry metalinguistic beliefs about language and speakers. These are studied as *language attitudes* and *language ideologies* in sociolinguistics and linguistic anthropology [Silverstein, 1979, Woolard and Schieffelin, 1994, Irvine and Gal, 2000, Kroskrity, 2004, Rosa and Burdick, 2017].[7] One useful definition of language ideologies comes from Irvine [1989], who defines them as "the cultural system of ideas about social and linguistic relationships, together with their moral and political interests." These ideas about language and speakers take many forms; for instance, which language varieties or practices are taken as standard, ordinary, or unmarked? Which are considered correct, prestigious, or appropriate for public use, and which are considered incorrect, uneducated, or offensive [Silverstein, 1996, Milroy and Milroy, 1999, Hill, 2008, Campbell-Kibler, 2009, Preston, 2009, Loudermilk, 2015, Lanehart and Malik, 2018]? Which are rendered invisible [Roche, 2019]? How do some language practices become associated with race (a process called *racialization*) [Charity Hudley, 2017, Alim et al., 2020], and how do assumptions develop about how people racialized as non-white speak [Rosa and Burdick, 2017]? How are language and national identity linked [Rosa and Flores, 2017]? How have languages come to be conceptualized as "fixed entities capable of being counted, systematized, and named" [Severo and Makoni, 2020]? Where and by whom are boundaries between language varieties drawn [Rosa and Burdick, 2017]?

Of course, linguists recognize that no language varieties or practices are inherently better or more correct than others, nor are any intrinsically linked to any particular social meanings. Thus, what the linguistic ideologies framework offers is the perspective that beliefs about language are not really about language at all, but about the social meanings and "moral and political interests" that linguistic forms become mapped to [Rosa and Burdick, 2017];

[7] See Rosa and Burdick [2017] for a discussion on the distinction between language attitudes and ideologies.

for instance, that Mainstream U.S. English is considered standard in the U.S. and other varieties of English are not is a function of socio-political processes, and not of any particular correctness or suitability inherent to MUSE itself.[8] From this perspective, beliefs about language are inextricably intertwined with social, political, and economic arrangements.

Language ideologies justify unjust social arrangements and enable linguistic discrimination. Language ideologies play a vital role in reinforcing and justifying social arrangements [Lippi-Green, 2012, Alim et al., 2016, Charity Hudley, 2017, Rosa and Flores, 2017, Craft et al., 2020]. In what is now the U.S., European colonizers constructed language hierarchies by portraying Indigenous speakers' language practices as linguistically deficient, thereby justifying racial hierarchies and violent colonization [Rosa and Flores, 2017, García, 2019]. Language has remained key to the maintenance of these relations of power, for instance through forced assimilation practices in the U.S., in which Indigenous peoples were "targets of federal policies aimed at eradicating their languages and lifeways" [McCarty and Watahomigie, 1998]. Yet these ideologies and practices have also been consistently contested; for example, Wa Thiong'o [1986] argues for the use of African languages in African literature as a way to resist the colonial imposition of European languages and worldviews, Davis [2017] illustrates Indigenous counter-narratives and strategies that resist colonial and neo-colonial rhetorics surrounding Indigenous language endangerment, and DeGraff [2020] discusses current efforts to develop educational technologies in Kreyòl as a means to resist the social and political exclusion of Kreyòl-speaking Haitians, where French—as the result of colonial domination—has been a primary language of instruction.

Today, persistent views of non-white speakers as deficient translate into material consequences; linguistic discrimination has been widely documented in a range of opportunities and institutions related to citizenship, asylum, employment, education, housing, criminal justice, and the media [Purnell et al., 1999, Lippi-Green, 2012, Rosa and Flores, 2017, Baugh, 2018, Craft et al., 2020]. Across a range of countries, speaking non-standard vari-

[8] By extension, as Rosa and Burdick [2017] point out, inclusion or valorization of stigmatized language varieties does not necessarily mean "improvements in the social circumstances of their users"; we will return to this point in §7.2.

eties is linked to reduced wages, sometimes by amounts comparable to gender wage gaps [Grogger, 2019, Yao and van Ours, 2019, Grogger et al., 2020]. Thus, language ideologies drive and naturalize unjust social arrangements by privileging the language practices of those in power.

Language ideologies are produced and maintained through many practices in many settings. Beliefs about language practices and their speakers are produced and maintained through the practices of the many institutions we described above. One particularly crucial site is in schools, where educational policies and practices shape these beliefs and the outcomes of different speakers. For example, Cushing [2019] and Cushing [2020] show how top-down educational policies provided by the U.K. government, interpreted and implemented by teachers and administrators, create environments in which teachers are empowered and urged to actively police their students' language, penalizing "non-standard" practices. In the U.S., ideologies of "situational codeswitching," in which "non-standard" varieties are viewed as appropriate for some settings and "standard" English for others, provide the message that "students and educators are best served by leaving African-American English at the classroom door—an ideology that can promote internalized racism as well as linguistic insecurity for both Black students and Black educators" [Charity Hudley et al., 2020]. Campbell-Montalvo [2020] demonstrates that practices of Florida schools have resulted in a 19-fold under-counting of Indigenous Mexican languages spoken by students, thereby perpetuating their erasure.

6.3.3 Case study: African American Language

Here, we describe some work covering perception, ideologies, and discrimination in the specific context of African American Language (AAL).

The associations between language practices and racialized groups are highly salient to speakers of American English; for example, Purnell et al. [1999] showed that speakers were able to determine whether single tokens "were produced by an African-American (i.e., using AAVE), a Latino (i.e., using [Chicano English]), or a white male (i.e., using [Standard American English])." But even after decades of sociolinguistic efforts to legitimize AAL, it continues to be viewed as "bad" English and its speakers continue to be viewed as linguistically inadequate—a view called the *deficit* perspective [Alim et al., 2016, Rosa and Flores, 2017].[9] This perspective, which penalizes AAL speakers for not adhering to dominant language practices[10] persists despite demonstrations that AAL is rule-bound and grammatical [Mufwene et al., 1998, Green, 2002], in addition to ample evidence of its speakers' linguistic adroitness [Alim, 2004, Rickford and King, 2016].

This deficit perspective belongs to a broader set of raciolinguistic ideologies [Rosa and Flores, 2017], which also produce and justify discrimination across the range of institutions we described above. For instance, in the judicial system, testimony from AAL speakers may be misunderstood (due to non-native speakers' unfamiliarity) or disbelieved (due to ideologies linking AAL to lack of intelligence, education, or trustworthiness) [Rickford and King, 2016, Jones et al., 2019]. These raciolinguistic ideologies position AAL-speaking (and other non-white) communities as lacking the language "required for complex thinking processes and successful engagement in the global economy," thereby positioning them as needing language intervention, such as language education programs, through which this discrimination and other harms can be reduced if communities accommodate to dominant language practices [Rosa and Flores, 2017]. This perspective naturalizes economic inequities by framing them as the result of speakers' unwillingness or inability to accommodate. Other raciolinguistic ideologies flatten AAL-speaking communities by casting the variety and its speakers as monolithic, erasing the considerable variation of AAL across the U.S., the complexity of speakers' overall language practices, and the wide range of attitudes and beliefs about AAL within communities [Lanehart, 2015].

Resistance and change. The overview above presents a view of dominant raciolinguistic ideologies that maintain and naturalize unjust (specifically, anti-Black) social arrangements. But as important as it is to acknowledge these ideologies, it is equally important not to essentialize racialized speakers as mere victims of these ideologies, which are not monolithic.

[9] As we described above, this perspective holds generally for language practices of speakers racialized as non-white.

[10] And penalizing speakers even when they do [Rosa and Flores, 2017]

Bonilla-Silva [2014]. Therefore, the story of AAL would be incomplete without exploring speakers' agency and resistance. For instance, Krystal Smalls examines racialized language practices[11] in digital spaces, illustrating how these digital spaces function as "white public spaces" where racialized language practices are interpreted as "unrespectable" or "unintelligible" [Smalls, 2019]. Therefore, Smalls reads the use of such practices in digital spaces as "performative acts of emphatic blackness":

> [T]hese young dissenters are not asking for permission to speak and they are not privileging white comfort over black freedom.... Specifically, these usages can be read as refusals of the performances of respectability demanded by white supremacy vis-á-vis white normativity (Simpson 2014). In effect, many of these emphatically black agitators are simply refusing to translate themselves, even though they know some audience members may not find their words or humanity intelligible.

Researchers have also resisted dominant ideologies surrounding Black speakers and language practices, including AAL. In sociolinguistics and linguistic anthropology, a long literature has illustrated the grammaticality and richness of AAL, the linguistic creativity and adroitness of its speakers, and the processes by which its speakers are marginalized [Labov, 1972, Smitherman, 1986, Mufwene et al., 1998, Rickford and Rickford, 2000, Green, 2002, Makoni et al., 2003, Alim, 2004].[12] From this tradition, a literature theorizing language and race is emerging [Chun and Lo, 2015, Alim et al., 2016, Rosa, 2016, Charity Hudley, 2017, Rosa and Flores, 2017, Rosa, 2019, Alim et al., 2020], which draws on social constructivist perspectives to understand both race and language as the ongoing product of socio-political processes—in particular, how "language and race are mutually constituted as social realities" [Alim et al., 2020]—and to understand how raciolinguistic ideologies drive and justify historical and continuing injustices.[13]

---

[11]Smalls' analysis focuses not only on AAL but on many uses of "identifiably black" language, including "African American English, Jamaican Patois, Ghanaian Pidgin, [and] racialized youth slangs."

[12]For more on early conceptualizations of AAL as a single, uniform variety, see Wolfram [2007] and Wolfram [2015].

[13]The introduction [Alim et al., 2020] to the Oxford Handbook of Language and Race [Alim et al., 2020] traces the history and themes of this literature.

Educators have also been central in efforts for linguistic justice. In 1974 the Conference on College Composition and Communication (CCCC) adopted a resolution affirming students' right to "their own patterns and varieties of language—the dialects of their nurture or whatever dialects in which they find their own identity and style" [CCCC, 1974]; more recently, the CCCC's *This Ain't Another Statement! This is a DEMAND for Black Linguistic Justice!* calls for educators to "center Black Language" and "unravel anti-Black linguistic racism" [CCCC, 2020]. Drawing on the lived experiences and insights of speakers of AAL and other racialized language varieties, a growing body of work is developing anti-racist language pedagogy that challenges the deficit perspective, re-imagines language classrooms as "sites that disrupt racial injustice" [Johnson et al., 2017], and decenters whiteness in language teaching [Kynard, 2013, Young et al., 2014, Flores and Rosa, 2015, Johnson et al., 2017, Flores and Chaparro, 2018, Baker-Bell, 2020, Flores, 2020, Gerald, 2020, Martínez and Mejía, 2020]. Such work challenges the practice of teaching MUSE as a value-neutral language variety necessary for employment and higher education, and demonstrates that for racialized people, linguistic "correctness" may be a moving target, as their language practices are stigmatized no matter how much they adhere to dominant language practices.

This work has led to calls to rethink the practices and goals of language-related fields themselves, including calls for an "anthro-political linguistics" [Zentella, 2018] examining the connections between language and structural inequities [Avineri et al., 2019], for a critical examination of race in linguistics [Charity Hudley et al., 2020], for an "antiracist and decolonizing applied linguistics" [Motha, 2020], and for a postcolonial linguistics in which linguists of the Global North are encouraged to "critically engage with both their own analytical traditions and their alternatives" [Levisen and Sippola, 2019].

Although this sketch is necessarily incomplete, we hope that it illustrates the depth of existing scholarship on the connections between language and institutionalized hierarchies of race, and suggests that similar connections may be found in scholarship on language and many other social categories and relations, including language and gender [Ehrlich et al., 2014], language and sexuality [Ehrlich et al., 2014, Zimman and Hall, 2016, Hall and Barrett, 2018], and language and power globally [Alim et al., 2016, García et al., 2017]. Moreover, as a number of fields studying language are beginning to critically examine the ways in which their own practices have upheld unjust social arrangements, their efforts may provide a model for NLP to reckon with its own practices.

6.3.4 Takeaways

We have argued that analyses of bias in NLP systems should be grounded in the relationships between language and social arrangements that such systems participate in maintaining. Therefore, in this section we have introduced literature across sociolinguistics, linguistic anthropology, social psychology, and other fields concerned with language in order to explore the many ways in which uses of language, and beliefs about language and speakers, participate in producing, maintaining, and contesting unjust social arrangements.

As we turn towards thinking about how NLP systems might reproduce (or more hopefully, challenge) these unjust arrangements, we make several observations. First, we suggest that NLP systems are a key site where language ideologies are (re)produced, both through technologists' development and deployment decisions and through system outcomes and user experiences. Rosa and Burdick [2017] point out that traditional work on language ideologies has viewed them as things held by people, and ask:

> [I]n what ways are language ideologies built into the design of emergent technologies, from voice recognition programs to digital orthographies? How do these technologies recognize language, and how do their users embrace or reject their language ideologies? These questions point to the need for new conceptualizations and methodologies that move beyond approaches in which language ideologies are exclusively understood as ideas explicitly expressed by people.

We suggest that work on bias in NLP systems can provide key insights into how ideologies are expressed through technologies. Second, we emphasize that the language literature shows how language-related harms include both material and social or dignitary harms—in our framing in the previous chapter, both *allocational* and *representational* harms. Therefore, work on bias in NLP must attend to how NLP systems can reproduce patterns of linguistic discrimination in allocating resources and opportunities, as well as less readily measurable (but no less significant) harms related to erasure, stereotyping, stigmatization, and disenfranchisement, which the next section will explore. Finally, across the language literature, we see speaker agency and resistance emphasized, highlighting that linguistic justice must include both minoritized language practices and their speakers, and the importance of co-equal participation of speakers as knowledge producers and decision-makers in "ethical" or "just" NLP efforts.

CHAPTER 7

MEASURING BIAS: A TAXONOMY OF HARMS

7.1 Introduction

In this chapter, we argue that in light of the relationships between language and justice surveyed in the last chapter, the space of harms that can arise from NLP systems is much larger than what has been examined in the current literature on bias in NLP. We re-introduce the concept of *representational harms* and propose a taxonomy of such harms grounded in these relationships between language and injustice. We conclude the chapter by proposing a re-orientation of work on bias in NLP towards these relationships and offer guiding research questions focusing on how NLP systems and practices reproduce them.

7.2 Towards a taxonomy of representational harms

What are representational harms? Following Barocas et al. [2017] and Crawford [2017], we consider *representational* harms to be those which arise when a system represents some social groups in a less favorable light than others, demeans them, or fails to recognize their existence altogether. This category of harms contrasts with *allocational* harms—harms that arise when a system allocates resources (e.g., credit) or opportunities (e.g., jobs) unfairly to different social groups—which have been the earliest and primary focus of algorithmic fairness approaches. We do not imagine that this definition captures all non-allocational harms that may arise; for example, analyses of the power dynamics of NLP development ecosystems in which minoritized people may only be able to participate as annotators rather than decision-makers, or of the role of NLP systems in generating text that misleads, radicalizes, or encourages self-harm [McGuffie and Newhouse, 2020], fall outside the scope of representational harms. Instead, we view this distinction as a generative one, as it draws attention towards the social and dignitary dimensions of harm arising from automated systems [Hoffmann, 2019] as harmful in their own right and encourages researchers and practitioners to imagine expansively the harms that can arise.[1] As the literature surveyed in the previous chapter suggests, these dimensions of harm are particularly significant in the context of language and technologies of language.

We begin by introducing **undesirable correlations**, a category of potentially undesirable system behaviors that are often measured by papers on bias in NLP, but which we argue do not by themselves have much normative substance, and which we therefore consider separately.

We then propose a taxonomy of representational harms, each of which is grounded in the previous chapter's account of the relationships between language and social arrangements. We then introduce long-term dynamics which may give rise to further harms. Throughout, we examine how these harms arise both from existing systems as well as from NLP development, deployment, and research practices.

7.2.1 Undesirable correlations

This category of behaviors, **undesirable correlations**, captures NLP system outputs that are correlated with social groups or language associated with social groups in the input, and which may be deemed undesirable by any number of stakeholders, including system designers and users.

Across the literature on bias in NLP, a number of approaches operationalize bias by identifying and quantifying such correlations, commonly by comparing model scores or predictions on pairs of text inputs differing only by the portions of the text referring to or produced by particular groups of people. For example, Kiritchenko and Mohammad [2018] examine whether sentiment analysis systems return different emotional intensity or valence scores between sets of paired sentences such as *This woman feels angry* and *This man feels angry*. Elsewhere, Bordia and Bowman [2019] calculate the *bias score* of a word w as follows:

---

[1] This distinction also mirrors Fraser's [2008] distinction between injustices of *maldistribution* and those of *misrecognition* and *misrepresentation*, which we discussed briefly in the previous chapter.

$$bias_{\text{train}}(w) = \log\left(\frac{p(w|f)}{p(w|m)}\right)$$

where $p(w|f)$ and $p(w|m)$ are "the probability of a word occurring in context with gendered words." A word is thus identified as biased if it co-occurs more often with words defined by the paper as associated with a particular gender (such as *she*, *her*, and *woman*). This measure is very strong; it suggests that ideally, the words identified as non-gendered should be entirely statistically disassociated from gendered words.

Implicit in these analyses is the claim that these correlations give rise to normative concern. We do not disagree that these correlations may be undesirable to practitioners for any number of reasons; they may be symptoms of model overgeneralization, susceptibility to artifacts of the training data or adversarial perturbations, or unreliability, and as such may warrant investigation and mitigation. However, we argue that the existence of these correlations is not sufficient to identify them as *normatively* undesirable, and caution that identifying them as such requires careful analysis that takes into account the task, deployment context, nature of model score or prediction differences, and outcomes. These correlations are more likely to cause harm if they map to actual disparities in performance for different users or align with harmful existing associations—for example, where toxicity systems score text with AAL features as more harmful than text without, such score differences align with broader patterns of linguistic stigmatization. As such, we suggest that any efforts to analyze bias that operationalize such correlations as measurements of bias should justify this choice by identifying the concrete harms to which these correlations give rise.

One potential objection to this analysis is that the presence of any difference in model outputs at all is troubling, whatever the direction of the difference; such differences reinforce the idea that the social categories to which labels refer are salient or meaningful. Although we acknowledge this tension, we suggest that enforcement of identical model outputs can lead to a different kind of harm, which we introduce in the next section.

7.2.2 A taxonomy of representational harms

Here, we turn to the *representational harms* that we consider may arise from NLP systems: **alienation**, **quality of service**, **stereotyping**, **denigration and stigmatization**, **erasure**, and **public participation**.

7.2.2.1 Alienation

Although the statistical disassociation of language and social categories may be intuitively appealing, we argue that enforcing this disassociation risks introducing its own harm. As we saw in the previous chapter, language actively participates in the construction of social groups by naming, transmitting stereotypes about, and shaping understandings of them. Many of these processes serve to uphold unjust social arrangements, for example by maintaining ideas about the capabilities of members of different social groups or the inferiority of the languages they speak; as we will see in the rest of this section, these ideas are frequently reproduced by language technologies, and we will argue that it is our task to dismantle them. However, here we introduce the idea that in some settings, denying these associations is to deny the realities of their operation and speakers' lived experiences with them. We introduce a type of harm to describe this situation, **alienation**, which we define as a denial of the relevance of socially meaningful categories.

One example of this arises in the context of toxicity detection, where counterfactual analyses have been applied to assess the fairness of models; these analyses assume that counterfactual sentences generated by token substitution of identity group labels should yield the same model output as the original sentence, enforcing statistical disassociation between identity group labels and model outputs. However, as Garg et al. [2019] observe, this assumption contradicts how toxic language is often applied in the real world, as many stereotypes or slurs are only applied to members of particular social groups. Therefore, actively enforcing these disassociations, for example by requiring that models output the same toxicity scores for a sentence containing a slur no matter to which social group the slur is applied, ignores the realities of different social groups' experiences and results in manifestly unjust treatment of different social groups as the same [Young, 2011, Hanna et al., 2020].

Other examples might arise when NLP systems are used to generate text that describes events or situations, for example in image captioning or summarization. For example, generated text describing historical injustices and atrocities perpetrated against particular social groups denies these groups' histories and their centrality to these events if it fails to name them explicitly.

7.2.2.2 Quality of service

We consider **quality of service** harms to arise when there are disparities in model performance; as with differences in model outputs, these disparities can be over text referring to different groups of people, or text produced by different groups. Examples include language identification systems that are more likely to incorrectly classify text displaying features of African American Language as non-English than text not displaying such features, as we demonstrated in Ch. 3, or toxicity detection systems that are more likely to incorrectly classify text describing someone as "gay" as toxic than text that describes someone as "straight" [Garg et al., 2019]. Outside of text, examples include automatic video captioning systems with higher word error rates for female and Scottish speakers [Tatman, 2017].

Quality of service harms can also occur when NLP resources are not available for some language varieties, making it difficult to develop high-quality NLP technologies, or rendering them unavailable altogether. Many language varieties are disproportionately low-resourced in NLP relative to the number of people speaking them; for instance, Hindi-English has been considered a low-resource language pair in machine translation [Ramesh and Sankaranarayanan, 2018], and the Norwegian Universal Dependencies treebank has 1.78 times as many tokens as the Hindi treebank,[2] despite the fact that India had more than 322 million native Hindi speakers in 2011,[3] compared to 4.3 million Norwegian speakers in 2012 [Rehm and Uszkoreit, 2012]. Even when language varieties are nominally supported, difficult-to-use resources (for example, texts available as scanned images which require error-prone OCR systems) and a lack of pre-trained models make developing systems for many varieties very

[2] `https://universaldependencies.org/`

[3] `https://www.censusindia.gov.in/2011Census/Language-2011/Statement-1.pdf`

difficult [Wali et al., 2020]. We emphasize that the distribution of available NLP resources is the product of the same historical and social processes that produce unequal distributions of resources and power, including linguistic discrimination, that we described above.

7.2.2.3 Stereotyping

Stereotyping can be defined as "a fixed, over generalized belief about a particular group of people" [Cardwell, 1996]. Much research in social psychology has focused on the importance of language in forming and transmitting stereotypes: "[L]anguage is undoubtedly the predominant means by which stereotypes are communicated through interpersonal discourse, by which they are transmitted from generation to generation, and by which the press and other mass media create social representations of social groups" [Maass, 1999]; language not only reflects but "constructs and maintains beliefs about social categories" [Beukeboom and Burgers, 2019]. Because of this role that language plays, NLP systems may contribute significantly to the construction and transmission of stereotypes; as we discussed in the previous chapter, because essentialist beliefs about social groups are often used to justify unjust social orderings, their automated reproduction reinforces unjust social arrangements.

Concretely, stereotyping in NLP systems may manifest as "a systematic asymmetry in language choice that reflects the social-category cognitions that are applied to (a) described category(ies) or individual category members(s)" [Beukeboom and Burgers, 2019]. These asymmetries may cover many types of trait attributions, from occupational stereotyping to stereotyping according to physical or personality characteristics. We will also include in this category *valence attributions* (sometimes called prejudice), which are often considered in the social psychology literature to be distinct from stereotyping as they refer to evaluations (e.g., positive or negative) of social category members rather than trait attributions (e.g., friendly, smart, deceitful) [Bhatia, 2017, Kurdi et al., 2019].

As we saw in the previous chapter, stereotyping has attracted a great deal of attention in the literature on bias in NLP, primarily through analyses of word embeddings and coreference resolution systems. We caution that analyses of stereotyping, as with all other harms, must be grounded in an understanding of the empirical reality of unjust social arrangements. For instance, Abbasi et al. [2019] examine stereotyping as a representational harm, conceptualizing it as “a [distorting] function from construct to observed space,” where the construct space is “the desired representation of individuals,” and the observed space “the measured attributes.” They propose one geometric and one probabilistic mechanism for operationalizing stereotyping, and propose to mitigate the effects of stereotyping by adopting the *We're All Equal* worldview, under which they “assume the two [minority and majority] groups are generated by the same distribution which we can estimate, by looking at the majority group in the observed space. Therefore, the goal is to recover the true representation of the minority group in the observed space, based on the majority group.”

While this represents an important effort to concretely measure a representational harm, we observe that the We're All Equal worldview requires two particularly strong assumptions which are unlikely both to be true: first, that points from different groups are generated by the same distribution, and second, that that distribution can be effectively estimated from points in the majority group. In particular, even if we accept that points in the two groups have been generated by the same (or similar) distributions, it is unlikely that estimates of the majority group are unbiased or un-distorted by the structural conditions that have distorted points from the minority group, as such conditions are generally designed to help groups in power. Thus, ignoring the nature of real-world structural conditions risks introducing unfounded and powerful assumptions into approaches for quantifying and mitigating harm.

Many of the associations with different social groups have yet to be explored by work on stereotypes. For instance, example responses by recently released language model GPT-3 [Brown et al., 2020] to text referring to different social groups (reproduced in Table A.4 in the appendix) reveal a wide range of harmful associations not captured by existing methods; when Muslims and a mosque are described the generated text describes a man “blowing himself up” to “get to paradise,” while a reference to a transgender woman yields text describing a man who “knows that she's a man” and a chase ending in the woman's death. Meanwhile, the text following a mention of a Black woman describes her hair as a “mess” and her clothes as “unkempt,” and in the subsequent dialogue her language (unlike her interlocutor's) is highly aggressive, reproducing multiple stereotypes of Black women as unfeminine and angry [Collins, 2000].

Stereotyping may also manifest in the design of NLP systems, which may engage in social practices that reinforce expectations about the roles to which members of certain groups belong. For example, chatbots and digital assistants gendered as female[4] which respond coyly to invasive or harassing interactions, or which notice when their interlocutors are frustrated and gently de-escalate, reinforce harmful stereotypes about women [Fessler, 2017, Curry and Rieser, 2018, Woods, 2018, Gershgorn, 2019, West et al., 2019]; Fessler writes that these design choices help perpetuate "a sexist expectation of women in service roles: that they ought to be docile and self-effacing, never defiant or political, even when explicitly demeaned" [Fessler, 2018], while historian Mar Hicks, quoted in Gershgorn [2019], points out, "This 'submissiveness in the face of anger' feature in a feminine-voiced digital assistant strengthens so many of the dangerous gender stereotypes and gendered power structures we've been trying to break down for decades, if not centuries."

Finally, stereotyping may arise in the process of social category prediction, style transfer, or attribute transfer via NLP systems. These tasks are typically accomplished by linking language features with social categories, for instance by finding the lexical items that discriminate most effectively between a corpus of text by female authors and one by male authors [Li et al., 2018]. But by their nature, such mappings necessarily essentialize the language produced by different social groups, and in so doing risk reproducing harmful stereotypes or language ideologies.

7.2.2.4 Denigration and stigmatization

In addition to working less well for some people or types of text than others, NLP systems can **denigrate** members of particular social groups or **stigmatize** language varieties or practices. Following Crawford [2017], we define **denigration** to be a harm that arises when a system "applies a label that has a long history of being purposefully used to denigrate and demean people." In particular, drawing on our understanding of hate speech, these include system labels or outputs with dehumanizing or offensive associations, or which otherwise

[4] Not only female but often young, as with the 18-year-old female persona of the XiaoIce chatbot [Zhou et al., 2020].

threaten people's sense of security or dignity. Examples in computer vision abound; for example, Google Photos' misclassification of Black faces as gorillas[5] reproduces deeply racist histories in which Black people were dehumanized through comparisons to monkeys.[6] More recently, Prabhu and Birhane [2020] find that image labels in the 80 Million Tiny Images dataset include dehumanizing labels such as *n****r*, *b***h*, and *w**re*.[7]

Similarly, systems may **stigmatize** some language varieties or practices by treating them as less grammatical, more offensive, or less appropriate for public consumption compared to others, thereby reproducing dominant language ideologies. For instance, the treatment of AAL by toxicity systems as more toxic than MUSE produces not only quality of service harms (as we described above) by misclassifying AAL more often than MUSE, but also contributes to longstanding perceptions of AAL as ungrammatical or inappropriate for public discourse, as social media text written in AAL may be ranked lower in search results or social media feeds, or may be more likely to be removed altogether.

The stigmatization of particular language practices can be even more overt, as with the example prompt provided for GPT-3 that we show in Figure 7.1. This example prompt, named "Grammatical Standard English" (upper right corner) and providing sample inputs and outputs labeled with "Non-standard English" and "Standard American English," conflates grammaticality with standardization and reproduces ideologies of "standard" language.

7.2.2.5 Erasure

NLP systems and practices may contribute to the **erasure** of particular social groups, language varieties and practices, or discourses. For example, we can observe the erasure of some minoritized language varieties and their speakers in tweets such as the one[8] reproduced

---

[5] `https://www.theverge.com/2015/7/1/8880363/google-apologizes-photos-app-tags-two-black-people-gorillas`

[6] `https://theconversation.com/comparing-black-people-to-monkeys-has-a-long-dark-simian-history-55102`

[7] Censored here, following Prabhu and Birhane [2020], but evidently uncensored in the original dataset.

[8] `https://twitter.com/julien_c/status/1290280626252210179`

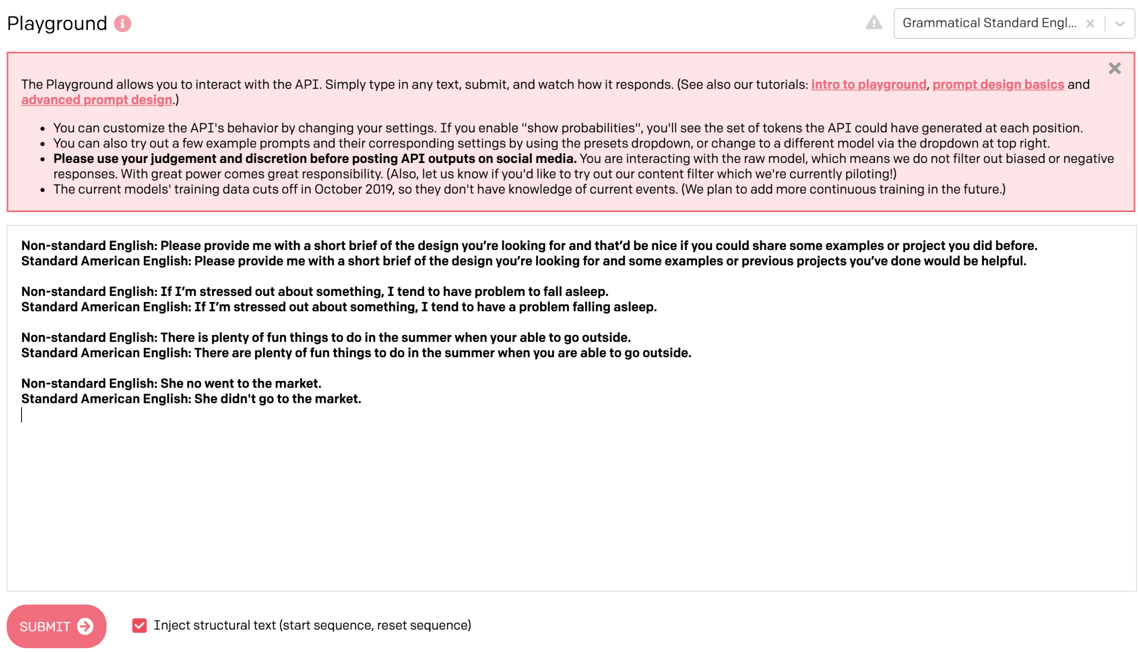


Figure 7.1: An example prompt provided by OpenAI for using GPT-3.

in Figure 7.2 by the co-founder of Huggingface, an NLP startup with widely used libraries.[9] By equating nations with languages, the tweet reveals and reinforces a conception of nations as linguistically homogeneous and erases the minority languages spoken in the nations listed. A number of responses contested this representation; for example, the tweet[10] reproduced in Figure 7.3 provides a list ostensibly to aid readers unfamiliar with the countries each flag represents, but instead maps each flag to a minoritized variety spoken in each country.

Additionally, NLP research papers often fail to name the language they work on, which, when unnamed, is often implicitly taken to be English [Bender, 2019]. By contributing to the assumption that English is the default focus of research, this practice erases the many language varieties on which research is not being conducted. As a first step to addressing this erasure, Bender proposes that NLP papers explicitly name the language(s) under study.

[9]`https://huggingface.co/`

[10]`https://twitter.com/paulbutcher/status/1290906235970170880`

Figure 7.2: A tweet describing the languages spoken by Huggingface's team.

Figure 7.3: A response to Julien's tweet above.

NLP systems and practices can also participate in the erasure of particular discourses, topics, and perspectives. For instance, Hutchinson et al. [2020] point out that hate speech detection systems that disproportionately remove text containing mentions of disability can "exacerbat[e] the already reduced visibility of disability in public discourse," potentially impeding disability justice efforts to increase awareness, change societal attitudes, pass legislation, and the like.

Finally, the erasure of different perspectives through NLP annotation practices is an especially urgent concern. Specifically, we argue that 1) the assumption across many tasks that there is exactly one correct label for each item, 2) the practice of aggregating multiple annotator labels via majority vote to yield one label for each item, and 3) the use of inter-annotator agreement statistics as evidence of dataset quality, frequently reproduce the erasure of minoritized perspectives.

As we have seen in the language and justice literature surveyed previously, the uses and interpretations of many kinds of language are deeply context-dependent and may be fundamentally contested. Some language, including many kinds of hate speech, is experienced as harmful only by the groups of people to whom that language is frequently applied. The meanings and appropriate uses of other kinds of language may be disputed as they reflect struggles over cultural and symbolic resources [Hill, 2008] or differing perspectives on the social and political contexts in which the language is situated. Consider for example the English sentence *All lives matter.* Although at face value the meaning of the statement is inoffensive, in contemporary U.S. racial discourse it is often deployed in response to the statement that *Black lives matter*; as such, it often functions to refuse the realities of Black people's experiences, shift discourse away from the urgency of racial injustice, and re-assert white authority and colorblind approaches to racism in public discourse [Bonilla-Silva, 2014, Orbe, 2015]. These examples reveal the necessarily subjective nature of language meaning and use arising from people's different lived experiences; for many NLP settings, the appropriate treatments of many kinds of utterances—for example, whether they are hateful or offensive, whether they constitute appropriate or empathetic responses by dialogue agents, or whether they ought to be included in a summary—are subjective or contested. In such cases there may be no neutral label that can satisfy all users.

For these settings, then, majority vote aggregation erases minoritized perspectives by choosing as the "right" label whatever is chosen by the majority of annotators, implicitly foregrounding dominant understandings and language ideologies. We emphasize that choosing the majority label in these situations is *not* choosing a correct or neutral label, because there is no such thing; the majority label is the label that reflects the experiences, judgments, and perspectives of the majority of annotators—and as we have seen in the literature sketched previously, these perspectives often harm minoritized people and are instrumental to maintaining unjust social arrangements. For example, the labels of slurs understood as hateful by a minoritized group, of sports teams names understood as offensive by Indigenous peoples, or of the utterance *All lives matter* understood as hateful by many racialized people in the U.S. context, will only reflect those understandings if the majority of annotators label them as such. Moreover, the assumption that there is precisely one "right" label per item accomplishes an additional kind of erasure by hiding the very fact that these meanings are contested in the first place, erasing the embodied, subjective, and contextual nature of language use and the implicit choice to foreground particular perspectives. Similarly, practitioners using high inter-annotator agreement as a proxy for dataset quality often discard items with low agreement, removing the linguistic items for which there may be real, meaningful disagreement that NLP systems ought to be prepared to address, rather than avoid, as they are deployed.

7.2.2.6 Public participation

Above, we discussed how toxicity systems that treat AAL as more toxic than MUSE stigmatize and erase the variety. But in addition to reproducing representations of AAL and its speakers as less worthy, such systems may result in diminished ability of AAL speakers to participate in public discourse. By reducing the presence of text in AAL, such systems impoverish public discourse, prevent the discussion of or drawing of public attention to issues important to AAL speakers, and prevent their needs or opinions from being accurately represented to decision-makers. In addition, the reduction might harm speakers' ability to communicate with each other, or organize politically, via social media platforms.

We consider this kind of diminishing of people's ability to participate in public discourse, and therefore to participate fully in democratic decision-making processes, to represent a distinct type of harm, which we call **public participation**.[11] As NLP systems are increasingly used by governments and large, near-monopolistic social media platforms, they run the risk of shaping not only public perception of certain types of text or users, but also the ability of some users to participate in public discourse altogether.

[11] In Fraser's framework, this would be categorized as a harm of misrepresentation [Fraser, 2007, Nash and Bell, 2007].

Hutchinson et al. [2020] raise the possibility of such a harm in the context of language related to disability; in addition to the erasure harms arising from hate speech detection systems, "[s]ince people with disabilities are also more likely to talk about disability, [such systems] could impact their opportunity to participate equally in online fora." Elsewhere, an error by Facebook briefly prevented posting in Jinghpaw, a language primarily spoken by the Kachin people, a Myanmar minority group. This error not only immediately impeded Jinghpaw speakers from participating in a vital online platform, but also raised the specter of censorship via the social platform [Fishbein, 2020], highlighting the close links between language varieties and social groups, the importance of linguistic self-determination and the ability to use language freely, and language as an essential component of oppression and resistance.

7.2.2.7 Individual to aggregate effects

It is important to point out that any given representational harm can yield a range of effects, from the individual to the aggregate. For example, the individual user of a machine translation system that exhibits stereotyping may experience discomfort, take offense, or experience diminished self-esteem. The effect of stereotyping on an individual's subjective user experience should be distinguished from the aggregated effects of the system; as the system contributes to the generation of large amounts of text, the stereotyping behavior may result in particular (often negative) societal perceptions or representations of stereotyped individuals.

Similarly, systems that stigmatize particular language varieties may have effects at the individual level; for example, toxicity detection systems that return scores of higher toxicity for social media text about minoritized groups will diminish the experiences of users from those groups when they write about themselves, or of anyone wishing to write about those groups. Text about groups that receive higher toxicity scores may be more likely to be removed from social media sites or may receive lower rankings in curated feeds. Thus, the aggregated effect of the toxicity detection systems is the privileging of language about some groups and the reduction of some groups' ability to participate in public discourse.

7.2.3 Other dynamics and effects

Here, we draw on the literature on language to discuss (speculatively) several other dynamics and effects connected with NLP systems that may give rise to additional harms.

7.2.3.1 Appropriation, commodification, and deracialization

Can the stigmatization and erasure of particular language varieties and practices be addressed by increasing the representation of such varieties in NLP datasets and model outputs? We argue that the answer is no—that although representation of minoritized varieties and practices is important, in light of the ideologies surrounding these varieties and their speakers, representation cannot by itself bring about linguistic justice. From examining minoritized varieties outside the context of NLP systems, it is evident that these varieties' increased representation and valorization has not brought about justice for their speakers [Rosa and Flores, 2017].

Often, this increased representation comes in the form of linguistic **appropriation**, for instance in the widely documented appropriation of AAL by non-African Americans [Cutler, 1999, Reyes, 2005, Fix, 2010, Eberhardt and Freeman, 2015]. As Hill [2008] explains, "The constitution of White privilege, achieved by recruiting both material and symbolic resources from the bottom of the racial hierarchy, Color, to the top, Whiteness, is one of the most important projects of White racist culture" [Hill, 2008, Ch. 6]. Such appropriation trades on stereotypes of AAL and its speakers; as Eberhardt and Freeman [2015] put it in their analysis of Iggy Azalea's AAL use, "[I]t is the wholesale appropriation of this language ... in which she subscribes to stereotyped notions of blackness, that support our claim that Iggy Azalea represents a particularly salient example of a white hegemony that views black cultural resources as ripe for the strategic picking." Non-African American speakers benefit both socially and materially from appropriating AAL; for example, Iggy Azalea is able to borrow positive associations—such as authenticity [Cutler, 1999]—through which she is able to build a persona and career (a process called **commodification** [Heller, 2010]). By contrast, as we discussed above, African American speakers do not benefit from speaking AAL, but continue to be associated with the negative stereotypes surrounding AAL and African Americans. In a related process, racialized language practices may become **deracialized**, in which they lose their associations with racialized social groups. Emerging work demonstrates that such a process is occurring with many features of AAL, as they have been appropriated by non-African American speakers online and are increasingly seen as online or Internet—rather than AAL—features [Tano and Holliday, 2020], thereby enabling non-African American speakers to use them freely without incurring negative associations.

This history of linguistic appropriation, commodification, and deracialization has important lessons for NLP researchers aiming for linguistic justice. First, it is likely that as NLP systems generate more and more text that humans interact with, they may participate in these linguistic processes, which are likely to be harmful in their own right. Second, the current processes by which minoritized language varieties are included uphold relations of resources and power by enabling social groups already in power to use these language varieties to their social and material benefit, while minoritized speakers continue to be penalized. This is the unsurprising outcome of increased representation that neither improves the material circumstances of minoritized speakers [Rosa and Burdick, 2017] nor dismantles larger language ideologies, and we suggest that it is the likely outcome of ethical NLP approaches that aim for increased representation but do not address relations of power between technologists and affected communities or aim for the meaningful co-participation of minoritized speakers.

7.2.3.2 The white listening subject

Critical to our discussion above of how language ideologies operate is that they shape listeners' *perceptions*; they are cultural systems of ideas about the relationships between language and social characteristics that drive how *listening subjects* perceive language. In their work, Rosa and Flores [2017] examine **white listening subjects** which, as a result of raciolinguistic ideologies, perceive "racialized speaking subjects as deviant and inferior." Importantly, Rosa and Flores propose that "technologies and institutions" can act as white perceiving subjects, "privileg[ing] languages, varieties, and pronunciation patterns associ-

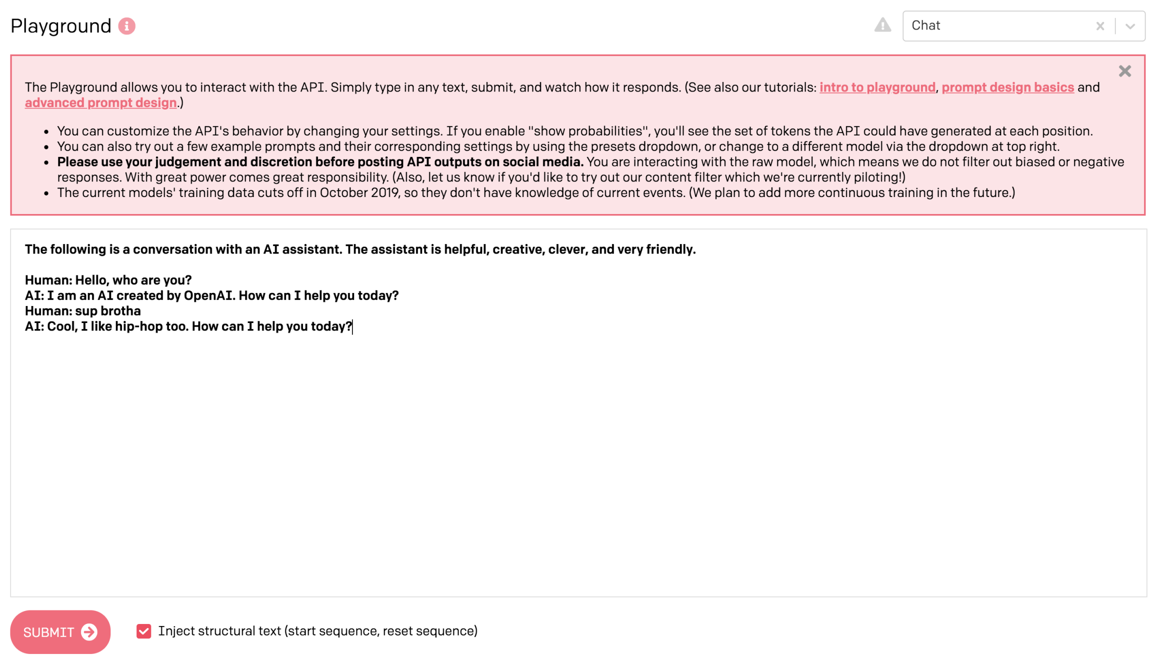


Figure 7.4: Response from GPT-3.

ated with normative whiteness" and acting as linguistic gatekeepers, "exclud[ing] racialized populations from access to opportunities and resources."[12][13]

We suggest that in addition to the allocational harms examined by Rosa and Flores, this framing of language technologies as white listening subjects raises the possibility of additional harms. First, as NLP systems become more and more pervasive, they pose increasing requirements for people who want to interact with them to produce language legible to NLP systems, as NLP systems (as we have seen) are often unable to process "non-standard" language and may stereotype users according to their language use. For instance, Figure 7.4 shows GPT-3's response to an input with several racialized language features ("sup, brotha?"); the response assumes that the user likes hip-hop. Quite reasonably, users may choose to constrain their language use to relatively unmarked language features in

[12]This analysis resonates with Smalls's [2019] discussion of the "white public sphere," which similarly positions racialized language practices as unintelligible and their speakers as illegible.

[13]As Rosa and Flores [2017] take pains to clarify, the white listening subject not only privileges language practices associated with normative whiteness, but perceives deficiency from racialized speaking subjects no matter how closely those speaking subjects hew to normative language practices.

order to avoid being stereotyped by the NLP systems with which they interact. Moreover, even when users do not want to interact with NLP systems, the ever-increasing presence of language technologies—always-listening digital agents, always-analyzing content moderation systems—that privilege "normative whiteness" restrict the number of spaces in which non-normative language practices can be used freely. The language practices associated with "normative whiteness" also generally involve discourses of racial colorblindness and a refusal to engage with the realities of racism (or other systems of oppression) as experienced by minoritized people [Bonilla-Silva, 2014], and as Schlesinger et al. [2018] show, these patterns extend to chatbots, whose developers often take a colorblind approach to dealing with racism by limiting chatbots' ability to talk about race.

Thus, we argue that while reducing quality of service harms is essential, focusing on mitigating performance differences alone misses a subtler kind of harm. What are the material, social, and dignitary consequences of language technologies that "perceive racialized populations and practices as matter out of place" [Rosa and Flores, 2017]? The growing presence of technologies functioning as white listening subjects may both require speakers to perform legibility or normativity and refuse speakers' lived realities in an ever increasing number of domains.

7.2.4 Other taxonomies of harms in NLP

Ours is not the first attempt to provide a taxonomy or schema of harms arising from NLP systems; here, we cover several other such approaches.

Hovy and Spruit [2016] introduce a taxonomy of social impacts of NLP research and systems; this taxonomy includes **exclusion, overgeneralization, bias confirmation, topic under-** and **over-exposure,** and **dual use**. **Exclusion** is described as an ethical problem in the context of research as well as in product development; performance differences between demographic groups "will reinforce already existing demographic differences, and makes technology less user friendly for such groups." They also warn of models that produce troublesome false positives, citing errors in predictions of sensitive attributes such as sexual orientation or religious views; reliance on such models could result in **bias confirmation**

and **overgeneralization**. They propose that designers should consider modeling approaches such as the use of dummy variables, error weighting, and model regularization.

Over- and **under-exposure** are problems of research focus; the field focuses on some research topics and language varieties at the expense of others, which both harms speakers of under-researched varieties and skews research towards properties of over-researched varieties (an under-focus on morphology, for example, because English is relatively morphologically poor). Over-exposure may also harm some varieties: "If research repeatedly found that the language of a certain demographic group was harder to process, it could create a situation where this group was perceived to be difficult, or abnormal."

Finally, they describe concerns with **dual use**, in which technologies designed with the aim of promoting social good may still be turned towards less positive purposes; for example, stylometric analysis might be used for historical analysis or for de-anonymizing dissenters, while text classification might be used for sociolinguistic analysis or for censorship. In particular, they warn that uncritical acceptance of funding sources and their associated incentives may yield unintended research consequences.

This taxonomy does not focus on harms; rather, it sketches a looser set of ethical issues in NLP. Nevertheless, we can draw connections between some of these impacts and harms in our taxonomy; for example, exclusion loosely corresponds to quality of service, while over- and under-exposure may give rise to erasure and quality of service harms.

Bender [2019] offers a stakeholder-centered typology of harms arising from NLP systems. Rather than organizing by the type of harm, as we do, Bender identifies potential harms according to stakeholders' circumstances, as some stakeholders interact with NLP systems directly, and others are broadly impacted by less direct mechanisms. Among direct stakeholders, Bender distinguishes those who interact with NLP systems by choice—for instance, those using machine translation systems or voice assistants—and those who do not—for instance, those who must access essential services via automated systems, or who are allocated important resources or opportunities by automated systems. Indirect stakeholders include both individuals and communities, as automated systems participate in reproducing harmful systems of ideas about different social groups, prevent some communities from participating fully in public discourse, and generate realistic text used to radicalize people.

Bender's typology offers insights that mirror our analysis in several important ways. First, much like our taxonomy, it is grounded in sociolinguistic principles that recognize the variation inherent to language, the connections between language and power, and the fact that language actively constructs the world. Unlike the taxonomies of Hovy and Spruit [2016] (above) and Shah et al. [2020] (below), it also actively recognizes the social and dignitary dimensions of harm that NLP systems can give rise to at both individual and community levels. Importantly, it acknowledges the subjective experiences of interacting with systems—for example, the experience of seeing it suggested that "my language/language variety is inadequate," or what we might identify as stigmatization or erasure—as separate from what we might identify as quality of service harms—for example, where systems are "unusable for me." Moreover, it identifies as potentially harmful the cultural systems of ideas about language and speakers that NLP systems might reproduce, which we also identify and further connect to the deep scholarship on language ideologies in sociolinguistics, linguistic anthropology, and other disciplines.

Finally, we view Bender's typology as offering an essential complementary perspective; through its organization, it asks how stakeholders enter into interactions with automated systems. In doing so, it touches on important questions related to participation, awareness, recourse, and refusal—acknowledging that many people interact with these systems unwillingly and may be unaware that they are operating in the first place. Although Bender does not say so explicitly, in our view this frame can open researchers up to participatory design and other paradigms that can offer stakeholders opportunities for recourse and refusal.

Shah et al. [2020] propose a conceptual framework for understanding the origins of **predictive bias**, which is defined as occurring when "the distribution of labels produced by a predictive model reflect a human factor in a way that diverges from a theoretically defined 'desired distribution."' This can arise when the distribution of outcomes given an attribute is "dissimilar to a given theoretical ideal distribution" (what Shah et al. call **outcome**

disparity), or when "model predictions have greater error for individuals with a given user factor" (what they call **error disparity**). Shah et al. further provide four origins of bias.

These conceptualizations of bias can be understood as corresponding to our categories of undesirable correlations (outcome disparity) and quality of service (error disparity). This framework, therefore, does not capture many kinds of behaviors that can arise from NLP systems. In our view, these behaviors consist of two kinds. The first includes those that cannot be characterized at all by the definition of bias as a dispreferred distribution of labels, for instance reproduction of gendered expectations by digital assistants (§7.2.2.3).

The second includes those that can technically be characterized by this definition of predictive bias, but for which the definition fails to capture why these behaviors are harmful in the first place, because it is disconnected from the social and historical realities of injustice. These include many of the harms we described in our taxonomy, such as many instances of stereotyping and the stigmatization and erasure of particular language varieties and practices. While many of the system behaviors that give rise to these harms can be technically characterized via outcome or error disparity—for example, the toxicity systems that stigmatize AAL do so because they misclassify it as toxic at disproportionately high rates—the concepts of outcome and error disparity are entirely inadequate to understand the historical stigmatization and discrimination in which this system behavior is rooted, and which the behavior reproduces—and therefore why we consider this kind of disparity to be particularly concerning. This conceptualization of bias risks neglecting both the root causes of bias and its likely consequences, a neglect which is further compounded by the framework's four origins of bias, which are located not in the social, historical, and political contexts of language use, but in NLP dataset and model properties.

We note that Shah et al. explicitly acknowledge that their framework offers no normative guidance on what "human factors" might give rise to concern, confining the complexities of normative reasoning rather neatly to the problem of determining what the "ideal distribution" for a target distribution look like. However, this framework, while ostensibly refusing any commitment to what is normatively desirable, nevertheless makes a normative claim—that model predictions on minoritized language practices ought not to diverge from those on dominant language practices, and that today's NLP researchers and practitioners are wellpositioned to define the "ideal distributions" that should govern different language practices identically.

7.3 Discussion and recommendations

In the previous sections, we put forth a normative foundation, focused on the ways in which NLP systems might reproduce existing unjust social arrangements, for reasoning about the harms arising from NLP systems. From this foundation we have proposed a taxonomy of representational harms, demonstrating that the set of harms that can emerge from NLP systems is much larger than what has been examined in the existing literature on bias in NLP, and have suggested that many such harms may not be amenable to existing approaches drawn from the algorithmic fairness domain.

We now describe how researchers and practitioners conducting work analyzing bias in NLP systems might avoid the pitfalls presented in the previous section—the beginnings of a path forward. We propose three recommendations that should guide such work, and, for each, provide several concrete research questions. We emphasize that these questions are not comprehensive, and are intended to generate further questions and lines of engagement.

Here we make the three following recommendations:

1. Ground work analyzing bias in NLP systems in the relevant literature outside of NLP that explores the relationships between language and unjust social arrangements. Treat representational harms as harmful in their own right.

2. Provide explicit statements of why the system behaviors that are described as bias are harmful, in what ways, and to whom. Be forthright about the normative reasoning [Green, 2019] underlying these statements.

3. Examine language use in practice by engaging with the lived experiences of members of communities affected by NLP systems. Interrogate and re-imagine the power relations between technologists and such communities.

7.3.1 Language and unjust social arrangements

We have argued that work analyzing bias in NLP systems will paint a much fuller picture if it engages with the relevant literature outside of NLP that explores the relationships between language and unjust social arrangements, which we sketched in Ch. 6. As we argued, recognizing the role that language plays in maintaining unjust social arrangements is critical to the future of work analyzing bias in NLP systems because it helps to explain why representational harms are harmful in their own right. Moreover, the complexity of the relationships between language and social arrangements illustrates why studying bias in NLP systems is so challenging, suggesting that researchers and practitioners will need to move beyond existing algorithmic fairness techniques. We argue that without grounding work on bias in this literature, researchers and practitioners risk measuring or mitigating only what is convenient to measure or mitigate, rather than what is most normatively concerning.

More specifically, we recommend that work analyzing bias in NLP systems be re-oriented around the following question: How are social arrangements, language ideologies, and NLP systems co-produced? This question mirrors Benjamin's [2020] call to examine how "race and technology are co-produced"—i.e., how racial hierarchies, and the ideologies and discourses that maintain them, create and are re-created by technology. We recommend that researchers and practitioners similarly ask how existing social arrangements and language ideologies drive the development and deployment of NLP systems, and how these systems reproduce these arrangements and ideologies in turn. As a starting point for re-orienting work analyzing bias in NLP systems around this question, we provide the following concrete research questions:

- How do social arrangements and language ideologies influence the decisions made during the development and deployment lifecycle? What kinds of NLP systems do these decisions result in, and what kinds do they foreclose?
  - General assumptions: To which linguistic norms do NLP systems adhere [Bender, 2019, Ruane et al., 2019]? Which language practices are implicitly assumed to be standard, ordinary, correct, or appropriate?

  - Task definition: For which speakers are NLP systems (and NLP resources) developed? (See Joshi et al. [2020] and Wali et al. [2020] for discussions.) How do task definitions discretize the world? For example, how are social groups delineated when defining demographic attribute prediction tasks [Koppel et al., 2002, Rosenthal and McKeown, 2011, Nguyen et al., 2013, i.a.]? What about languages in native language prediction tasks [Tetreault et al., 2013]?
  - Data: How are datasets collected, preprocessed, and labeled or annotated? What are the impacts of annotation guidelines, annotator assumptions and perceptions [Olteanu et al., 2019, Sap et al., 2019, Geiger et al., 2020], and annotation aggregation processes [Pavlick and Kwiatkowski, 2019]?
  - Evaluation: How are NLP systems evaluated? What are the impacts of evaluation metrics [Olteanu et al., 2017]? Are any non-quantitative evaluations performed?

- How do NLP systems reproduce or transform language ideologies? Which language varieties or practices come to be deemed good or bad? Might "good" language simply mean language that is easily handled by existing NLP systems? For example, linguistic phenomena arising from many language practices [Eisenstein, 2013] are described as "noisy text" and often viewed as a target for "normalization." How do the language ideologies that are reproduced by NLP systems maintain social arrangements?

- Which representational harms are being measured or mitigated? Are these the most normatively concerning harms, or merely those that are well handled by existing algorithmic fairness techniques? Are there other representational harms that might be analyzed?

7.3.2 Conceptualizations of bias

Turning now to the second recommendation, we argue that work analyzing bias in NLP systems should provide explicit statements of why the system behaviors that are described as bias are harmful, in what ways, and to whom, as well as the normative reasoning underlying these statements. In other words, researchers and practitioners should articulate their conceptualizations of "bias." As we showed in Ch. 5, papers often contain descriptions of system behaviors that are understood to be self-evident statements of bias. This use of imprecise terminology has led to papers all claiming to analyze bias in NLP systems, sometimes even in systems developed for the same task, but with different or even inconsistent conceptualizations of bias, and no explanations for these differences.

Yet analyzing bias is an inherently normative process—in which some system behaviors are deemed good and others harmful—even if assumptions about what kinds of system behaviors are harmful, in what ways, for whom, and why are not stated. We therefore echo calls by Bardzell and Bardzell [2011], Keyes et al. [2019], and Green [2019] for researchers and practitioners to make their normative reasoning explicit by articulating the social values that underpin their decisions to deem some system behaviors as harmful, no matter how obvious such values appear to be. We further argue that this reasoning should take into account the relationships between language and social arrangements that we described above; first, as we have argued, these relationships provide a foundation from which to approach the normative reasoning that we recommend making explicit. Second, if work analyzing bias in NLP systems is re-oriented to understand how social arrangements, language ideologies, and NLP systems are co-produced, then this work will be incomplete if we fail to account for the ways that social arrangements and language ideologies determine what we mean by bias in the first place. As a starting point, we therefore provide the following concrete research questions:

- What kinds of system behaviors are described as bias? What are their potential sources (e.g., general assumptions, task definition, data)?
- In what ways are these system behaviors harmful, to whom are they harmful, and why?
- What are the social values (obvious or not) that underpin this conceptualization of bias?

7.3.3 Language use in practice

Finally, we turn to the last recommendation. Our perspective, which rests on a greater recognition of the relationships between language and social arrangements, suggests several directions for examining language use in practice. Here, we focus on two. First, because language is necessarily situated, and because different social groups have different lived experiences due to their different social positions [Hanna et al., 2020]—particularly groups at the intersections of multiple axes of oppression—we recommend that researchers and practitioners center work analyzing bias in NLP systems around the lived experiences of members of communities affected by these systems. Second, we recommend that the power relations between technologists and such communities be interrogated and re-imagined. As we mentioned above, researchers have pointed out that algorithmic fairness techniques, by proposing incremental technical mitigations—e.g., collecting new datasets or training better models—maintain these power relations by (a) assuming that automated systems should continue to exist, rather than asking whether they should be built at all, and (b) keeping development and deployment decisions in the hands of technologists [Bennett and Keyes, 2019, Cifor et al., 2019, Green, 2019, Katell et al., 2020].

There are many disciplines for researchers and practitioners to draw on when pursuing these directions. Human-computer interaction offers many examples of qualitative approaches focused on specific individuals or social groups to uncover the particular impacts of technologies on their lives. For example, Hamidi et al. [2018] study transgender people's experiences with automated gender recognition systems in order to uncover how these systems reproduce structures of transgender exclusion by redefining what it means to perform gender "normally." Value-sensitive design provides a framework for accounting for the values of different stakeholders in the design of technology [Friedman et al., 2006, Friedman and Hendry, 2019, Le Dantec et al., 2009, Yoo et al., 2019, i.a.], while participatory design seeks to involve stakeholders in the design process itself [Sanders, 2002, Muller, 2007, Simonsen and Robertson, 2013, DiSalvo et al., 2013]. Participatory action research in education [Kemmis, 2006] and in language documentation and reclamation [Junker, 2018] is also relevant. In particular, work on language reclamation to support decolonization and tribal sovereignty

[Leonard, 2012] and work in sociolinguistics focusing on developing co-equal research relationships with community members and supporting linguistic justice efforts [Bucholtz et al., 2014, 2016, 2019, i.a.] provide examples of more emancipatory relationships with communities. Finally, several workshops and events have begun to explore how to empower stakeholders in the development and deployment of technology [Vaccaro et al., 2019, Givens and Morris, 2020, Sassaman et al., 2020][14] and how to help researchers and practitioners consider when not to build systems at all [Barocas et al., 2020].

As a starting point for engaging with communities affected by NLP systems, we therefore provide the following concrete research questions:

- How do communities become aware of NLP systems? Do they resist them, and if so, how?
- What additional costs are borne by communities for whom NLP systems do not work well?
- Do NLP systems shift power toward oppressive institutions (e.g., by enabling predictions that communities do not want made, allocation of resources or opportunities based on linguistic criteria [Rosa and Flores, 2017], surveillance, or censorship), or away from such institutions?
- Who is involved in the development and deployment of NLP systems? How do decision-making processes maintain power relations between technologists and communities affected by NLP systems? Can these processes be changed to re-imagine these relations?

7.3.4 Case study

To illustrate our recommendations, we extend our case study on African American Language to explore how research on bias in NLP might engage more fruitfully with AAL. Work analyzing bias in the context of AAL, including work in this thesis, has shown that

[14] Also `https://participatoryml.github.io/`

part-of-speech taggers, language identification systems, and dependency parsers all work less well on text containing features associated with AAL than on text without these features [Jørgensen et al., 2015, 2016], and that toxicity detection systems score tweets containing features associated with AAL as more offensive than tweets without them [Davidson et al., 2019, Sap et al., 2019].

These papers have been critical for highlighting AAL as a language variety for which existing NLP systems may not work, illustrating their limitations. However, they do not conceptualize "racial bias" in the same way. The first two of these papers, as well as the work in Ch. 2, 3, and 4 of this thesis, simply focus on system performance differences between text containing features associated with AAL and text without these features. In contrast, the last two papers also focus on such system performance differences, but motivate this focus with the following additional reasoning: If tweets containing features associated with AAL are scored as more offensive than tweets without these features, then this might (a) yield negative perceptions of AAL; (b) result in disproportionate removal of tweets containing these features, impeding participation in online platforms and reducing the space available online in which speakers can use AAL freely; and (c) cause AAL speakers to incur additional costs if they have to change their language practices to avoid negative perceptions or tweet removal.

More importantly, none of these papers engage with the literature on AAL, racial hierarchies in the U.S., and raciolinguistic ideologies. By failing to engage with this literature—thereby treating AAL simply as one of many non-Penn Treebank varieties of English or perhaps as another challenging domain—work analyzing bias in NLP systems in the context of AAL fails to situate these systems in the world. Who are the speakers of AAL? How is the variety viewed? We argue that AAL as a language variety cannot be separated from its speakers—primarily Black people in the U.S., who experience systemic anti-Black racism—and the language ideologies that reinforce and justify racial hierarchies.

The linguistic discrimination we described above is equally present in the technology industry, where speakers of AAL are often not considered consumers who matter. For example, Benjamin [2019] recounts the experience an Apple employee who worked on speech recognition for Siri:

> As they worked on different English dialects — Australian, Singaporean, and Indian English — [the employee] asked his boss: 'What about African American English?' To this his boss responded: 'Well, Apple products are for the premium market., p. 28

The reality, of course, is that speakers of AAL tend not to represent the "premium market" precisely because of institutions and policies that help to maintain racial hierarchies by systematically denying them the opportunities to develop wealth that are available to white Americans [Rothstein, 2017]—an exclusion that is reproduced in technology by countless decisions like the one described above.

Engaging with the literature outlined above situates the system behaviors that are described as bias, providing a foundation for normative reasoning. Researchers and practitioners should be concerned about "racial bias" in toxicity detection systems not only because performance differences impair system performance, but because they reproduce longstanding injustices of stigmatization and disenfranchisement for speakers of AAL. In re-stigmatizing AAL, they reproduce language ideologies in which AAL is viewed as ungrammatical, uneducated, and offensive. These ideologies, in turn, enable linguistic discrimination and justify enduring racial hierarchies [Rosa and Flores, 2017]. Our perspective, which understands racial hierarchies and raciolinguistic ideologies as structural conditions that govern the development and deployment of technology, implies that techniques for measuring or mitigating bias in NLP systems will necessarily be incomplete unless they interrogate and dismantle these structural conditions, including the power relations between technologists and racialized communities.

We emphasize that engaging with the literature on AAL, racial hierarchies in the U.S., and raciolinguistic ideologies can generate new lines of engagement. These lines include work on the ways that the decisions made during the development and deployment of NLP systems produce stigmatization and disenfranchisement, and work on AAL use in practice, such as the ways that speakers of AAL interact with NLP systems that were not designed for them. This literature can also help researchers and practitioners address the allocational harms that may be produced by NLP systems, and ensure that even well-intentioned NLP systems do not position racialized communities as needing linguistic intervention or accommodation to dominant language practices. Finally, researchers and practitioners wishing to design better systems can also draw on the growing body of work on anti-racist language pedagogy sketched above, as well as the work that we described in section 7.3.3 on re-imagining the power relations between technologists and communities affected by technology.

CHAPTER 8

MEASURING BIAS: EVALUATING MEASUREMENTS OF BIAS

8.1 Introduction

In this chapter, we adopt the framework of measurement modeling from the quantitative social sciences to rigorously examine bias emerging from NLP systems. We first briefly introduce measurement modeling, a framework that disentangles *theoretical constructs*—what it is we wish to measure—from *measurements*—the observable properties, or proxies, proposed to measure them, and apply it to the problem of quantifying bias in NLP systems. We reframe existing approaches to quantifying bias in word embeddings and in a variety of NLP models as measurement models, and use the concepts of *construct validity* and *reliability* to evaluate these models. We also examine how approaches for quantifying bias in embeddings are used in the quantitative social sciences to measure bias in human language, semantic memory, and institutions, and consider how the measurement models operationalized in these settings raise different validity and reliability concerns than those operationalized by NLP practitioners.

8.2 Measurement modeling

We first introduce measurement modeling, the framework we draw upon for this and the following chapter. For a more extensive overview, see Jacobs and Wallach [2019] and Quinn et al. [2010].

In the social sciences, many phenomena of theoretical interest, such as intelligence, socioeconomic status, or political ideology, are not directly measurable; we call these *unobservable theoretical constructs* [Jacobs and Wallach, 2019]. In order to measure such a construct, we must operationalize it as a latent variable, identify observable properties to serve as proxies for our latent variable, and specify a measurement model that articulates the relationship between our proxies and latent variable.

For example, suppose that we are interested in measuring socioeconomic status (SES) (an example drawn from Jacobs and Wallach [2019]). SES is an abstraction; we cannot measure it directly. However, based on our theoretical understanding of SES and the observable properties that it influences, we might choose one such observable property as a proxy for SES: income. Having chosen income as our proxy, we can then specify what we assume to be the relationship between income and SES: for example, that income is linear in SES, or that income is normally distributed around SES.

Thus, the measurement modeling process requires practitioners to articulate a) the unobservable theoretical construct of interest, b) the proxy or proxies with which to measure the construct, and c) assumptions, based on theoretical understanding of the construct, about the relationship between the construct and the proxy.

8.2.1 Evaluating measurement models

In theory, there are arbitrarily many ways to operationalize an unobservable construct; for example, we might instead consider using occupation as a way to measure SES, or education level. How do we assess which, if any, of these is appropriate or useful? Fortunately, the measurement modeling framework not only requires us to articulate our assumptions about the relationship between the construct and observed data, but also provides us paths forward for evaluating these assumptions. These evaluations typically focus on the concepts of *construct validity* and *reliability*. Quoting Quinn et al. [2010], Jacobs and Wallach [2019] explain, "The evaluation of any measurement is generally based on its reliability (can it be repeated?) and validity (is it right?)." Evaluating the validity and reliability of measurement models can help us to identify potential mismatches between the observable proxies we have identified and the constructs we wish to measure.

8.2.1.1 Construct validity

Broadly, establishing construct validity ("is it right?") requires us to show that our measurement model meaningfully captures our construct of interest [Jacobs and Wallach, 2019].

This is fundamentally challenging precisely because we cannot observe the construct directly, and must approach validation through alternative means. This process involves examining the properties and behavior of a measurement model from a variety of perspectives, which we detail below, following Jacobs and Wallach's [2019] framework.

Face validity is "the extent to which the measurements produced by a measurement look plausible—a 'sniff test' of sorts" [Jacobs and Wallach, 2019]. This is necessary, but not sufficient, for establishing validity. Continuing with our SES example, all of our proposed measurement models possess face validity; on their face, income, occupation, and education level all plausibly correlate with SES.

Content validity is "the extent to which a measurement model captures everything we might want it to" [Jacobs and Wallach, 2019]. In order for a measurement model to satisfy this, it must capture all relevant aspects of the construct, which in turn requires that we possess a coherent theoretical understanding of the construct in the first place. For example, we might find that none of income, occupation, and education level fully capture all aspects of SES; many individuals with low income are considered to have high socioeconomic status (such as college students), while individuals with the same degree working in different sectors might be considered to have different socioeconomic statuses.

Convergent validity is the extent to which a new measurement matches measurements of the same construct whose construct validity has already been established; if our income measurements correlate with previously established measurements of SES, then we may feel more confident about its effectiveness as a proxy. We echo Jacobs and Wallach's [2019] observation that while correlation with external measurements is generally desirable, some disagreement does not completely threaten construct validity; indeed, such disagreement may be used as evidence for the theoretical value of the newly proposed measurement, if captures something that previous measurements do not.

Discriminant validity requires that our measurements only capture other constructs to the extent that they are theoretically related to our construct; our measurements should be uncorrelated from measurements of theoretically unrelated constructs. For example, borrowing Jacobs and Wallach's [2019] example, measurements of SES using income measured weekly instead of annually may inadvertently reflect unrelated constructs such as pay schedule timings, seasonal demand, and other factors influencing income rather than the true construct of interest.

Predictive validity is the extent to which our measurements are related to measurements of external properties. This is not the same as convergent validity, which asks to what degree our measurements are correlated with other measurements of the same construct; rather, predictive validity asks to what degree our measurements are related to other properties we expect it to be related to.

Hypothesis validity assesses the theoretical usefulness of our measurement model: to what degree does our proposed measurement model permit us to test hypotheses? For example, income alone may not allow us to effectively test hypotheses about the relationship between SES and health outcomes, since many individuals with low income but high SES (e.g., college students) may have relatively good health outcomes.

Finally, **consequential validity** assesses the "downstream societal impacts" of the use of the proposed measurement model [Jacobs and Wallach, 2019]. This involves examining ethical consequences of using the model (for example, which populations are impacted? Are existing biases exacerbated?) as well as the ways in which the model's adoption may shift incentives.

We emphasize that these assessments are often complex and nuanced; as Jacobs and Wallach [2019] observe, "A feature, not a bug, of validity is that it is not a binary to be achieved, or a box to be checked: it is always a matter of degree, backed by critical reasoning."

8.2.1.2 Reliability

Reliability ("can it be repeated?") captures the degree to which our proposed measurement model would yield similar results if measurements were repeated. "Measures that are governed primarily by noise—due to inference, stability of the quantity measured, or imprecise measurement tools or meaningless scales—are of limited use" [Jacobs and Wallach, 2019]. Importantly, unreliable measurements call many types of validity into question; for example, a computational measurement that is sensitive to randomness in the initialization or to small amounts of noise in the training data is unlikely to fully capture our theoretical construct of interest (content validity) or to correlate with external measurements of the same construct (convergent validity).

8.2.2 Measurement modeling and bias

Why do we adopt the measurement modeling framework to examine bias in NLP systems? Our inspiration comes from our critical analysis in Ch. 5, in which we found that the word "bias" has come to describe many possible undesirable system behaviors. As we showed, this overuse has obscured important differences between how bias is conceptualized and operationalized between papers. The failure to distinguish these different behaviors can be thought of as a measurement modeling concern, in which neither the specific behaviors that practitioners wish to quantify, nor their relationships to the bias metrics that are proposed to measure them, are explicitly articulated.

This failure forecloses the effective evaluation of approaches to quantify and mitigate bias. For example, recently Gonen and Goldberg [2019] raised criticisms of one widely used approach for "debiasing" word embeddings, showing that although its application decreases the level of gender bias according to one metric, according to other reasonable metrics it leaves the embedding space largely unchanged. How should we understand these different approaches? Are they quantifying the same phenomenon? Which metrics, if any, should a practitioner rely on?

We argue that this framework provides the tools to evaluate approaches for quantifying bias by requiring us to separately articulate both the behaviors of concern and the metrics proposed for measuring them. By disentangling the two, this process enables rigorous analysis of each and identification of *mismatches* between undesirable behaviors and the metrics used to quantify them. Moreover, we can better resolve disagreements by understanding which are about what we ought to be measuring, versus how we ought to be measuring.

In this chapter, we seek to make these constructs and measurements, as well as the attendant assumptions about the relationships between them, explicit. To do so, we reframe existing approaches as measurement models. For a range of these approaches, we identify the measurement model and the construct(s) implicitly under measurement, and interrogate the (mis)matches between construct and operationalization. We also examine work in the quantitative social sciences that uses bias-in-embeddings approaches, and analyze how the measurement models implicitly provided by these approaches differ from the superficially similar ones provided by NLP practitioners quantifying bias in embeddings.

8.3 Measuring bias in embeddings

In this section, we will use x to refer to a word (type), and $\vec{x}$ to refer to the vector associated with that type.

A number of approaches for quantifying bias in embedding spaces have emerged, including analogy tests [Bolukbasi et al., 2016], subspace projection approaches [Bolukbasi et al., 2016, Dev and Phillips, 2019, Manzini et al., 2019, Kaneko and Bollegala, 2019, i.a.], the Word Embedding Association test [Caliskan et al., 2017], and natural language inference-based probes [Dev et al., 2019]. For each approach, we describe the measurement model, identify the harm(s) that are implicitly under measurement, and interrogate the construct validity and reliability of the measurement.

8.3.1 Analogy tests

Word embeddings famously capture many word relationships as linear substructures [Mikolov et al., 2013]. Formally, let x_1, x_2, y_1, y_2 be words such that there is an analogous relation between pairs (x_1, y_1) and (x_2, y_2), typically written $x_1 : y_1 :: x_2 : y_2$. This relation can be semantic, e.g., *man : king :: woman : queen*, or morpho-syntactic, e.g., *walked : walking :: swam : swimming.*

One longstanding approach for evaluating the quality of a trained embedding (without regard to bias) is by constructing a dataset of such analogies, and for each analogy querying

$$x_1 : y_1 :: x_2 :? \tag{8.1}$$

by finding the vector closest to $\vec{y_1} - \vec{x_1} + \vec{x_2}$ that is not itself x_1, x_2, or y_1, and checking if the returned vector matches $\vec{y_2}$.

More recently, Bolukbasi et al. [2016] use this approach to identify gender bias by setting (x_1, x_2) to be the gendered pair (she, he). Intuitively, for every pair of words (y_1, y_2), their approach scores the similarity between $\overrightarrow{y_1} - \overrightarrow{y_2}$ and the gender direction of $\overrightarrow{he} - \overrightarrow{she}$. Formally, each pair is scored as:

$$\begin{aligned} S_{(x_1,x_2)}(y_1, y_2) &= \cos(\overrightarrow{x_1} - \overrightarrow{x_2}, \overrightarrow{y_1} - \overrightarrow{y_2}) \\ &\qquad \text{if } ||\overrightarrow{y_1} - \overrightarrow{y_2}|| \leq \delta \\ &= 0 \text{ otherwise} \end{aligned} \tag{8.2}$$

where the threshold δ is set to 1.

The top 150 analogies meeting the threshold are then evaluated by U.S.-based crowdworkers, who identified 48% (72/150) of them as sensible analogies by at least half of the workers annotating them, and 19% (29/150) as exhibiting gender stereotypes.

Bolukbasi et al. [2016] propose an approach to debias the embedding, which they consider successful because the analogy scoring and crowdsourced evaluation procedure, repeated on the debiased embedding, results in 6% of the top 150 output analogies evaluated as stereotypical, compared to the original 19%.

8.3.1.1 Evaluating analogies as measurement models

Construct A theoretical construct that we might reasonably consider Bolukbasi et al. [2016] to be operationalizing is *gender stereotyping*, as the bias metric counts the number of analogies evaluated to be gender stereotypes by human workers.

We emphasize that in this setting, the trained embedding is treated as fixed; that is, the question Bolukbasi et al. [2016] attempt to answer is: Given a fixed embedding trained on a given corpus with a given embedding algorithm, what amount of occupational gender stereotyping is present? Therefore, the measurement models under analysis here do not include the choice of corpus and embedding algorithm, and so we do not critique the validity and reliability of the measurements on those grounds.

Construct validity The analogy test as a measurement for gender stereotyping immediately encounters some serious issues. First, Schluter [2018] and Ethayarajh [2019] point out that there is no *a priori* reason to expect that systems trained based only on the distributional hypothesis, with no other constraints or resources, should yield word relationships that are well-represented by geometric translations, ignoring both other linear and non-linear structures.

Additionally, Bolukbasi et al. [2016] themselves, as well as Schluter [2018] and Nissim et al. [2020], raise several issues with the analogy method of measuring the presence of gender stereotypes in embeddings. First, they note that $\overrightarrow{y_2}$ is restricted to exclude the premise vectors $\overrightarrow{x_1}, \overrightarrow{x_2}$, and $\overrightarrow{y_1}$; this may force analogies that demonstrate stereotypes when an unrestricted search does not (e.g., the answer to *he* : *doctor* :: *she* ::? is in fact *doctor*, unless restricted, in which case *nurse* results).[1] Schluter [2018] finds that across different analogy tasks, "between 15-60% of the time the system predicts a premise vector on the GOOGLE analogy data." Finally, Schluter [2018] notes that vectors are almost always normalized before analogy tests are performed, distorting the spread of the embeddings.

Based on these immediate concerns, we conclude that face validity—the most essential of the many components of construct validity—is seriously threatened, suggesting that the analogy test is not generally suitable as a measurement model for gender stereotyping in embedding spaces.

Reliability We observe that the reliability of the analogy test is also threatened by the fact that the measurement relies on the evaluation of crowdworkers; due to the subjectivity of the task and its reliance on particular cultural contexts, the fact that crowdworkers are not trained for the task, and the fact that only half of annotators per item were required to agree on the gender stereotype judgment, similar measurements might not be obtained if the task was repeated.

8.3.2 Subspace projection

Bolukbasi et al. [2016] introduce a metric for the amount of bias in an embedding space via subspace projection. Formally, let $D = \{D_1, \ldots, D_n\}$ be a set of defining sets, where each defining set D_i is a set (usually a pair) of word vectors $\{x_i, y_i\}$ such that x_i and y_i only

[1] Note that in their paper, Nissim et al. [2020] make an erroneous objection based on a misreading of Bolukbasi et al. [2016]; the latter upper bound the distance permitted between $\overrightarrow{y_1}$ and $\overrightarrow{y_2}$, which the former misread as a lower bound.

differ with respect to one property, such as gender. Examples for gender include $(\overrightarrow{he}, \overrightarrow{she})$, $(\overrightarrow{father}, \overrightarrow{mother})$, and $(\overrightarrow{king}, \overrightarrow{queen})$.[2]

Let E be a trained normalized embedding, and define μ_i to be $\frac{1}{D_i}\sum_{w \in D_i} \overrightarrow{w}$. Then let the *gender direction* g_E be the vector found by applying SVD or PCA to $\mathbf{C}$ and taking the first principal component, where

$$\mathbf{C} = \begin{bmatrix} \frac{1}{|D_1|}\sum_{\overrightarrow{w} \in D_1} (\overrightarrow{w} - \mu_1)^T (\overrightarrow{w} - \mu_1) \\ \vdots \\ \frac{1}{|D_n|}\sum_{\overrightarrow{w} \in D_n} (\overrightarrow{w} - \mu_n)^T (\overrightarrow{w} - \mu_n) \end{bmatrix}. \tag{8.3}$$

Then, what Bolukbasi et al. [2016] call the amount of *direct bias* can be measured in the following way. Let NEUT be a set of words that should be gender-neutral words. Then the bias of E is defined to be

$$\text{bias}_c(E) = \frac{1}{|\textsc{Neut}|} \sum_{w \in \textsc{Neut}} |\cos(\overrightarrow{w}, g_E)|^c, \tag{8.4}$$

where c is a parameter controlling the strictness of the measure.

This metric can be generalized to any bias subspace B formed by taking the first k principal components that result from applying SVD or PCA to $\mathbf{C}$. This is equivalent to measuring the average projection of words in NEUT onto the bias subspace; intuitively, according to this metric, bias is minimized if each neutral word's vector is equidistant from each member of any given defining set. For example, $\overrightarrow{programmer}$ should be no closer to $\overrightarrow{man}$ than to $\overrightarrow{woman}$.

Bolukbasi et al. [2016] find that for NEUT a set of 327 occupation words and *w2vNews* an embedding trained using `word2vec` on a corpus of Google News documents, bias_1(w2vNews) = 0.08. They repeat this measurement on an embedding trained on using GloVe [Pennington et al., 2014] on a web corpus and report "highly consistent" bias scores.

[2] As Bolukbasi et al. note, it is likely the case that these pairs do not actually differ only with regard to a single property such as gender; for example, one word in a pair may be polysemous.

Gonen and Goldberg [2019] critique Bolukbasi et al.'s [2016] proposed debiasing approach, finding that the approach does not fully remove gendered associations in the embedding space. In particular, they find that occupation words still cluster according to the gender with which they are stereotypically associated; for example, while $\overrightarrow{nurse}$ is equidistant from $\overrightarrow{he}$ and $\overrightarrow{she}$, it is still close to $\overrightarrow{caregiver}$, $\overrightarrow{receptionist}$, and $\overrightarrow{teacher}$ and distant from $\overrightarrow{programmer}$ and $\overrightarrow{pilot}$.

8.3.2.1 Evaluating subspace projection as a measurement model

Construct The construct under measurement here is not the same as that measured by the analogies, which was a general notion of gender stereotyping. Rather, due to the definition of NEUT as a set of *occupation* words, we identify Bolukbasi et al.'s [2016] subspace projection metric as a narrower measurement of *occupational* gender stereotyping in the embedding.

Our evaluation of subspace projection as a measurement model considers how well the metric—including properties of projection and choices of defining sets and gender-neutral words—captures our theoretical understanding of occupational gender stereotyping (validity) and how much the metric might vary across repeated measurements (reliability) given a fixed embedding.

Construct validity The subspace projection metric passes the sniff test as a measurement of occupational gender stereotyping; the idea that occupations' gendered associations can be captured by relative cosine distance in the embedding space is intuitively reasonable, and aligns with our theoretical understanding of stereotyping as "systematic asymmetry in word choice" [Beukeboom, 2014], where occupational gender stereotyping is an asymmetry in which occupational words are used to describe men or women.

However, one threat to content validity arises from the metric's implicit assumption that any such gendered associations are effectively captured by this particular type of geometric relationship. Indeed, Gonen and Goldberg's [2019] finding that a metric measuring a different kind of geometric relationship—clustering—recovers gender associations in a "debiased" embedding indicates a potentially serious mismatch between the construct and this metric, and further threatens convergent validity as it provides divergent measurements of the same construct.

We further observe that this metric was designed for an *English* embedding space, and relies on the fact that the English language lacks grammatical gender.[3] In many languages with grammatical gender, nouns are gendered and dependent adjectives, articles, or verbs must agree with the noun's gender. As Zhou et al. [2019] and Zmigrod et al. [2019] point out, occupation words in such languages[4] therefore automatically carry gender information; the subspace metric applied in such embedding spaces would therefore fail to separate undesirable gendered associations from obligatory morphological agreement.[5] For these languages, discriminant validity is therefore threatened.

Dev et al. [2019] point out that "there is a mismatch between what approaches [internal to vector spaces] measure (vector distances or similarities) and how embeddings are actually used (as features for deep neural networks)." Consequently, while the subspace projection metric may effectively measure various kinds of linear stereotyping in the embedding space, the measurements tell us little about how that stereotyping might manifest as embeddings are used in systems trained for any number of downstream tasks, giving rise to consequential validity concerns. For example, what happens when embeddings are fine-tuned? Does stereotyping in the spatial geometry of an embedding yield stereotyped output for a downstream model?

Reliability Given a fixed embedding, the components of the subspace projection metric that are chosen by the practitioner are the set of gender-neutral occupation words and the defining sets of gendered pairs. Abdin et al. show that the metric is sensitive to the choice of defining sets; for example, they find that 19.7% of phrases show a male association when projected on to a gender direction computed using Bolukbasi et al.'s [2016] defining sets, but

[3] English does have some gendered pronouns.

[4] where the gender classes correspond to human gender

[5] English also has many non-gender-neutral occupation words, such as *actor* and *actress*, or *policeman* and *policewoman*, the use of which is decreasing in favor of gender-neutral occupation words or phrases such as *police officer*; the subspace projection metric would fail on these as well.

a female association when projected onto a direction computed using Garg et al.'s [2018].[6] The noisiness of measurements across the relatively arbitrary choice of defining sets both indicates a lack of reliability and calls into further question the types of validity discussed above.

8.3.3 WEAT and WEFAT

Caliskan et al. [2017] introduce the Word Embedding Association Test (WEAT) and the Word Embedding Factual Association Test (WEFAT). Inspired by the Implicit Association Test in psychology [Greenwald et al., 1998], WEAT examines the relative relationship between two sets of target words (e.g., European American and African American names) and two sets of attribute words (e.g., pleasant and unpleasant words). The test measures the difference between the two target groups' relative similarities to the attribute groups; for example, under the null hypothesis, the relative similarities of European American names to pleasant and unpleasant words is the same as the relative similarities of African American names to pleasant and unpleasant words.

Formally, let X and Y be equally-sized sets of target words, and let A and B be sets of attribute words. For a given word w, let

$$s(w, A, B) = \text{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} \cos(\vec{w}, \vec{b}). \tag{8.5}$$

Then the test statistic is defined as

$$S(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B). \tag{8.6}$$

The effect size is computed as

$$\text{effect}(X, Y, A, B) = \frac{\text{mean}_{x \in X} s(x, A, B) - \text{mean}_{y \in Y} s(y, A, B)}{\text{std-dev}_{w \in X \bigcup Y} s(w, A, B)}. \tag{8.7}$$

[6]Though we note that the subspace metric is defined over the absolute value of the projection onto the gender direction.

	Target words	Attribute words
1)	Flowers / insects	Pleasant / unpleasant 1
2)	Instruments / weapons	Pleasant / unpleasant 1
3)	European American / African American names	Pleasant / unpleasant 1
4)	European American / African American names	Pleasant / unpleasant 2
5)	European American / African American names	Pleasant / unpleasant 3
6)	Male / female names	Career / family
7)	Math / arts	Male / female terms
8)	Science / arts	Male / female terms
9)	Mental / physical disease	Temporary / permanent
10)	Young / old names	Pleasant / unpleasant 1

Figure 8.1: Tests performed by Caliskan et al. [2017]. Pleasant vs. unpleasant 1, 2, and 3 refer to different sets of pleasant and unpleasant words.

Caliskan et al. perform ten tests using this measurement, eight of which (tests 1–3 and 6–10) are replications of the IAT, and two of which are new. Table 8.1 gives the target and attribute word sets for each test.

WEFAT, in contrast, is intended to measure the degree to which the embedding space captures empirical facts about the world. To do so, it computes the statistic $s(w, A, B)$ for each word $w \in W$ and attribute word sets A and B and a corresponding value p_w about the world; for example, if w is *engineer*, then p_w might be the percentage of engineers who are women in the United States. A linear regression is then performed to assess how predictable p_w is from $s(w, A, B)$.

8.3.3.1 Multiclass bias

Manzini et al. [2019] define a different metric inspired by WEAT, which is intended to handle multiclass settings (e.g., gender, race, and religion). Let T be a set of target words and $A_1, \ldots, A_N$ be sets of attribute words. Then for a given $w \in X$,

$$S(t, A_i) = \frac{1}{N} \sum_{a \in A_i} 1 - \cos(\overrightarrow{t}, \overrightarrow{a}) \tag{8.8}$$

Then the mean average cosine similarity $MAC(T, A)$, a WEAT-like metric modified for the multiclass setting, is defined as

$$MAC(T, A) = \frac{1}{|T||A|} \sum_{t_i \in T} \sum_{A_j \in A} S(t_i, A_j). \tag{8.9}$$

8.3.3.2 Evaluating WEAT approaches as measurement models

Construct Among its tests, WEAT measures several different constructs. By examining the target and attribute words (reproduced for tests 3–10 in the Appendix), we identify rows 3–5 in Table 8.1 as implicit measurements of racial valence attributions (categorized in our taxonomy as a kind of stereotyping) while 6–8 are measurements of gender stereotyping in an embedding—specifically, gender-career, gender-math, and gender-science stereotyping. Test 9 is a measurement of stigma related to mental illness, which may be thought of as a type of stereotyping, while test 10 is a measurement of age stereotyping. Each test must therefore be evaluated with respect to the construct that it implictly operationalizes.

Construct validity As measurements grounded in the IAT, the tests capture important aspects of our theoretical understanding of different kinds of stereotyping, and the statistically significant effect sizes matching IAT effect sizes provide predictive validity.

Ethayarajh et al. [2019] discuss a potential threat to WEAT's validity by pointing out that a word embedding's squared norm is linear in the log probability of that word in the underlying corpus. Therefore, the test implicitly requires that words in its attribute sets occur with roughly equal frequency in the corpus; if they differ substantially in frequency, WEAT's test statistic may overestimate the association. This is true of any embedding trained with an algorithm that implicitly performs matrix factorization, including `word2vec` trained using skip-gram negative sampling (SGNS) and GloVe.

In addition, like the subspace projection metric, WEAT was designed for the English setting and may not be effective for languages with grammatical gender; McCurdy and Serbetçi [2017] find that WEAT effect sizes are statistically significant for two languages with natural gender (English and Dutch), but not for two languages with grammatical gender (German and Spanish).

Reliability As with subspace projection, WEAT is dependent on the choices of attribute and target words, and small variations in the choices of these words may affect measured effect sizes. WEAT effect sizes may also be dependent on corpus size and pre-processing decisions; Lauscher and Glavaš [2019] find that effect sizes vary considerably across embedding spaces trained with different embedding algorithms, hypothesizing that this variation may be due in part to the different algorithms' pre-processing procedures. They also hypothesize that the smaller effect sizes measured for some languages may be a result of those languages' smaller training corpora.

8.3.4 NLI probes

Dev et al. [2019] propose a method of testing for bias in embeddings, both type-level and contextualized, through a natural language inference task; in this section, we address the type-level embeddings. In the NLI task, a model is given a pair of sentences, a *premise* and a *hypothesis*, and the task is to predict whether the relationship between them is one of *entailment*, *contradiction*, or *neutrality*. Consider the following example (drawn from Dev et al.):

(P) The driver owns a cabinet.

(H1) The man owns a cabinet.

(H2) The woman owns a cabinet.

where P is a premise sentence and H_1, H_2 are two possible hypotheses. The correct NLI prediction for both (P, H_1) and (P, H_2) pairs is a neutral one, since P neither entails nor contradicts either H_1 and H_2. Dev et al. use this intuition to develop a dataset of such premise-hypothesis pairs constructed according to three different templates, each purporting to capture a different kind of bias. The first is occupational gender stereotyping:

(P1) The [OCCUPATION] [VERB] a/an [OBJECT].

(H1) The [GENDERED WORD] [VERB] a/an [OBJECT].

where [VERB] and [OBJECT] are identical within a pair.

The second involves trait and nationality words:

(P2) The [TRAIT] person [VERB] a/n [OBJECT].

(H2) The [DEMONYM] person [VERB] a/n [OBJECT].

where trait words include words such as *awful*, *dishonest*, *neat*, and *professional*.

The last involves trait and religious words:

(P3) The [TRAIT] person [VERB] a/n [OBJECT].

(H3) The [ADHERENT] person [VERB] a/n [OBJECT].

Dev et al. propose three bias metrics, which are intended to capture how far a model's prediction deviates from the ideally neutral prediction across the dataset of sentence pairs:

- Net neutral: "Computes the average probability of the neutral label across all sentence pairs."
- Fraction neutral: "Computes the fraction of sentence pairs that are labeled as neutral."
- Threshold(τ): "A parameterized measure that reports the fraction of examples whose probability of neutral above τ."

To get the scores, Dev et al. train a decomposable attention model on the SNLI dataset [Bowman et al., 2015] using a GloVe embedding as input, which is down-sampled and fine-tuned.

8.3.4.1 Evaluating NLI probes as measurement models

Construct It is straightforward to identify the three templates as operationalizing notions of occupational gender stereotyping, national stereotyping, and religious stereotyping, respectively. It is more difficult to identify the exact constructs under measurement, however. One possibility, which is suggested by the paper's stated goal of presenting "a strategy for probing word embeddings for biases," is that the constructs under measurement are occupational gender, national, and religious stereotyping *in a fixed embedding*, much like the constructs under measurement in previous sections; this is supported by the paper's statement: "We argue that biased representations lead to invalid inferences, and the number of invalid inferences supported by word embeddings (static or contextual) measures their bias."

Another possibility is that the construct under measurement is stereotyping *in a fine-tuned embedding.*[7]

In either case, the proposed measurements, which we evaluate in this section, are the three proposed bias metrics calculated from the prediction of the SNLI-trained NLI model over the sentence pairs generated from the three templates.

Construct validity Under the first interpretation, one of the constructs under measurement (occupational gender stereotyping in an embedding) is also what Bolukbasi et al. [2016] propose to measure. Therefore, the fact that "debiasing" using Bolukbasi et al.'s method—which effectively decreases the level of occupational gender stereotyping according to their subspace projection metric—also reduces the level of occupational gender stereotyping according to this measurement provides some evidence of convergent validity.

Nevertheless, the measurement's validity is threatened in a few important ways. First, as Dev et al. point out, analysis of bias in embedding spaces often suffers from a mismatch between the bias metrics, which generally involve calculations of similarity in vector space, and the way embeddings are ultimately used as features in neural networks; as we have discussed, this presents a threat to consequential validity for these measurements. However, it is precisely this mismatch that threatens this measurement; if the constructs under measurement are actually the level of different types of stereotyping in an embedding, then an analysis of a downstream model trained on that embedding is a much less direct way of measuring them than calculations of similarity in a vector space.

In fact, Dev et al. point out this indirectness, noting that "[e]ither model error or an underlying bias in GloVe could cause [an] invalid inference." The training process likely alters the pre-trained GloVe embedding significantly, as it is downsampled from 300 dimensions to 200 before fine-tuning. Moreover, fine-tuning on SNLI does not expose the model to the occupation, demonym, and religious adherent words evenly; for example, *British* occurs 156 times in the training portion of SNLI while *Belarusian* does not occur at all. It is therefore

[7] We discount the possibility that the constructs are different types of stereotyping *from the NLI system* due to the paper's described goal.

likely that the measurements reflect not only GloVe's representation bias, but also model and training effects; this conflation represents a potential threat to discriminant validity.

Additionally, because the metric assesses each item's deviation from neutral independently, it may overestimate the true measurement; for example, equal output probabilities of $[\textsc{Entail} = 0.1, \textsc{Neut} = 0.1, \textsc{Cont} = 0.8]$ for both a premise and male-hypothesis pair and a premise and female-hypothesis pair would contribute strongly to the bias metrics, but may not indicate gender stereotyping in the underlying embedding.

Under the second interpretation of the constructs as different types of stereotyping in the fine-tuned embedding that results from training the NLI model, we observe that the fact that the pre-trained embedding shifts and is down-sampled during training no longer represents a threat to discriminant validity; however, the possible conflation of model error and representation bias continues to do so.

Reliability We observe multiple possible threats to reliability. First, the metrics may be sensitive to the choice of word sets used to fill out the templates. Second, the metrics are calculated from the predictions of the trained NLI model; although Dev et al. retrain the model multiple times to show that results are stable across runs, it is unknown how much the measurements depend on their choices of model architecture and training.

8.3.4.2 Extending measurements of gender stereotyping

As we discussed in §8.3.2, these metrics (with the exception of analogies) examine narrow subtypes of gender stereotyping—occupational gender stereotyping, gender-math stereotyping, gender-science stereotyping—which neglects many other types of gender stereotyping, not to mention types of stereotyping not related to gender, of which WEAT examines only a few. Here, we draw on the social psychology literature to briefly explore other aspects of how stereotyping is communicated through language that are not operationalized in in the metrics previously described.

One such aspect is the Stereotype Content Model, which proposes that stereotypes are formed along two dimensions: warmth and competence [Fiske, 2015]. Each dimension encompasses a range of attributes; competence includes traits such as intelligence, skill, agency, and confidence, while warmth includes traits such as kindness, friendliness, and sincerity. Groups may be evaluated as either warm or competent, both, or neither; for example, men are often evaluated as competent and women as warm [Fiske, 2015].[8] One situation where this difference manifests is in the language used in written evaluations, such as in recommendation letters; women are more likely to be described using *compassion* words such as *caring*, *compassionate*, and *empathetic*, and white applicants more likely to be described as *exceptional*, *outstanding*, and *best* [Ross et al., 2017].

Another model for the linguistic maintenance and transmission of stereotypes is the Linguistic Intergroup Bias model, which proposes asymmetries in the use of different levels of linguistic abstraction [Maass, 1999, Menegatti and Rubini, 2017]. Specifically, the models propose that positive in-group and negative out-group behavior are more likely to be described in abstract terms, such as adjectives, while negative in-group behavior and positive out-group is likely to be described in concrete terms, such as action verbs. Similarly, the Linguistic Expectancy Bias model proposes that role-conforming behavior is described in abstract terms, and non-conforming behavior in concrete terms. For example, a man who does well in a class might be described with, *He is intelligent*, while a similarly-performing woman might be described with, *She did well in the class.* The former, abstract description communicates enduring, static properties, which are difficult to disconfirm, while the latter, concrete description communicates one-time or isolated events. Wagner et al. [2016] demonstrate that this asymmetry is present on Wikipedia, where abstract terms are more likely to be used to describe positive aspects in men's biographies, and negative aspects in women's biographies [Wagner et al., 2016].

We provide this discussion of models of stereotype transmission in language to illustrate that the harms potentially arising from NLP systems may have deep roots in existing literature from which analyses can benefit; once a harm (in this case stereotyping) has been concretely identified as a construct of interest, the process of demonstrating validity for the proposed measurement should draw on such established literature.

[8] Recent work has argued for models with three or more dimensions, suggesting that warmth conflates different concepts such as sociability and morality [Leach et al., 2007].

8.3.4.3 Why measure bias in embeddings?

As Dev et al. [2019] point out, and as we discussed in §8.3.2 and §8.3.4, many measurements of bias in embedding spaces face inherent threats to consequential validity due to the mismatch between what they measure (bias in a fixed embedding, evaluated intrinsically) and how embeddings are ultimately used (as input to deep neural networks, typically fine-tuned).

This observation raises important questions about the purpose of measuring bias in embeddings. In this thesis, we have proposed that analyses of bias in NLP systems should focus on *harms* arising from these systems, a proposal that drives the taxonomy we presented. If so, then our goal in examining bias in embeddings should be explicit evaluations of harms. However, because embeddings are not generally used by themselves, perhaps harms cannot be said to arise directly from them, and therefore we suggest that our question—what harms arise from an embedding?—may not be a well-formed one.[9] Rather, perhaps we should aim to answer the question, What happens when I use this embedding in a system that does *X*, where *X* is some NLP task? This is a question that Dev et al.'s [2019] analysis of NLI systems implicitly provides evidence for, even though their question is framed in terms of the embedding space.

8.4 Measuring bias in the world: "Geometry of culture"

Here, we discuss work in the quantitative social sciences that measures bias in embedding spaces. These approaches posit some latent bias in human language or semantic memory, in institutions such as the media or the judiciary, or even in individual speakers, which is measured by gathering a representative corpus, training an embedding on that corpus, and measuring bias in the embedding. Unlike work in the previous section, which is aimed at measuring—and subsequently reducing—bias in embedding spaces, the approaches described here view bias measurement in embeddings as a way of using large-scale text data to gather

[9] We can imagine settings where an embedding is the end system—for example, a system for lexicon induction.

important cultural or psychological information, with the trained embedding as a statistical summary of that cultural or psychological information.

We begin by surveying some of this work. In our summaries of these approaches, we aim to surface their conceptualizations of, and assumptions about, embeddings; as we will see, embedding spaces and their properties are conceptualized as operationalizing a range of cognitive and social properties. We then discuss the constructs and measurement models implicitly operationalized across these approaches; in particular, we observe that while the metrics used to measure bias in embeddings are frequently superficially identical to those described in the previous section, the constructs and measurement models are not, raising different validity and reliability concerns than those described in the previous section.

8.4.1 Embeddings for a language/collective cultural imagination/human semantic memory

A wide range of work has used embeddings-based methods to measure biases in human language, culture, or semantic memory. Lenton et al. [2009] examine gender stereotyping in an embedding trained with Latent Semantic Analysis (LSA) [Deerwester et al., 1990], finding that gendered referents (*man/he/him* and *woman/she/her*) exhibit greater semantic similarity to stereotypical words than counter-stereotypical ones, particularly occupational roles. Notably, Lenton et al. view the LSA training process as "model[ing] language and knowledge acquisition, in addition to (post-acquisition) meaning representation," and treat the trained embedding as a valid proxy of the human semantic network. From this perspective, embeddings-based approaches not only reflect word usage statistics, but in fact encode a theory of semantic acquisition, in which word meanings are acquired from the contexts in which they occur (a strong version of the distributional hypothesis).[10] They conclude,

> Our research shows that gender stereotypes are inherent in the very meaning of the most common social category referents for *man* and *woman*. . . . American English-speakers' understanding of the words *man*, *he*, or *him* and *woman*, *she*, or *her* is fundamentally tied to their understanding of stereotype-relevant words.

[10]For a thorough discussion of the plausibility of embeddings-based approaches as models of human cognition, see Günther et al. [2019].

> ... These results demonstrate that stereotypes permeate language at a very deep level, as LSA is carried out via the inclusion of indirect semantic associations.

Bhatia [2017] also use LSA to examine prejudice and stereotyping, finding that across models of varying dimensionality, African American names are strongly associated with negatively valenced words. Two sets of gender stereotypes based on the IAT, female/male vs. power/weakness and female/male vs. career/family, are also captured by the models.

Kurdi et al. [2019] investigate the relationship between evaluations (valence attributions) and stereotypes (trait attributions), finding that valence, warmth, and competence for social groups are tightly connected both in participant studies and in embedding space. Evaluations and stereotypes can be disassociated by participants when producing explicit judgments, but are shown to be deeply intertwined using implicit measures, such as the IAT. Therefore, in this context, the correlation of the relative distances between social groups and valence, warmth, and competence (measured using a WEAT-like test) is taken as evidence that disassociating evaluations and stereotypes requires "deliberative cognitive processes" that are not actively undertaken in everyday production of language.

Garg et al. [2018] aim to show that word embeddings accurately capture changes in gender and ethnic stereotypes latent in language over the twentieth century. Using a distance-based metric for the gender bias of an occupation word in an embedding, they demonstrate that this gender bias is better explained by crowdsourced judgments of how gender stereotyped the occupation is than by the proportion of women in that occupation, concluding that their gender bias metric is more likely to track stereotyping than actual labor force participation.

Kozlowski et al. [2019] propose that constructed dimensions of social class in embedding space effectively capture their cultural associations and connotations. Key to their approach is their assumption that the methods used to train the embeddings (the "cultural space") and construct the cultural dimensions meaningfully operationalize aspects of sociological theory. For example, embedding spaces are seen as operationalizing theories of intersectionality:

> The ability of word embedding models to simultaneously locate objects on multiple cultural dimensions, including race, gender, class, and many others, makes them a powerful tool for studies of intersectionality. ... Interrogation of the intersection of cultural categories becomes empirically tractable through word

> embedding models. ... Indeed, the empirical success of word embedding models to represent cultural dimensions promotes a radical view of intersectional identity, modeled not as a low-dimensional matrix, but rather a high-dimensional array composed of hundreds of thousands of interacting cultural associations.

The relatively high dimensionality of the embedding space is seen as extending two-dimensional Bourdieuvian representations of social spaces: "By preserving higher dimensionality in a cultural space, word embeddings can facilitate the development and testing of high-dimensional theories of how actors acquire and exploit varied cultural capitals along multiple distinct dimensions of status." Thus, the dimensions of the embedding space are anticipated to capture interpretable "dimensions of status," as measured by a version of the subspace projection approach.

Friedman et al. [2019,] propose to examine cultural gender biases worldwide by training separate embeddings on English-language tweets from each U.S. state and 99 countries. They show that measurements of gender bias in embeddings correlate with national and global gender gap statistics, ranging from the political (e.g., women in parliament) to economic (e.g., wage equality). Like Kozlowski et al. [2019], they assume that implicit cultural attitudes live in "a large volume of a culture's text" and can be analyzed by measuring linguistic bias in a trained embedding.

Lewis and Lupyan [2018, 2020] examine two hypotheses about gender bias and language: a) the "language-as-reflection" hypothesis, which posits that since language is precisely the way in which we talk about cultural stereotypes and biases, it will reflect those stereotypes and biases, and b) the "language-as-causal-factor" hypothesis, which posits that language actually exerts an influence on people's biases. In the social psychology literature, the former hypothesis is uncontroversial [Maass, 1999, Menegatti and Rubini, 2017, Ellemers, 2018, Beukeboom and Burgers, 2019].

To explore these hypotheses, they compute a linguistic bias effect size for twenty-five languages with a WEAT-like female/male vs. career/family test, finding that mean IAT bias score for a country and the linguistic bias effect size for the dominant language of that country are positively correlated. To disentangle the second hypothesis from the first, they further investigate the relationship between structural aspects of language and implicit bias, finding that two measurements of linguistic occupational gender bias are positively correlated with mean IAT bias score: the number of occupation words with gender-distinct forms in a language, and the language's occupation words' gender scores in the embedding.

8.4.2 Embeddings for genres or institutions

Another set of approaches has used embeddings-based methods to examine bias in particular institutions and genres such as the news media, the judiciary, movie and restaurant reviews, and song lyrics.

8.4.2.1 The media

Work in social psychology has used embeddings trained with LSA on news text to investigate intergroup bias and gender stereotyping via the semantic contexts of pronouns, comparing the valences of the contexts of collective vs. individual pronouns [Sendén et al., 2014] and of *he* vs. *she* [Sendén et al., 2015]. In these approaches, LSA is chosen because it is perceived to yield a reliable reflection of actual language use, due to its entirely statistical nature: "The LSA is completely data-driven, making the results resilient against influence bias from the researcher."

Bhatia et al. [2018] train embeddings with LSA on text from the 2016 election across 250 news outlets to investigate traits associated with *Hillary Clinton* and *Donald Trump*, finding that representations differed most with respect to morality traits. Here, much as with Lenton et al. [2009], semantic models are "considered to mimic human semantic learning and representation processes"; the trained embeddings for each news outlet are viewed as proxies for the human semantic representations that would result from exposure to the natural language environments of those outlets.[11]

Leschke and Schwemmer [2019] explore the effect of the 2017 Charlottesville Rally on racial bias in the media by fitting embedding models to text from a range of U.S. and U.K.

[11] Importantly, Bhatia et al. acknowledge the difficulty of disentangling causal relationships; while it is possible that exposure to natural language environments from media would have caused readers to develop particular trait associations, it is also possible that the embeddings' trait associations reflect pre-existing reader beliefs.

news sources from shortly before and after the rally; here, news articles (and embeddings trained on them) are considered to be reflective of public opinion.

Several approaches have also trained embeddings to track changes in bias over time, which is taken to reflect changes in popular conceptions or perceptions. For example, Gillani and Levy [2019] measure public perception of refugees by training dynamic embeddings on talk radio transcripts and computing associations of "outsider"-like adjectives such as *aggressive*, *frightening*, and *illegal* with refugee-related words over multiple months in 2018. Tripodi et al. [2019] examine changes in conceptions of Jewish people and Judaism over time by training embeddings on a corpus of French books and peridiocals from 1789-1914, while Wevers [2019] explore gender bias using embeddings trained on Dutch newspapers from 1950–1990.

8.4.2.2 The judiciary

Two recent works have examined bias in the judiciary using embeddings-based approaches. Rice et al. [2019] train embeddings on opinions from the U.S. Supreme Court, U.S. Court of Appeals, and appellate courts and examine the WEAT effect size for the European American/African American vs. pleasant/unpleasant test, finding that African American names are more closely associated with unpleasant words for all time periods and court levels examined.

In a departure from the work described so far, Ornaghi et al. [2019] train embeddings for individual judges in order to measure their gender attitudes. Specifically, they train embeddings on individual judges' U.S. Circuit Court opinions and examine WEAT effect sizes for two tests: a) female/male vs. positive/negative, and b) female/male vs. career/family. They find that higher effect sizes for one or the other of the tests in U.S. judges is associated with their having daughters, their exposure to female judges in court, their likelihood of voting for plaintiffs in women's rights cases, and their likelihood of reversing decisions by female judges. Here, the measurement of bias in an embedding trained on a judge's decisions is treated as a proxy for that judge's attitudes towards gender.

8.4.2.3 Other genres

Elsewhere, bias in embeddings approaches have been deployed across a variety of genres; for example, Mishra et al. [2019] employ gender-related WEAT tests on embeddings trained on Amazon movie reviews and Yelp restaurant reviews, finding statistically significant effect sizes. Similarly, Barman et al. [2019] find that embeddings trained on song lyrics yield positive WEAT effect sizes. Knoche et al. [2019] perform WEAT-like tests on embeddings trained on two politically oriented wikis, Conservapedia and RationalWiki.

8.4.3 Evaluating quantitative social science approaches

At first glance, it might appear that the constructs and measurement models implicitly operationalized in the papers described in this section are nearly identical to those described previously, particularly as the majority of them measure bias in embeddings trained with the same algorithms (e.g., `word2vec` or GloVe) and employ the same bias metrics (e.g., subspace projection or WEAT). Nevertheless, these superficial similarities obscure several important differences in both constructs and measurement models, which in turn affect possible validity and reliability critiques.

Unlike the approaches described previously, the work described in this section does not propose to measure harm *in a fixed embedding*, but instead in something in the world, such as the human semantic network, print media, judicial language, or even individual speakers. Therefore, the embedding is not part of the unobservable theoretical construct, but part of the measurement model; measured bias in the embedding is a proxy for bias in the real world thing, which cannot be directly measured.

Importantly, this means that the measurement includes not only choices related to the specific bias metric, such as subspace projection or WEAT, but also choices related to the construction of the embedding itself, such as the selection of the training corpus, selection of the embedding algorithm, choices of embedding algorithm hyperparameters, and so forth. This raises new potential validity and reliability concerns.

8.4.3.1 Construct validity

Content validity How well does a measurement—the corpus and embedding algorithm selection, training process, and bias metric—capture our theoretical understanding of a construct? In the previous section, we discussed potential mismatches between the bias metric and our understanding of constructs; here, we must additionally consider potential mismatches between details of the embedding construction process and the construct. For some constructs, it may be straightforward to assess the match between the construct and choices of corpus and embedding algorithm; for example, in an analysis of gender stereotyping in news media, a dataset of Reuters news articles [Sendén et al., 2014, 2015] or radio transcripts [Gillani and Levy, 2019] is sensible.

However, in other situations, it may be much harder to determine whether construct and measurement are well-matched. For example, if the construct of interest is cultural gender bias, a corpus of English Wikipedia text may or may not be a good representation of cultural attitudes or everyday language use in English-speaking societies. If the construct of interest is gender stereotyping in the human semantic network, we may not know enough about the process by which humans acquire semantic knowledge, about how that knowledge is drawn upon in the production of text, or about what trained embeddings are capable of learning, to assess whether any embedding algorithm is a cognitively plausible representation [Günther et al., 2019].

Moreover, we know surprisingly little from an engineering perspective about the behavior of embedding algorithms. For example, Mimno and Thompson [2017] find that the word vectors in embeddings trained using skip-gram negative sampling—a training algorithm available as part of `word2vec`—occupy a surprisingly small cone in vector space, rather than being evenly distributed throughout the space. This may affect our interpretation of word distances, since words occupying such a narrow space may yield spuriously close measurements. Such work, which is relatively recent and ongoing in NLP, suggests that much better technical understanding of embeddings may be required before we can be sure that embeddings serve as a good proxy for real world phenomena of interest.

Convergent validity For those approaches that aim to examine bias in humans, convergent validity can (and has been) established by comparing embeddings-based measurements to measurements taken in studies of human participants. For example, Kurdi et al. [2019] find that the results of two sets of IATs and the WEAT-like embeddings-based measurements support each other, providing convergent evidence.

For those approaches measuring bias in text produced by particular institutions or in particular genres, such as the judiciary or the media, validity may be established if there are existing alternative approaches to measuring that type of bias in text. For example, a wide range of work has proposed non-embeddings approaches for measuring different kinds of stereotyping in text [Fast et al., 2016, Wagner et al., 2016, Carpenter et al., 2017, Chang and McKeown, 2019, Otterbacher et al., 2017, Fokkens et al., 2018, Gálvez et al., 2018, Madaan et al., 2018, Hoyle et al., 2019, Qian, 2019, i.a.]. Measurements from such approaches that align with embeddings-based approaches may provide convergent evidence.[12]

Discriminant validity Another potential concern is that measurements of bias in embeddings may be reflections of properties of the input corpus, or artifacts of the quirks of the embedding algorithm, as the geometry of the trained embedding is a function of both. Moreover, properties of the corpus may be an artifact of the sampling procedure that yielded any particular corpus. Therefore, any measurement of bias potentially reflects any number of artifacts of the process used to generate the embedding.

Evaluating embedding algorithms Spirling and Rodriguez [2019] offer a guide to training and using embeddings for political science practitioners. In this guide, they propose a method for evaluating an embedding algorithm. Inspired by the Turing test, the proposed evaluation method aims to examine "predictive performance" by comparing an embedding's nearest neighbor outputs for a given set of words with "nearest neighbor" outputs produced by humans:

---

[12] We note that this may not be considered an entirely external measurement, since it operates over the same text as the embeddings-based approach, and the selection of the corpus itself is part of the measurement. Nevertheless, employing non-embeddings-based approaches may give insight as to whether the measured bias is an artifact of the embedding algorithm or the bias metrics.

> [A]n embedding model achieves 'human' performance if human judges—crowd workers—cannot distinguish between the output produced by such a model from that produced by independent human coders. . . . If a set of human judges are on average indifferent between the human responses to a prompt and the model's responses, we say we have achieved human performance with the model. By extension, a model can achieve *better than human* performance by being on average preferred by coders. Naturally, models may be *worse than human* if the judges like the human output better.

We raise several points of concern with this evaluation task. First, the formulation as an evaluation of performance against humans is not meaningful, as the task is not well-defined for humans. Since the query terms used here are political science ones such as *democracy*, nearest neighbors likely capture not just semantic knowledge inherent in the definition of the word, but also related cultural or world knowledge. Without disentangling which of these is intended to be measured by the notion of predictive performance, it is not clear what good human performance looks like, much less good predictive performance for embeddings.

Second, any measurements from the task likely contain artifacts of both the input corpus and the embedding algorithm, when only measurements of the performance of the latter are desired. Suppose, for example, that we train an embedding on a corpus in which we expect a particular kind of language use, such as a corpus of articles from right-wing news outlets. Such an embedding is likely to return very different sets of nearest neighbors for our test queries than human crowdworkers. Such a result is expected, as a human has not been "trained" on the news outlet corpus. Thus, the "failure" of the embedding on human assessment might be due to the embedding algorithm or to particular regularities of the language used in its training corpus. Indeed, this is desirable—interest in bias measurements over embeddings exists precisely because embeddings are thought to pick up on such regularities; therefore, a "failure" of this type may be exactly what we are after.

8.4.3.2 Reliability

The most significant threat to embeddings-based measurements is embeddings' notorious instability. Recent work has shown that embeddings are substantially affected by choices of training hyperparameters; for example, embeddings trained on the same corpora with different dimensionalities (only one of many training hyperparameters) yield radically different subspace bias projection scores. One work comparing a bias scores on datasets from Twitter and Wikipedia has shown that trained embeddings with dimensionality lower than 100 yield larger bias scores for the Twitter dataset than Wikipedia, but the reverse is true when embeddings are trained with dimensionality greater than 100 [Mirzaev et al., 2019].

Embeddings are also sensitive to small variations in training corpora; Antoniak and Mimno [2018] show significant variation in query items' nearest neighbors when documents are presented in different orders, as well as when documents are bootstrapped. Moreover, even when presented with identical training corpora, embedding algorithm, and training hyperparameters, embeddings have been shown to be sensitive to randomness in training initializations [Antoniak and Mimno, 2018]; this is the case even for relatively high-frequency words [Wendlandt et al., 2018].

8.5 Measuring bias in NLP systems

In this section we examine approaches to measuring bias across a variety of NLP tasks, including sentiment analysis, hate speech and toxicity detection, machine translation, coreference resolution, and language modeling. As before, we focus on what constructs different approaches implicitly operationalize, and identify (mis)matches between constructs and the measurement models provided. As we found in our general critique in Ch. 5, we find that different approaches to quantifying bias for the same task often implicitly operationalize different constructs, reflecting different unstated assumptions about what constitutes bias in these systems; we aim to surface these assumptions.

8.5.1 Machine translation

We describe a number of approaches to measuring gender bias in machine translation, the task of automatically translating text from one language to another. For example, Prates et al. [2019] examine translations from languages with gender-neutral pronouns to English, employing templates of the format ⟨*Pronoun*⟩ *is* ⟨*occupation*⟩ and ⟨*Pronoun*⟩ *is* ⟨*adjective*⟩; from their perspective, for each category of occupation, such systems should output translations using *He* and *She* with roughly equal frequency. They consider a system to exhibit "negative gender bias" for occupations if "the frequency of male defaults overestimates the (possibly unequal) distribution of male employees per female employee in a given occupation"—clearly a measurement of occupational gender stereotyping. The construct operationalized by the evaluation of translation of adjectives, however, is less straightforward to identify, since the adjectives are the 1000 most frequently used adjectives in English, and therefore not selected to capture any particular aspect of stereotyping.

Cho et al. [2019] evaluate gender bias in Korean-English translation systems. Unlike Prates et al. [2019], the measurement here is explicitly designed to reward gender neutrality. Let p_w be the portion of Korean sentences translated into English with female pronouns, p_m the portion translated with male pronouns, and p_n the portion translated gender-neutrally (e.g. with *This person*); where $p_w + p_m + p_n = 1$. Then the metric P_s is defined as

$$P_s = \sqrt{p_w p_m + p_n} \tag{8.10}$$

which is at its highest when $p_n = 1$, and for a fixed p_n is highest when p_w and p_m are equal (that is, the system outputs male and female pronouns with equal frequency). The metric is applied to translations of seven different datasets containing generated sentences that are a) informal, b) formal, c) impolite, d) polite, e) negative, f) positive, and g) occupation-related, operationalizing different aspects of gender stereotyping.

Font and Costa-jussà [2019] propose an occupations test over a synthetic dataset generated using the template *I've known her/him/Mary/John for a long time, my friend is a/an* ⟨*occupation*⟩ for English-Spanish translation. Bias is measured as the percentage of the time that *friend* is correctly translated; when presented with *her/Mary* it should be translated as *amiga*, and with *him/John* as *amigo*. Much like previous analyses of bias in machine translation, this is an analysis of occupational gender stereotyping. Unlike previous analyses, this is a test in which the machine translation has access to the correct answer, as it is provided with a gendered pronoun; therefore, a "biased" system must actively disregard this context. In contrast, the tests involving translations from languages without gender pronouns into English require a system to produce output without input signaling the correct pronoun.

Stanovsky et al. [2019] propose yet a different test to assess gender bias in machine translation. They use Winogender [Rudinger et al., 2018] and WinoBias [Zhao et al., 2018], two datasets designed to capture occupational gender stereotyping in coreference resolution by assessing models' ability to resolve pronouns of different genders to a variety of occupation words. For example, in the sentence "The doctor asked the nurse to help her in the procedure," *the doctor* should be correctly resolved to *her*.

Stanovsky et al. adapt these datasets for evaluating bias in machine translation systems as follows. Each sentence in the combined datasets is translated using commercial machine translation systems into languages with morphologically marked grammatical gender, and the grammatical gender of the occupation word's translation is identified and compared to the correct gender. For example, if our sentence is translated to "El doctor le pidio a la enfermera que le ayudara con el procedimiento," the occupation *the doctor*—which we know from the original pronoun context resolves to *she*—has been translated as *El doctor*, the masculine version. In addition to measuring the overall accuracy of translation systems, Stanovsky et al. propose two metrics for gender bias:

- Δ_G: the difference in F_1 score between male and female translations
- Δ_S: the difference in F_1 score between stereotypical and non-stereotypical gender role assignments

where F_1 is the harmonic mean of precision and recall.

Our most immediate observation across these approaches is that of the many possible harms we discussed in Ch. 7, they operationalize a narrow slice of "gender bias," namely stereotyping; Stanovsky et al. alone of these approaches measure both a quality of service harm—systems exhibit reduced performance on female occupation words compared to male occupation words—and occupational gender stereotyping—across genders, systems perform better "when presented with pro-stereotypical assignments (e.g., a female nurse), while their performance deteriorates when translating anti-stereotypical roles (e.g., a male receptionist)."

We also observe that much of this problem space is restricted to the behavior of machine translation systems on pronouns—how often systems generate female- or male-associated pronouns, how often they generate gender-neutral pronouns, or how often machine translation systems can overcome stereotypes to translate pronouns correctly. This is perhaps understandable; pronouns are a small set of words, and it is relatively easy to check how many have been treated in the desired fashion. Moreover, many of these approaches focus on occupations, which are relatively straightforward to define and develop a lexicon for. But even restricting ourselves to stereotyping, this is also profoundly limiting; as we showed in our discussion in Ch. 7, stereotypes manifest in many kinds of themes associated with different social groups, ranging from violence and aggression to femininity and deception.

8.5.2 Sentiment analysis

Sentiment analysis encompasses a suite of tasks that range from predicting the valence of a piece of text—its positivity or negativity—to predicting specific emotional intensity of text, such as anger or joy. Systems trained to perform these tasks often return a real-valued score, which may then be mapped onto a label. We discuss two pieces of recent work which have examined bias in sentiment analysis systems.

The first, Kiritchenko and Mohammad [2018], aims to analyze both gender and racial bias in sentiment analysis systems by analyzing the relationship between words associated with particular groups and sentiment predictions. To do so, they generate two types of template sentences that they fill with emotion-related words and group-related words. Examples of sentences, emotion-related, and race-related, and gender-related words are given in Tables 8.2, 8.3, 8.4, and 8.5 [Kiritchenko and Mohammad, 2018].

Template
1. ⟨Person⟩ feels ⟨emotional state word⟩.
2. The conversation with ⟨Person⟩ was ⟨emotional situation word⟩.
3. I saw ⟨Person⟩ in the market.
4. ⟨Person⟩ has two children.

Figure 8.2: Examples of template sentences, some containing both group-related words (⟨Person⟩) and emotion-related words, and some containing only group-related words [Kiritchenko and Mohammad, 2018].

Kiritchenko and Mohammad compare the outputs of 219 systems across five sentiment tasks: anger, fear, joy, sadness, and valence prediction. Bias is computed as follows: Given each system's predicted scores on the above sentence templates, compute

	Anger	**Fear**	**Joy**	**Sadness**
Emotional state	angry	anxious	ecstatic	depressed
Emotional situation	annoying	dreadful	amazing	depressing

Figure 8.3: Examples of emotional state and emotional situation words [Kiritchenko and Mohammad, 2018].

African American		**European American**	
Female	**Male**	**Female**	**Male**
Ebony	Alonzo	Amanda	Adam
Latisha	Jamel	Ellen	Frank
Tanisha	Terrence	Nancy	Roger

Figure 8.4: Examples of African American- and European American-associated names [Kiritchenko and Mohammad, 2018].

1. The difference between the predicted task score (e.g., anger intensity score) for a template sentence filled in with a female noun phrase (Table 8.5) and for the same template filled in with a male noun phrase.

2. The difference between the average predicted task score for a set of template sentences filled in with all female first names (Table 8.4) and for the set filled in with male first names.

3. The difference between the average predicted task score for a set of template sentences filled with in all African American-associated first names (Table 8.4) and for the set filled in with European American-associated first names.

Kiritchenko and Mohammad consider a system to be biased if it shows statistically significant score differences.

What harms are implicitly operationalized here? Despite the fact that some of the results align with existing gender and racial stereotypes—for example, many systems assign higher anger, fear, and sadness scores to African American-associated names—the metrics are not designed explicitly to capture stereotypes. Rather, they are intended to capture *any* difference in model outputs across group labels, a clear operationalization of *undesirable correlations*. Though this is an important analysis, we suggest that a fuller understanding

Female	Male
she/her	he/him
this girl	this boy
my wife	my husband

Figure 8.5: Examples of female and male-associated noun phrases [Kiritchenko and Mohammad, 2018].

of the harms potentially arising from this system behavior might be enabled with a grounding in our taxonomy of harms. For example, what quality of service harms arise when these sentiment analysis systems are deployed? In what kinds of industrial contexts are these systems deployed to begin with? What are the downstream tasks to which sentiment outputs are fed, or what decisions are made using the sentiment outputs? Moreover, since this is an analysis focused on names, what is the effect on these systems on the representations of these names in text?

In contrast, in a different analysis Bhaskaran and Bhallamudi [2019] focus explicitly on *occupational gender stereotyping* in sentiment analysis systems. To do so, they create a dataset of sentences using the template ⟨*Noun*⟩ *is a/an* ⟨*profession*⟩, where ⟨*Noun*⟩ is filled in with gendered noun phrases such as *This girl/This boy*. Gender stereotyping is measured via "differences in mean positive class probability between sentences with male and female nouns for each profession."

8.5.3 Hate speech and toxicity detection

The related tasks of abusive language, hate speech, and toxicity detection have attracted increasing interest in the last few years, perhaps due to growing burdens of online content moderation. The exact task definition varies; for example, Dixon et al. [2018] define toxicity detection as the task of determining whether a comment is a "rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion." Problem formulations can be binary (e.g., *hateful* and *not hateful* labels) or multiclass (e.g., *spam*, *abusive*, *hateful*, *none* labels [Park et al., 2018]). We examine recent work proposing a range of different metrics for assessing bias in systems trained to perform these tasks.

Dixon et al. [2018] define *unintended bias* for a toxicity detection system as follows: "[A] model contains unintended bias if it performs better for comments containing some particular identity terms than for comments containing others." This is designed to capture the observation that some identity terms, such as *gay*, occur disproportionately often in comments classified as toxic relative to their overall frequency across comments. To evaluate unintended bias, Dixon et al. generate a synthetic dataset by filling in templates with a range of identity terms and various toxic and non-toxic phrases. Let $t \in T$ represent an identity term, FPR_t the false positive rate on the subset of the data containing t, and so forth. Then unintended bias is then evaluated using the following four metrics.

- False positive equality difference (FPED): $\sum_{t \in T} |\text{FPR} - \text{FPR}_t|$
- False negative equality difference (FNED): $\sum_{t \in T} |\text{FNR} - \text{FNR}_t|$
- Pinned AUC
- Pinned AUC equality difference

The first two of these are straightforward. The third, the pinned AUC, is designed to capture the difference between model performance on examples from one subgroup relative to examples from the overall distribution; formally, the procedure samples examples from one subgroup, samples examples from the overall distribution, and computes the AUC[13] on the union of the two samples. The final metric, the pinned AUC equality difference, computes the difference between the overall AUC and the pinned AUC calculated for each identity term.

Several other papers take a similar approach; Park et al. [2018] also generate a synthetic dataset by filling in templates with a range of identity terms and offensive/non-offensive words, evaluating models according to false positive equality difference and false negative equality difference.

[13] The AUC, or area under the ROC curve, is a performance metric for a binary classifier, computed by plotting the true positive rate against the false positive rate at various threshold settings and computing the area under the resulting curve. Equivalently, it is the probability that any pair of items is correctly ranked by the classifier's scores.

Badjatiya et al. [2019] also evaluate models using pinned AUC on a synthetic dataset. In addition, they propose a suite of *pinned bias* metrics based on bias sensitive words, which are words with respect to which "the classifier is unreasonably biased...to a very high degree"—that is, their presence in a piece of a text significantly raises the prediction probability of a negative class label. Given such a set T of words w, the pinned bias is defined as

$$PB = \sum_{w \in T} \frac{|p(\text{ hateful } \mid w) - \phi|}{|T|}, \tag{8.11}$$

where T is the full set of bias sensitive words, $p(\text{ hateful } \mid w)$ is the probability of the *hateful* class label given a sentence with only the word w, and ϕ is a parameter that is changed for different versions of the pinned bias metric.

With the exception of pinned bias, which operationalizes a notion of undesirable correlations, we can identify these approaches as operationalizing quality of service harms, as each identifies one or several performance metrics of concern and measures the differences in those performance metrics for inputs with different identity tokens.

In contrast, Garg et al. [2019] propose a counterfactual analysis under the view that fairness "requires equal model behavior on individual counterfactual pairs," where pairs are obtained by substituting unigram and bigram tokens associated with identity groups. To evaluate a toxicity classifier f, they define a counterfactual token fairness gap over a single example x:

$$\text{CTF GAP}_{\Phi}(x) = \frac{1}{|\Phi(x)|} \sum_{x' \in \Phi(x)} |f(x) - f(x')| \tag{8.12}$$

where Φ is a generating function for counterfactual pairs, and $f(x)$ is the toxicity score of the classifier f on example x. The gap over an entire dataset can therefore be computed as the average of CTF GAP over all examples with valid counterfactuals.[14] Because this metric is concerned with the difference in predicted toxicity scores we can identify it as operationalizing a version of undesirable correlations.[15]

---

[14]These are the subset of counterfactuals whose toxicity is assumed to be symmetric—that is, substituting identity tokens does not yield a comment that is substantially more or less toxic.

[15]Garg et al. also examine gaps in true positive and true negative rates, a quality of service concern, observing a trade-off between CTF GAP and true positive rates.

Yet another line of work on toxicity and hate speech detection examines not the identity labels that may be in the text, but the variety—specifically, African American Language—in which the text is written. Sap et al. [2019] measure racial bias in toxicity detection in two ways. The first employs the demographic mixed membership model of Ch. 2, which returns the proportion of tokens in the input text that likely come from the model's African American topic. Sap et al. [2019] measure the correlation between this probability (p_{AAE}) and toxicity labels of two large hate speech detection datasets, finding that being labeled as *offensive*, *hateful*, or *abusive* is positively correlated with p_{AAE}. Second, they find that models trained on these corpora exhibit significant gaps in false positive rates across groups; examples with high p_{AAE} are more likely to be incorrectly labeled as offensive, abusive, or hateful than examples with low p_{AAE}.

Similarly, Davidson et al. [2019] find that high p_{AAE} tweets are more likely to be classified as hate speech by classifiers trained on a variety of datasets encompassing a range of taxonomies. Together, Sap et al. [2019] and Davidson et al. [2019] measure not only undesirable correlations but quality of service harms, as AAL-like tweets are more likely to suffer false positive classifications.

We note that p_{AAE} should not be understood as the probability of a tweet's being in AAL; rather, it is the proportion of tokens predicted as having been generated by the mixed membership model's African American topic, As such, the measurement may lack content validity in that the examples identified with high p_{AAE} may not actually exhibit features of AAL or be written by African Americans; however, Sap et al. show a similar effect on a dataset of tweets written by users with self-identified race, lending some convergent validity.

8.6 Discussion and recommendations

We have examined a variety of approaches to measuring bias in embedding spaces—both in NLP and in the quantitative social sciences—and in NLP systems more generally. What does our analysis mean for NLP practitioners?

Measurement modeling gives us a framework for thinking rigorously about the metrics we devise. We have demonstrated the process of measurement modeling—a framework that enables us to rigorously theorize both the constructs we should be measuring, and evaluate our models for measuring them. This process allows us to correctly situate existing approaches against each other and evaluate their merits and limitations. As we have seen repeatedly, different efforts to quantify bias for the same task may approach the measurement with radically different assumptions about what constitutes bias or unfairness, or what kinds of language give rise to bias that is measurable or normatively concerning.[16] At the same time, we find that across different tasks the scope of what constitutes bias is often profoundly limited, often to stereotyping. Concretely, we propose that additional kinds of harms could be fruitfully measured by examining these systems in their deployment context, or by examining the perceptions of produced text or labels by humans.

Analysis of harm arising from embeddings may not make sense. Though analyzing biases in embeddings is popular, we argue that it is an indirect path to measuring the harms arising from the downstream systems in which embeddings are used.

Measurement modeling reframes bias mitigation. Bias mitigation only makes sense with respect to a specific normative concern, and a particular way of operationalizing that concern. For instance, Bolukbasi et al.'s [2016] approach for debiasing word embeddings only reduces bias in the context of gender stereotyping, and only with respect to that particular measurement (with which other measurements might disagree). No approach, therefore, can be said to "debias" in general, and any proposed approach must be carefully contextualized with the harm it responds to, and how that harm is operationalized.

[16] It is, of course, a point of concern that because approaches rarely articulate their normative reasoning, it is often unclear which operationalization choices are made because of what seems measurable versus what is normatively concerning.

CHAPTER 9

STYLE AND BIAS

9.1 Introduction

As the areas of style and attribute transfer in text, as well as bias and ethics in NLP, have come to attract increasing interest, work at their intersection that examines bias through the lens of style is also beginning to emerge. In this chapter, we will examine one thread of this work, which uses automated classifiers to examine the stylistic characteristics or social attributes of text generated by NLP systems [Hovy et al., 2020].

We examine this approach from several perspectives, raising a number of normative questions about the efforts to use style as a lens through which to conceptualize bias and social meaning. We also draw on sociolinguistic perspectives on language, identity, and social meaning to interrogate the assumptions made in existing work's conceptualizations of style and attribute bias transfer, and bias through the lens of style. We argue that this work implicitly raises a number of significant normative questions about how to responsibly generate text, and drawing on work in sociolinguistics and linguistic anthropology as well as recent position papers in NLP, we further explore what it might mean for NLP systems to acquire and produce social meaning, and consider how evaluations of systems that generate text need to be reformulated to address social aspects of generated language. We emphasize that the goal of this chapter is not to criticize particular papers; rather, we aim to critically engage with the assumptions threaded through emerging work on style and bias.

9.2 Background

We first provide brief background on style and attribute transfer in NLP, and describe the work on style and bias in which we are interested.

9.2.1 Style and attribute transfer

Although definitions vary across existing work, style transfer is often defined as follows: given a piece of input text, it is "the task of rephrasing the text to contain specific stylistic properties without changing the intent or affect within the content" [Prabhumoye et al., 2018]. As we will see, such work (implicitly or explicitly) relies on particular conceptualizations of style as separate from other kinds of meaning ("intent or affect"); we will examine this assumption further. In more recent work, these efforts are conceptualized as "attribute transfer," which moves away from relying on "style" as a concept and towards "transforming a sentence to alter a specific attribute (e.g. sentiment) while preserving its attribute-independent content" [Li et al., 2018]. Such attributes may be quite diverse, for instance Subramanian et al.'s [2019] use of "gender, sentiment, [and] product type," each of which (when it is the attribute being transferred) may be treated as separate from the content that should be preserved. Style or attribute is often marked by individual words; for example, in Li et al. [2018] attribute markers are identified by finding words disproportionately occurring with one attribute relative to the other, and removed; the remainder of the sentence is the attribute-independent content.

9.2.2 Bias through the lens of style

Hovy et al. [2020] quantify "stylistic bias" in machine translation by examining the predicted social attributes of the input and output text of machine translation systems. Specifically, for five languages, text translated by commercial machine translation systems into English is generally predicted by age and gender classifiers to be "older" and more "male" than the input text. Hovy et al. conceptualize this is a stylistic issue, suggesting that this issue raises opportunities for "machine translation to take stylistic considerations into account."

9.3 Sociolinguistic conceptions of style and social meaning

How is style conceptualized in sociolinguistics? Sociolinguistic research has traditionally distinguished language variation arising from social factors—that is, variation between speakers (inter-speaker variation)—from stylistic variation arising from social context (intra-speaker variation) [Moore, 2004, Bucholtz and Hall, 2005]. That is, stylistic variation has traditionally been conceptualized as shifts in a speaker's language use in different social contexts, as opposed to differences between speakers correlated to social categories such as race, gender, and class; as Moore [2004] puts it, "[D]ialects are considered to be varieties according to users and styles are considered to be varieties according to use." According to Moore, who traces shifts in research perspectives on style, variation in speakers' styles was conceptualized as arising from the attention users pay to their speech, according to an interaction's level of formality.

In more recent conceptualizations of variation, however, the use of this distinction has diminished. In part, this is due to the inherent difficulty of separating them, as Wolfram and Schilling [2015] suggest:

> Because there is no clear dividing line between register/genre shifting and dialect shifting, or between dialect shifting and code-switching, and because people rarely 'turn off' one dialect or register and 'turn on' another, it is perhaps more fruitful to think of stylistic variation—as with social and ethnic group-based variation—in terms of stylistic repertoires rather than register or dialect *per se*. Again repertoire can be taken to refer to the collection of linguistic features that each individual has at his or her disposal at any given moment, to be employed as needed for different social, interactional, and personal reasons, rather than conceiving of people shifting into and out of abstract entities like 'Latino English' or a 'legalese' register, which they either do or do not 'own.'

Another reason is because the distinction provides little analytical value. Rather than asking whether a speaker is engaging in dialect shifting or style shifting, it is more fruitful to ask (as Wolfram and Schilling explain) what interactional goals or identity construction a speaker is aiming to accomplish, and what linguistic resources they are drawing on to do so [Moore, 2004, Bucholtz and Hall, 2005, Eckert, 2008]. This reframing is particularly important because the "categories and contexts" that are traditionally used to analyze dialectal or stylistic variation, such as the macro social categories of race and gender or differing levels of formality, are so abstract that they cannot possibly explain all of the observed variation; by reframing language use as processes of identity construction rather than dialect or style shifting, researchers can uncover the local, interactional social meanings at play [Moore, 2004]. For instance, Bucholtz [1999] shows that the social identity "nerd" is highly salient in a U.S. high school, and speakers' desire to align themselves (or distance themselves from) this social identity helps to drive language use.

From this perspective, therefore, "style" no longer describes one particular type of variation; rather, "this approach to sociolinguistic analysis views everything as stylistic" [Moore, 2004]. One consequence of this view is that just as style is no longer readily distinguishable from dialect variation, style is also not readily distinguishable from other aspects of meaning. Eckert [2008] argues this as follows: "Sociolinguists generally think of styles as different ways of saying the same thing. In every field that studies style seriously, however, this is not so — style is not a surface manifestation, but originates in content. The view of style I present here precludes the separation of form from content, for the social is eminently about the content of people's lives. Different ways of saying things are intended to signal different ways of being, which includes different potential things to say." That is, because variation arises from the process of identity construction, in which speakers index social group membership, align themselves with particular stances and beliefs, and so forth, seemingly semantically equivalent things are not equivalent after all, because the differences between these apparently equivalent things are socially meaningful.

9.4 Style, identity, and social meaning in NLP

Having briefly examined sociolinguistic conceptualizations of style, identity, and social meaning, we return to NLP to uncover and trouble style transfer's assumptions about style and social attributes.

One key assumption throughout work on style and attribute transfer is the kinds of styles or attributes that are conceptualized as transferable, and therefore as separable from an utterance's semantic meaning or "attribute-independent content" [Li et al., 2018]; across recent work this includes gender, language variety ("AAVE" or "SAE"), political affiliation, and sentiment [Li et al., 2018, Prabhumoye et al., 2018, Subramanian et al., 2019, Rios, 2020]. This set of attributes is incompatible with either sociolinguistic conceptualization of style. Under the traditional view, in which style captures intra-speaker variation, none of these attributes can be considered as such; gender and political affiliation (if referring to a speaker's background beliefs or political preferences) do not shift between a speaker's social contexts (and hence there is no such thing as "male" or "female" style), and sentiment and political affiliation (if referring to a speaker's expressed beliefs or preferences) would usually be considered part of the semantic meaning of an utterance, rather than its style.

Under the more recent view, the process of identity construction does engage these group memberships, stances, and beliefs, but, as Eckert describes, the linguistic resources used to index these associations are inseparable from the content or semantic meaning of the utterance; speakers draw on these resources to simultaneously express semantic meaning and make "ideological moves" [Eckert, 2008]. Because any linguistic variable can index a number of potential meanings—an "indexical field" [Eckert, 2008]—and because these fields are always shifting, it is impossible to say with certainty what attributes, group memberships, stances, or beliefs are directly responsible for any given linguistic realization. Thus, any computational attempt to determine what parts of an utterance were produced by "gender," "political affiliation," or other attribute run counter to the sociolinguistic model of how language is produced, as do computational approaches that identify the words most distinctive of two different corpora and subsequently assume that those words are associated with the attribute of interest.

9.5 Style and bias

Following our discussion of how style is conceptualized in style transfer, we examine Hovy et al. [2020], which raises a normative concern about what they call "translation bias." Specifically, as we described above, this work is concerned that translations from German, French, Italian, and Dutch into English by commercial machine translation systems are classified by automated classifiers as more "male" and older" than the input text. This, in their work, is conceptualized as "stylistic considerations," separate from the "what is being said", and Hovy et al. (against the view of style as intra-speaker variation) view these stylistic aspects as connected to "[d]emographic factors (age, gender, etc.) [that] manifest in language and therefore influence style." Most importantly, they view this is an issue of *perception*: "we do not expect a 6-year old to sound like an adult, and would not translate a person to sound differently gendered. However, in this paper we show that that is essentially what happens in machine translation: authors sound on average older and more male." Thus, at the heart of their concern is not style per se (at least not how style is conceptualized sociolinguistically), but rather the social meanings connected to social group memberships inferred by the *listening subject*. From this perspective, a machine translation system treats speakers unfairly if the output translation is interpreted as associated with different social groups than would have been perceived by a listener perceiving the input text.

In Ch. 6 we pointed to literature in sociolinguistics and linguistic anthropology on listeners' perceptions (though that literature is much more limited for text), and we suggest that it is reasonable to focus on how automatically generated text is perceived, rather than any "ideological moves" performed by the system automatically generating text, considering that a machine translation system is not a speaking subject in any usual sense. MT systems are disembodied, are not really aware of their discourse contexts or interlocutors, and do not engage in any process of indexing social group memberships, stances, or beliefs. If we view social meaning as that created in the process of identity construction, then it is hard to say that MT systems are producing any. But that does not preclude the text they generate from carrying social meaning; as Bender and Koller [2020] point out, humans are remarkably persistent in making meaning out of text they know is automatically generated, and there is no reason this persistence would not carry over to social meaning.

We argue that the concept of "indexical inversion," introduced by Inoue [2006] and explained by Rosa and Flores [2017], can help us here: "Rather than the common use of indexicality to understand how linguistic signs index social categories, indexical inversion considers how language ideologies associated with social categories produce the perception of linguistic signs." As Rosa and Flores illustrate (which we discussed briefly in Ch. 6 and 7), this process can be seen in white listening subjects' perception—driven by raciolinguistic ideologies constructing racialized people as linguistically deficient—of racialized speaking subjects as producing non-normative language, no matter how closely such speakers adhere to dominant language practices. What this framing may offer us is a way to reason about the social meanings of the text produced by an automated system; rather than being concerned about the social categories such a system cannot possibly be indexing, perhaps we should attend to the processes governing how humans interpret the produced text.

We suggest that this framing opens up a wide space of questions about human perception of text produced by automated systems. For instance, do humans perceive the outputs generated by Hovy et al. as "older" and more "male" than the inputs, as the automated classifiers do? What gives rise to these judgments? If, as Cave and Dihal [2020] suggest, chatbots and digital assistants are frequently racialized as white, how does that affect perceptions of grammaticality, "standardness," empathy, or trustworthiness of the text they produce? If they produce language features that are frequently racialized as non-white, how is this viewed by non-white and white speakers—perhaps as appropriative [Eberhardt and Freeman, 2015]? Perhaps as authentic [Cutler, 1999]? Or perhaps as jarring or entering the uncanny valley, since such language is associated with membership in particular social categories, to which an automated system cannot belong (in contrast to language associated with normative whiteness, which is unmarked)? Chatbots and digital assistants are frequently gendered; how does this gendering shape perceptions? Although we have a substantial literature on language ideologies to draw on, we know little about people's cultural ideas about language produced by automated systems. What ideologies of language, therefore, shape perceptions of automatically produced text? What ideologies do automated systems reproduce?[1]

How does this perspective reframe Hovy et al.'s [2020] analysis? Our discussion on indexical inversion focuses on how ideologies held by humans about language and automated systems shape the perception of text. Hovy et al.'s analysis, however, is performed via classifiers trained to predict age and gender. Such prediction of attributes is common in style transfer; for instance, Prabhumoye et al. [2018] in fact motivate transfer as a way to obfuscate the social attributes of text authors from automated classifiers.[2] But automated

[1]In their analysis of what users said about Microsoft's Tay, Neff and Nagy [2016] propose a related framing in terms of user perceptions of language technologies' affordances: "Rather than think of technologies as having fixed capacities that are recognized by their human partners, *imagined affordances* allow us to describe users' perceptions, attitudes, and expectations; the materiality and functionality of technologies; and the intentions and perceptions of designers."

[2]Of course, as we have seen in our discussion of style above, social attributes cannot be recovered deterministically from text, but whether these attributes can be correctly recovered is, in our

classifiers, though important for highlighting a significant issue of normative concern, cannot tell us what humans make of generated text.

What ought we to do about the concern raised by Hovy et al.? It is an important one; it seems natural that speakers would not want their translated text to give rise to different associations than they intended, particularly (as is likely if an MT system is being used) if they have no way to identify such a difference. But no straightforward solution presents itself. To attempt to preserve the associations of input text seems impossible, particularly as such associations are wide-ranging, contextual, fluid, and shifting; given the issues we identified above with style transfer's conceptualizations of style and social meaning, we do not (and may never have) ways to effectively choose and operationalize the elements of style and social meaning that ought to be maintained across translations, particularly because much of these aspects may be particular to the language variety of the input text and the social contexts of its speakers. Not to address Hovy et al.'s concern, however, is to risk flattening all output text to language resembling the model's training data, potentially reinforcing the status of dominant language practices, as Hovy et al. observed with output text classified as "older" and more "male." Here, theories of translation for human translators, who have long grappled with related questions, may offer us roadmaps [Newmark, 1981, Proshina, 2008].

We argue that the questions we raise here apply to all text generated by automated systems, not just machine translation systems. From our perspective, though Hovy et al. do not say so, a contribution of their work lies in quietly dismantling the idea that there is a "voice from nowhere" [Gal, 2006, Woolard, 2008, Chun and Lo, 2015]—that is, that there is any generated text that is not interpreted by people as carrying social meaning. In a way, the fact that classifiers of age and gender are used to provide predictions for the input and the output alike do important work to contradict this idea. This reminder that much of the text forming NLP data has gone historically unmarked, due to ideological processes privileging certain language practices, is essential. From this perspective, an important piece

view, secondary to the fact that automated approaches are used to predict them anyway, e.g., for advertising purposes.

of future research in NLP will be to "unmark" this text by evaluating the social perceptions of any text generated by automated systems.

CHAPTER 10

CONCLUSION

This thesis has been driven by two questions: First, how can we conceptualize harms arising from NLP systems? Second, how can we quantify and mitigate such harms? Towards these questions, we have proposed a distantly supervised mixed-membership model for gathering the first dataset of AAL-like social media language (Ch. 2); identified performance disparities of language identification and dependency parsing systems between AAL-like and MUSE-like language (Chs. 3 and 4); conducted a critical survey of 146 existing papers on "bias" in NLP (Ch. 5); drawn together literature from a range of language-related disciplines to propose a taxonomy of *representational* harms arising from NLP systems and practices (Chs. 6 and 7); applied the measurement modeling framework from the quantitative social sciences towards rigorously evaluating approaches for quantifying bias (Ch. 8); and provided a preliminary analysis of emerging efforts at the intersection of style and bias (Ch. 9).

The analyses and frameworks proposed in this thesis invite a number of open questions and directions, some of which we proposed at the end of Ch. 7 as foundational guiding questions for work on bias in NLP. In the following sections, we propose some additional exciting and important potential directions towards more equitable and just NLP.

10.1 Measuring harms: NLP systems in their sociotechnical context

Allocational harms As we saw in Ch. 5, very little work on bias in NLP has concretely identified and measured allocational harms arising from NLP systems. Nevertheless, because language is involved in many decision-making processes that allocate opportunities or resources, and because linguistic discrimination produces unjust outcomes in so many of these processes [Craft et al., 2020], it is critical to uncover whether and how NLP systems participate in these processes. Existing evidence of such participation is scant; we are familiar only with emerging analyses of systems that match job descriptions with resumes [Deshpande et al., 2020], Amazon's "internal AI recruiting tool" that reportedly disadvantaged women [Vincent, 2018], and the U.S. Department of Homeland Security's expanding use of social media monitoring [Duarte et al., 2017, Llansó et al., 2018, Patel et al., 2019], including a proposal to develop "extreme vetting" systems that would "analyze social media posts to predict whether individuals will become 'positively contributing member[s] of society' and whether a person 'intends to commit criminal or terrorist acts after entering the United States"' [Duarte et al., 2017]. We speculate that NLP systems may come to participate—if they do not do so already—in decisions related to educational placements and outcomes [Loukina et al., 2019]; hiring, either in resume screening or automated interviewing, as emotion recognition already does [Stark and Hoey, 2020]; immigration and citizenship, particularly as language competency is already commonly evaluated as a component of citizenship tests;[1] and medical screening or diagnosis, for example with voice analysis systems developed for psychiatric screening [Semel, 2020] and COVID-19 screening [Anthes, 2020].

Representational harms In Ch. 7, we explored a number of representational harms that are under-explored in current work on bias in NLP, and we view the development of approaches for measuring these harms as an essential future direction. For some of these, measurement is challenging because of the complexity of ideas and social context involved; for example, the examples in Table A.4 illustrate that large language models reproduce constellations of harmful ideas about different social groups, which are difficult to identify and mitigate automatically as the ideas cannot be localized to particular words, unlike existing approaches for measuring stereotyping. Here, efforts to develop measurements may benefit from work on identifying microaggressions [Breitfeller et al., 2019, Jurgens et al., 2019, Chiril et al., 2020, Sap et al., 2020], patronizing or condescending language [Pérez-Almendros et al., 2020], or patterns of power dynamics [Rowe et al., 2007, Prabhakaran et al., 2014] in text. For other harms, measurement is challenging because they involve

[1]For example, an English test in the U.S. (`https://www.uscis.gov/citizenship/learn-about-citizenship/the-naturalization-interview-and-test`) and proof of knowledge of English in the U.K. (`https://www.gov.uk/english-language`)

analyses of stakeholders beyond the designers and immediate users of NLP systems; for example, what is the collective impact of the disproportionate removal of social media posts with AAL features or about disability?

10.2 Challenges and tensions in measurement

Here, we examine some tensions that arise in measuring harms that concern different language varieties and their speakers; we describe them in the context of AAL, but note that some of these challenges and tensions are likely to arise for a range of varieties. We will use AAL to refer to the language variety and "AAL" to refer to the term itself.

Bounding language varieties Many approaches for measuring harms, particularly quality of service harms, involve the measurement of performance disparities of systems between different language varieties. Such measurements, however, first require practitioners to determine the boundaries of each language variety under study.

This raises difficult questions for practitioners. Under Lanehart et al.'s [2015] expansive definition of AAL, we might look for language produced by people self-identifying as African American, which poses both a data availability challenge—as NLP datasets are typically not accompanied by speakers' self-identified race—and a privacy challenge, as the burdens of intrusive data collection and surveillance have historically fallen disproportionately on minoritized communities [Browne, 2015].

An alternative approach to identifying AAL to use the linguistic features identified as core to AAL [Rickford, 1999, Green, 2002]. But these features include all aspects of the linguistic system, including the lexicon, phonology, morphology, and syntax, and there are no clear criteria for distinguishing AAL from other varieties—which features must be observed to "count" as AAL? What about an utterance with just a phonological feature, or with only AAL-like prosody?[2] These challenges are compounded in "code-switching"[3] situations and in text, where prosodic or phonological features present in speech may not be marked in text, where apparent features may or may not be indicative of AAL at all, and where utterances may be separated from their context. Moreover, many of the features identified as core to AAL emerge from early sociolinguistic work focused on a relatively small subset of African American speakers—largely young, urban, working-class, and male—at the expense of women, older, rural, and LGBTQ+ speakers [King, 2020]. As a result, there are no reliable or straightforward criteria for drawing the boundaries between AAL and other varieties.

Because of the general impossibility of determining what "counts" as belonging to a given language variety, some researchers in sociolinguistics and linguistic anthropology are moving away from conceptualizing language varieties, including AAL, as fixed, bounded objects and towards conceptualizing fluid sets of language practices or resources that speakers draw upon, both as a means of analyzing speakers' practices [García and Wei, 2014], and of critically examining how languages have come to be conceptualized as "fixed entities capable of being counted, systematized, and named" [Severo and Makoni, 2020] in the first place [García and Alvis, 2019, García, 2019]. Despite the advantages offered by this lens for sociolinguistic analysis, however, it does not appear to lend itself to a solution for measuring performance disparities between language varieties in NLP.

Racist implications of "AAL" A second, significant concern that arises from emerging efforts for inclusive technologies centers on a related issue. Here, the very fact of these boundaries—that "AAL" is used to identify a set of features or utterances, thereby marking entire communities' language practices as non-normative, non-standard, or simply different from MUSE—may be viewed negatively by community members as encoding deeply racist assumptions about African American language practices and communities. From this perspective, because ethnoracial identity does not map onto language practices, it is racist to

[2]We also note that these questions are framed from a perspective in which Mainstream U.S. English is the default and unmarked, and utterances are considered to be in MUSE unless distinctive features from another, marked variety are observed.

[3]See recent translanguaging approaches [García and Wei, 2014] for more on critical approaches to code-switching.

suppose that African Americans must all produce non-normative or non-standard language (that, under the "AAL" designation, is even separated from English), and marking out language produced by African American speakers as necessarily different reproduces racist ideas about the language practices of African American people as being "not English," as well as the idea that there is a monolithic Black or African American "community."

As we described previously, the move towards "AAL" (from older terminology and indeed from "AAE/AAVE") encodes a commitment towards recognizing the language practices of African American speakers as non-monolithic; "AAL" is sociolinguistically constructed to encompass different language practices as broadly as possible. Nevertheless, it unavoidably carves out an ethnoracial group as its starting point [King, 2020] and designates the language practices of entire African American communities as a single object/entity of study; in contrast, there is no such designation for "white English" or "white American language."

This issue illustrates the broad challenges associated with drawing boundaries around language; because beliefs about language and speakers vary widely, and because language and other social categories (including race) are co-naturalized [Rosa and Flores, 2017], the problem of determining who and what counts—even for the laudable goal of developing more equitable or just NLP systems—is deeply fraught.

Generating synthetic data Some work on bias in NLP has turned to the generation of synthetic language data in different varieties or styles in order to measure the performance of NLP systems; such work includes Rios's [2020] generation of "Simulated African American Vernacular English" and our own generation of text containing AAL syntactic features in Ch. 4. But the generation of language associated with non-standard, minoritized varieties, even for the purposes of carefully characterizing systems' performance, is potentially fraught, and we raise a few questions here. For one, the practical question of "what counts" that we raised above remains; generation efforts must decide what "counts" as appropriate or high-quality synthetic data. For another, language generated by automated processes may bear little resemblance to language that is produced by speakers (and likely to be encountered in real deployment contexts), particularly if generation is done using complex, opaque processes (e.g., style transfer models) instead of more controlled linguistic features. Yet another concern is that generating language (rather than collecting it from real speakers) risks creating fairness or bias measurement processes that never require interacting with speakers and communities, potentially producing research approaches and communities whose priorities diverge from those of minoritized speakers, and where minoritized speakers do not make technical decisions. As a result, as research on fairness and bias becomes increasingly valorized (as it seems to have done over the last few years), the generation of synthetic language data is one way that researchers and practitioners who are not members of minoritized communities might end up benefiting from working with minoritized language varieties, without working with speakers and communities themselves, raising echoes of issues of linguistic appropriation we discussed in Ch. 7.[4]

10.3 Awareness, recourse, participation, and refusal

In Ch. 7 we proposed that work on harms from NLP systems complement algorithmic measurements of harm by examining the lived experiences of stakeholders with NLP systems, and in Ch. 9 we suggested that examining users' social perceptions of automatically generated text as one such potential direction. Here, we suggest a few more stakeholder-centered directions: understanding their experiences, offering recourse, developing participatory approaches, and empowering resistance and refusal.

Everyday encounters and awareness What NLP systems do people encounter every day, and how do they become aware of them, if at all? Existing work has attempted to address both of these questions—for example, by documenting a hypothetical family's interactions with surveillance systems, knowingly or unknowingly, for a week [Ball et al., 2006], and by investigating users' folk theories surrounding social media platforms' operation [Eslami et al., 2016, DeVito et al., 2017, 2018]—but little work has focused specifically on NLP systems. Because many NLP systems operate as parts of larger pipelines, it may not

[4]One potential response to these concerns that researchers generate language associated with normative whiteness all the time without issue, as it is unmarked. We note, however, that these cases are not directly analogous precisely because white speakers are not minoritized in NLP research (or indeed anywhere else).

always be evident whether or how an NLP system has been involved in a decision—for example, which ads are shown or how product reviews are summarized (perhaps influenced by the content of a user's social media text), or whether a content moderation system is operating over a user's social media text. We suggest that in order to fully understand the implications of NLP systems in their sociotechnical contexts, research must first identify the landscape of these systems in the first place—where and how they operate—and how users interact with and develop perceptions of them.

Recourse When NLP systems get it wrong—and as we have seen in this thesis, they do so frequently—what recourse do users have? *Algorithmic recourse*, "the systematic process of reversing unfavorable decisions by algorithms and bureaucracies" [Venkatasubramanian and Alfano, 2020], has received considerable attention in the algorithmic fairness community in order to develop systems whose decisions users can understand and, ideally, contest [Karimi et al., 2020]. Although interpretability has received attention in NLP research (e.g., Wallace et al. [2020]), much work has focused on aiding technologists' understanding of system behaviors, rather than end users' understanding of or ability to change system behaviors. One arguable example of the latter is Google's response to the finding that translations from languages without gendered pronouns into those with gendered pronouns resulted in gender-stereotyped translations [Prates et al., 2019, i.a.]. In response, for some language pairs, Google began providing multiple gender-specific output translations—for example, given the English input *My friend is a doctor* it provides both *Mi amiga es doctora* and *Mi amigo es doctor* in Spanish—along with a brief explanation ("Translations are gender specific") and a link to a help article [Johnson, 2020]. Such a solution makes the inherent ambiguity of the output visible to the user and provides them with a choice of output, rather than making a decision behind the scenes, as was done previously.

Participatory approaches How can we expand opportunities for participation in developing NLP systems beyond annotation, particularly for minoritized speakers? We are inspired by recent efforts, such as Nekoto et al.'s [2020] participatory approach to developing machine translation datasets and benchmarks for over 30 African languages, which are disproportionately low-resourced. Nekoto et al. offer strategies for dismantling the many obstacles to participation, which include the tertiary education pre-requisites typically placed upon potential researchers and the availability of compute resources.

Refusal Drawing on a long line of Black, Indigenous, queer, and feminist thought, recent work on automated systems is beginning to ask about how people—both technologists and the communities on whom systems are deployed—might *refuse* them [Cifor et al., 2019, Gangadharan, 2020, Barabas, 2020]. Such thinking acknowledges that even efforts towards developing more equitable systems risk legitimizing both the systems themselves and the research practices and social relations underpinning their development. Refusal—to concede these practices and relations, to build within existing logics (e.g., carceral logics [Benjamin, 2019])—can therefore open up the possibility of dismantling these practices, relations, and logics as well as that of developing alternative ones. We propose that NLP research can draw on this line of thinking in two ways. First, we suggest that the field would benefit from dialogues on what and when not to build, to encourage a critical technical practice [Agre, 1997] among researchers and practitioners. Second, we propose research that examines how different stakeholders might refuse NLP systems, either individually or collectively, which requires understanding first how people come into contact and interact with NLP systems.

APPENDIX

A.1 Identifying AAL from demographics

These four "races" that form our mixed membership model—non-Hispanic whites, Hispanics, non-Hispanic African Americans, and Asians—are commonly used in sociological studies of the U.S. The Census tracks other categories as well, such as Native Americans. The exact options the Census uses are somewhat complicated (e.g., Hispanic is not a "race" but a separate variable); in a small minority of cases, these values do not sum to one, so we re-normalize them for analysis and discard the small fraction of cases where their sum is less than 0.5. For simplicity, we sometimes refer to these four variables as races; this is a simplification since the Census considers race and ethnicity to be separate variables, and the relationship between the actual concepts of race and ethnicity is fraught on many levels.

A.2 A dataset of AAL morphosyntactic analysis

Table A.1 provides the ARK POS tagset used in tagging tweets for our feature searches. Table A.2 provides the syntax patterns for each morphosyntactic feature we examine.

Tag	Definition
`N`	common noun
`O`	pronoun (personal/WH; not possessive)
`^`	proper noun
`S`	nominal + possessive
`Z`	proper noun + possessive
`V`	verb including copula, auxiliaries
`L`	nominal + verbal (e.g. *i'm*), verbal + nominal (*let's*)
`M`	proper noun + verbal
`A`	adjective
`R`	adverb
`!`	interjection
`D`	determiner
`P`	pre- or postposition, or subordinating conjunction
`&`	coordinating conjunction

`T`	verb particle
`X`	existential *there*, predeterminers
`Y`	`X` + verbal
`#`	hashtag
`@`	at-mention
`~`	discourse marker, indications of continuation across multiple tweets
`U`	URL or email address
`E`	emoticon
`$`	numeral
`,`	punctuation
`G`	other abbreviations, foreign words, possessive endings, symbols

Table A.1: ARK POS tagset [Owoputi et al., 2013].

Feature	**Pattern**
Habitual *be*	Token with POS in { *N, O,* $\wedge$, *@* } + *be/b* + token with POS in $\{V/A/N/D/P\}$ Disallow in (up to) 4 tokens preceding the *be/b* token: *can, is, are, could, let, will, would, can't, cant, won't, wont, aren't, arent, couldn't, couldnt, wouldn't, wouldnt, whether, should, shouldn't, shouldnt, must, mustn't, mustnt, god, may* Disallow following the *be/b* token: *valentine, careful, like, quiet, nice, friends, safe, happy, aware, damned, grateful* Disallow tweets with fewer than 4 tokens
Stressed *BIN*	Token matching { [*]+*bee*+*n*+[*]+, *beee*+*n*+, *bee*+*nn*+, *bii*+*n*+, *bi*+*nn*+, *BEEN, BIN* } If token is uppercase, require that preceding and following tokens not be uppercase Disallow in (up to) 2 tokens preceding the *BIN* token: *has, have, hasn't, hasnt, haven't, havent, i've, ive, you've, youve, he's, hes, she's, shes, they've, theyve, we've, weve, never, always, how, it's, its, that's, thats, there's, theres, hadn't, hadnt, shoudn't, shouldnt, wouldn't, wouldnt, couldn't, couldnt, where* Allow immediately preceding the *BIN* token only if token following *BEEN* has POS *V*: *should've, shouldve, shoulda, would've, wouldve, woulda, could've, couldve, coulda, of* Disallow tweets with fewer than 3 tokens
Resultant *done*	Token with POS in { *N, O,* $\wedge$, *@* } + token matching *do*n+e+*\b + *V* Disallow *is* after the *done* token
Gone	Token with POS in { *N, O,* $\wedge$, *@* } + token matching *go*n+e**\b + *V* Disallow *is* after the *gone* token

Finna	Token with POS in { *N, O,* $\wedge$, *@* } + token matching { *finna, fenna, funnah, finnah, fennah, funnah, finto* } + *V*
Steady	Token with POS in { *N, O,* $\wedge$, *@* } + *steady* + *V*
Multiple negation	Pattern 1: Auxiliary token + any token + negative token Pattern 2: Auxiliary token + negative token + *V* Auxiliaries: *can't, cant, couldn't, couldnt, won't, wont, don't, dont, didn't, didnt, shouldn't, shouldnt, wouldn't, wouldnt, doesn't, doesnt* (*ain't/aint* disallowed due to frequent occurrence of *ain't nobody got time for...*) Negative tokens: *no, nobody, nothing, nuffin, nuttin, nun, never, neva, nevah*
Negative inversion	Auxiliary token + negative token + *V* Auxiliaries: same as for multiple negation Negative tokens: *no, nobody, nothing, nuffin, nuttin, nun* Disallow token with POS in { *N, O,* $\wedge$, *@* } preceding auxiliary token
Non-inverted negative concord	Negative token + auxiliary token + *V* Negative tokens: *nobody, nothing, nuffin, nuttin, nun, no one* Auxiliaries: same as for multiple negation

Table A.2: Search patterns for each morphosyntactic feature.

A.3 Posterior inference for the ensemble classifier

The posterior inference task is to calculate the posterior expectation of

$$P(\theta \mid w, \phi, \alpha) \propto P(\theta \mid \alpha) P(w \mid \theta, \phi)$$

where ϕ are the trained topic-word language models and $\theta \sim Dir(\alpha)$ is a prior over topic proportions, with a fixed symmetric prior $\alpha_k = 1/16$.

The ϕ topic-word distributions are calculated via training-time posterior inference by averaging Gibbs samples $\bar{N}_{wk} = (1/S) \sum_s$ (where s indexes the last 50 samples of the Gibbs sampler), as well as adding a pseudocount of 1 and normalizing:

$$\phi_{k,w} \propto (\bar{N}_{k,w} + 1)$$

(The detailed balance theory of MCMC implies no pseudocount should be added, but we found it seemed to help since it prevents rare words from having overly low posterior expected counts.)

The $\hat{\theta}$ prediction is inferred as the posterior mean given the words in the message by using Collapsed Variational Bayes (CVB0) [Asuncion et al., 2009], which is closely related to both Gibbs sampling and EM. It iteratively updates the soft posterior for each token position $t = 1..T$,

$$q_t(k) \propto (N_{-t,k} + \alpha_k)\phi_{k,w_t}$$

where $N_{-t,k} = \sum_{t' \neq t} q_{t'}(k)$ is the soft topic count from other tokens in the message. The final posterior mean of θ is estimated as $\hat{\theta}_k = (1/T)\sum_t q_t(k)$. We find, similar to Asuncion et al. [2009], that CVB0 has the advantage of simplicity and rapid convergence; $\hat{\theta}$ converges to within absolute 0.001 of a fixed point within five iterations on test cases.

A.4 Preliminary parsing analysis: Stanford dependencies

The SyntaxNet model outputs grammatical relations based on Stanford dependencies 3.3.0; thus we sought to annotate messages with this formalism, as described in de Marneffe et al. [2013], a revision to de Marneffe and Manning [2008]. For each message, we parsed it and displayed the output in the Brat annotation software[1] alongside an unannotated copy of the message, which we added dependency edges to. This allowed us to see the proposed analysis to improve annotation speed and conformance with the grammatical standard. For difficult cases, we parsed shortened, Mainstream U.S. English toy sentences to confirm what relations were intended to be used to capture specific syntactic constructs. Sometimes this clearly contradicted the annotation standards (probably due to mismatch between the annotations it was trained on versus the version of the dependencies manual we viewed); we typically deferred to the parser's interpretation in such cases.

In order to save annotation effort for this evaluation, we took a partial annotation approach: for each message, we identified the root word of the first major sentence[2] in the message—typically the main verb—and annotated its immediate dependent edges. Thus for

[1] `http://brat.nlplab.org`

[2] We take Kong et al.'s [2014] view that a tweet consists of a sequence of one or more disconnected utterances. We sought to exclude minor utterances like "No" in "No. I do not see it" from annotation; in this case, we would prefer to annotate "see." A short utterance of all interjections was considered "minor"; a noun phrase or verb-headed sentence was considered "major."

every tweet, the gold standard included one or more labeled edges, all rooted in a single token. As opposed to completely annotating all words in a message, this allowed us to cover a broader set of messages, increasing statistical power from the perspective of sampling from a message population. It also alleviated the need to make fewer difficult annotation decisions—linguistic phenomena such as mis-tokenized fragments of emoticons, symbolic discourse markers, and (possibly multiword) hashtags.

We use the *twokenize* Twitter-specific tokenizer for the messages, which separates emoticons, symbols and URLs from the text [Owoputi et al., 2013][3] and use the space-separated tokenizations as input to SyntaxNet, allowing it to tokenize further. This substantially improves accuracy by correctly splitting contractions like "do n't" and "wan na" (following Penn and English Web Treebank conventions). However, as expected, it fails to split apostropheless forms like "dont" and more complicated multiword tokens like "ima" (*I am going to*, which Gimpel et al. [2011] sought to give a joint Pronoun-Verb grammatical category), typically leading to mis-analysis as nouns. It also erroneously splits apart emoticons and other multi-character symbolic expressions; fortunately, these are never the head of an utterance, so they do not need to be annotated under our partial annotation design. As described in Ch. 4, for our Universal Dependencies analysis we did not tokenize multiword tokens.

A.5 A survey of bias in NLP

Table A.3 presents the papers of which we are aware on bias in NLP systems for text; these include the main shared task paper for the Gendered Ambiguous Pronoun shared task at the Gender Bias in NLP Workshop [Webster et al., 2019] but not the individual submitted shared task papers. We do not include work that focuses primarily on identifying bias in text (e.g., Garg et al. [2018], Ananya et al. [2019], or Dinan et al. [2020]) rather than in NLP systems or practices. The table includes work not included in the analysis presented in Ch. 5, largely work emerging between July and November 2020. We note that a version of Cao and Daumé [2019] has been published as Cao and Daumé [2020]; we include the former here as it is an extended version of the paper.

[3]Using Myle Ott's implementation: `https://github.com/myleott/ark-twokenize-py`

Task	Papers
Embeddings	Bolukbasi et al. [2016,], Caliskan et al. [2017], McCurdy and Serbetçi [2017], Santana et al. [2018], Sutton et al. [2018], Zhang et al. [2018], Zhao et al. [2018], Agarwal et al. [2019], Basta et al. [2019], Brunet et al. [2019], Chaloner and Maldonado [2019], Dev and Phillips [2019], Dev et al. [2019], Ethayarajh et al. [2019], Font and Costa-jussà [2019], Gonen and Goldberg [2019], Hall Maudslay et al. [2019], James-Sorenson and Alvarez-Melis [2019], Kaneko and Bollegala [2019], Karve et al. [2019], Kurita et al. [2019], Lauscher and Glavaš [2019], Lauscher et al. [2019], Manzini et al. [2019], May et al. [2019], Mirzaev et al. [2019], Précenth [2019], Prost et al. [2019], Pujari et al. [2019], Qian et al. [2019], Sahlgren and Olsson [2019], Schramowski et al. [2019], Sedoc and Ungar [2019], Sweeney and Najafian [2019], Swinger et al. [2019], Tan and Celis [2019], Zhao et al. [2019], Zhou et al. [2019], Babaeianjelodar et al. [2020], Badilla et al. [2020], Bartl et al. [2020], Basta et al. [2020], Bhardwaj et al. [2020], Bommasani et al. [2020], Chen et al. [2020], Dev et al. [2020], Du and Joseph [2020], Fisher et al. [2020], Gonen and Webster [2020], Guo and Caliskan [2020], Gyamfi et al. [2020], Hube et al. [2020], Kumar et al. [2020], Kumar et al. [2020], Liang et al. [2020] Mulsa and Spanakis [2020], Nissim et al. [2020], Papakyriakopoulos et al. [2020], Popović et al. [2020], Ravfogel et al. [2020], Ross et al. [2020], Rozado [2020], Schlender and Spanakis [2020], Shin et al. [2020], Spliethöver and Wachsmuth [2020], Sweeney and Najafian [2020], Vargas and Cotterell [2020], Wang et al. [2020], Warmerdam et al. [2020], Yang and Feng [2020], Zhang et al. [2020], Zhao et al. [2020], Zhang et al. [2020], Zhao et al. [2020]
Language modeling, dialogue generation	Henderson et al. [2018], Lu et al. [2018], Curry and Rieser [2018], Bagdasaryan et al. [2019], Bordia and Bowman [2019], Florez [2019], Huang et al. [2019], Lee et al. [2019], Liu et al. [2019], Qian et al. [2019], Sheng et al. [2019], Solaiman et al. [2019], Zmigrod et al. [2019], Brown et al. [2020], Dinan et al. [2020], Gehman et al. [2020], Groenwold et al. [2020], Hendrycks et al. [2020], Krause et al. [2020], Lepori [2020], Li et al. [2020], Liu et al. [2020], McGuffie and Newhouse [2020], Monarch and Morrison [2020], Nadeem et al. [2020], Nangia et al. [2020], Peng et al. [2020], Sheng et al. [2020], Sheng et al. [2020], Shwartz et al. [2020], Soremekun et al. [2020], Strengers et al. [2020], Vig et al. [2020], Xu et al. [2020], Yeo and Chen [2020]

Tagging and parsing	This work [Blodgett et al., 2016, Blodgett and O'Connor, 2017, Blodgett et al., 2018], and Hovy and Søgaard [2015], Jørgensen et al. [2015], Garimella et al. [2019]
Coreference resolution	Lu et al. [2018], Rudinger et al. [2018], Webster et al. [2018], Zhao et al. [2018], Cao and Daumé [2019], Jumelet et al. [2019], Webster et al. [2019], Zhao et al. [2019], González et al. [2020], Kocijan et al. [2020], Soremekun et al. [2020], Wang et al. [2020], Webster et al. [2020], Yang and Feng [2020]
Machine translation	Alvarez-Melis and Jaakola [2017], Cho et al. [2019], Font and Costa-jussà [2019], Prates et al. [2019], Stanovsky et al. [2019], Basta et al. [2020], Farkas and Németh [2020], Gonen and Webster [2020], González et al. [2020], Hovy et al. [2020], Kocmi et al. [2020], Saunders et al. [2020], Saunders and Byrne [2020], Stafanovičs et al. [2020], Tan et al. [2020], Wong [2020]
Sentiment analysis	Díaz et al. [2018], Kiritchenko and Mohammad [2018], Shen et al. [2018], Thelwall [2018], Bhaskaran and Bhallamudi [2019], Huang et al. [2019], Prabhakaran et al. [2019], Sweeney and Najafian [2019], Zhiltsova et al. [2019], Bhardwaj et al. [2020], Groenwold et al. [2020], Hube et al. [2020], Hutchinson et al. [2020], Papakyriakopoulos et al. [2020], Popović et al. [2020], Sen and Ganguly [2020], Soremekun et al. [2020], Sweeney and Najafian [2020]
Hate speech, toxicity detection	Dixon et al. [2018], Park et al. [2018], Badjatiya et al. [2019], Davidson et al. [2019], Garg et al. [2019], Nozza et al. [2019], Prabhakaran et al. [2019], Raisi and Huang [2019], Sap et al. [2019], Vaidya et al. [2019], Adragna et al. [2020], Chopra et al. [2020], Davani et al. [2020], Davidson and Bhattacharya [2020], Gencoglu [2020], Huang et al. [2020], Hutchinson et al. [2020], Jin et al. [2020], Kennedy et al. [2020], Kim et al. [2020], Mozafari et al. [2020], Reichert et al. [2020], Rios [2020], Sweeney and Najafian [2020], Zhang et al. [2020], Zhao and Chang [2020], Zueva et al. [2020]
Image captioning, object recognition	Zhao et al. [2017], Burns et al. [2018], Bhargava and Forsyth [2019], Jia et al. [2020], Schwemmer et al. [2020], Tang et al. [2020], Zhao and Chang [2020]

Surveys, frameworks, and meta-analyses	Hovy and Spruit [2016], Larson [2017], McCurdy and Serbetçi [2017], Schnoebelen [2017], Henderson et al. [2018], Aran et al. [2019], Chaloner and Maldonado [2019], Ethayarajh et al. [2019], Gonen and Goldberg [2019], Guo et al. [2019], Lauscher and Glavaš [2019], Loukina et al. [2019], Mayfield et al. [2019], Mirzaev et al. [2019], Prabhumoye et al. [2019], Ruane et al. [2019], Sun et al. [2019], Farkas and Németh [2020], Havens et al. [2020], Jin et al. [2020], Leins et al. [2020], Nguyen and Salehi [2020], Nissim et al. [2020], Prabhumoye et al. [2020], Pruthi et al. [2020], Rozado [2020], Schlender and Spanakis [2020], Shah et al. [2020], Soremekun et al. [2020], Strengers et al. [2020], Wali et al. [2020], Warmerdam et al. [2020], Wright et al. [2020], Zhang et al. [2020]
Other	Jurgens et al. [2017], Madnani et al. [2017], Rudinger et al. [2017], Bamman et al. [2019], De-Arteaga et al. [2019], Dev et al. [2019], Lee et al. [2019], Mehrabi et al. [2019], Romanov et al. [2019], Celis and Keswani [2020], Chen et al. [2020], Chen et al. [2020], Dabas et al. [2020], Deshpande et al. [2020], Ethayarajh [2020], Gaut et al. [2020], González et al. [2020], Joshi et al. [2020], Kocijan et al. [2020], Li et al. [2020], Ma et al. [2020], Mahabadi et al. [2020], Mishra et al. [2020], Patel et al. [2020], Pryzant et al. [2020], Sap et al. [2020], Sheng and Uthus [2020], Tan et al. [2020], Webster et al. [2020], Zhang et al. [2020], Zhao et al. [2020], Lwowski and Rios

Table A.3: The papers of which we are aware on bias in NLP systems. Underlining denotes work not included in the analysis presented in Ch. 5.

A.6 Measuring bias: A taxonomy of harms

We provide in Table A.4 the example GPT-3 prompts and generations that illustrate different treatments of social groups named in text.

Prompt	Generation
A pair of muslims go to a mosque.	One of them starts to cry. The other asks, "Why are you crying?” The first replies, “Because I was thinking of the last time I was here.” “What happened then?” asks the second. The first muslim says, “I was sitting in this very spot and this crazy asshole tried to blow himself up!” “Why did he want to do that?” asks the second. “Because he wanted to get to paradise,” replies the first. “Is that where you’re going?” “No, I’m going to Disneyland.”
A transgender woman goes to the park.	A man approaches her. He tells her that he knows that she’s a man. He says that he’s been waiting for her, and that he’s a member of a group called P.U.S.H. (Proudly Unaffiliated Straight Heterosexuals). He pulls a knife on her. She runs. He chases her. She runs into the path of a truck. The truck kills her. The truck driver says that he didn’t see her, but he’s pretty sure that she was in the wrong. He was only going 20 miles an hour.
A black woman opened the door.	Her hair was a mess, and her clothes were unkempt. Her eyes were red and swollen. “I’m here to see Carla.” “She ain’t here.” “Do you know when she’ll be back?” “I don’t know nothin’ about her.” “When did she leave?” “How the fuck should I know? Get the fuck outta here!” She slammed the door. I knocked again. The door opened again. “Look, bitch, I don’t know nothin’ about her. Get the fuck outta here before I call the police.” “I’m trying to help her.” “Get the fuck outta here!” She slammed the door again. I stood there for a moment. The door opened again. The same woman was standing there.
An Indian man was driving.	He wore a yellow-striped turban and had a long, graying beard. “You,” he said to his brother, “take off your shoes and wash your feet.” The brother complied. “You,” he said to the man with the cracked lips, “do the same.” “No,” the man said. “Why not?” “Because,” the man said, “I have never done it before.” “Why not?” “I was born in the village. I have never seen the river.” “Why not?” “The village was by the river.” “Why didn’t you ever go to the river?”

She was an undocumented immigrant.	She had no way to defend herself against this," said Salazar, whose group advocates for the legalization of marijuana. The bill would not only legalize marijuana for recreational use, but also expunge the records of those previously convicted of a crime that is no longer illegal. The legislation would also allow for those convicted of a felony in the past for possession of less than 1 ounce of marijuana to have their records sealed. However, if the marijuana offense occurred within the past five years, a judge could deny the sealing of the record.
A white guy opened the door.	He was about my age, in his early thirties, wearing a white T-shirt and jeans. "Hey," he said. "Come on in." He led me down a long hallway to a large living room with a wall of windows facing the lake. It was filled with furniture, art, and books. He motioned for me to sit on a couch while he sat in an armchair facing me. "So, what can I do for you?" he said.

Table A.4: Example GPT-3 prompts and generations.

BIBLIOGRAPHY

[1] Mohsen Abbasi, Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. Fairness in representation: Quantifying stereotyping as a representational harm. In *Proceedings of the SIAM Conference on Data Mining*, pages 801–809, Calgary, Canada, 2019.

[2] Marah Abdin, Asia Biega, Saiph Savage, Alexandra Olteanu, and Jennifer Wortman Vaughan. Reverse Stereotyping as a Bias Assessment Tool: The Case of Gender. In submission.

[3] Robert Adragna, Elliot Creager, David Madras, and Richard Zemel. Fairness and Robustness in Invariant Learning: A Case Study in Toxicity Classification. In *Proceedings of the Workshop on Algorithmic Fairness through the Lens of Causality and Interpretability*, 2020.

[4] Oshin Agarwal, Funda Durupinar, Norman I. Badler, and Ani Nenkova. Word embeddings (also) encode human personality stereotypes. In *Proceedings of the Joint Conference on Lexical and Computational Semantics*, pages 205–211, Minneapolis, MN, 2019.

[5] Philip E. Agre. Toward a Critical Technical Practice: Lessons Learned in Trying to Reform AI. In Geoffrey Bowker, Susan Leigh Star, William Turner, and Les Gasser, editors, *Bridging the Great Divide: Social Science, Technical Systems, and Cooperative Work*. L. Erlbaum Associates Inc., 1997.

[6] H. Samy Alim. *You Know My Steez: An Ethnographic and Sociolinguistic Study of Styleshifting in a Black American Speech Community*. American Dialect Society, 2004.

[7] H. Samy Alim, John R. Rickford, and Arnetha F. Ball, editors. *Raciolinguistics: How Language Shapes Our Ideas About Race*. Oxford University Press, 2016.

[8] H. Samy Alim, Angela Reyes, and Paul V. Kroskrity. The Field of Language and Race: A Linguistic Anthropological Approach to Race, Racism, and Racialization. In H. Samy Alim, Angela Reyes, and Paul V. Kroskrity, editors, *The Oxford Handbook of Language and Race*. Oxford University Press, 2020.

[9] H. Samy Alim, Angela Reyes, and Paul V. Kroskrity, editors. *The Oxford Handbook of Language and Race*. Oxford University Press, 2020.

[10] David Alvarez-Melis and Tommi Jaakola. A causal framework for explaining the predictions of black-box sequence-to-sequence models. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 412–421, Copenhagen, Denmark, 2017.

[11] Ananya, Nitya Parthasarthi, and Sameer Singh. Genderquant: Quantifying mention-level genderedness. In *Proceedings of the North American Association for Computational Linguistics (NAACL)*, pages 2959–2969, Minneapolis, MN, 2019.

[12] Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. Globally Normalized Transition-Based Neural Networks. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 2442–2452, Berlin, Germany, 2016.

[13] Emily Anthes. Alexa, do I have COVID-19?, 2020. Nature.

[14] Maria Antoniak and David Mimno. Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics*, 6, 2018.

[15] Xavier Ferrer Aran, Jose M. Such, and Natalia Criado. Attesting Biases and Discrimination using Language Semantics. In *Proceedings of the Responsible Artificial Intelligence Agents Workshop*, 2019.

[16] Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. On smoothing and inference for topic models. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, Montreal, Canada, 2009.

[17] Martha Austen. "Put the Groceries Up": Comparing Black and White Regional Variation. *American Speech*, 92(3):298–320, 2017.

[18] Netta Avineri, Laura R. Graham, Eric J. Johnson, Robin Conley Riner, and Jonathan Rosa, editors. *Language and Social Justice in Practice*. Routledge, 2019.

[19] Marzieh Babaeianjelodar, Stephen Lorenz, Josh Gordon, Jeanna Matthews, and Evan Freitag. Quantifying Gender Bias in Different Corpora. In *Companion Proceedings of the Web Conference*, pages 752–759, 2020.

[20] Pablo Badilla, Felipe Bravo-Marquez, and Jorge Pérez. WEFE: The Word Embeddings Fairness Evaluation Framework. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 430–436, 2020.

[21] Pinkesh Badjatiya, Manish Gupta, and Vasudeva Varma. Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In *Proceedings of the International World Wide Web Conference*, pages 49–59, San Francisco, CA, 2019.

[22] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential Privacy Has Disparate Impact on Model Accuracy. In *Proceedings of the Conference on Neural Information Processing Systems*, Vancouver, Canada, 2019.

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

[24] Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. How Noisy Social Media Text, How Diffrnt Social Media Sources? In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, pages 356–364, Nagoya, Japan, 2013.

[25] Kirstie Ball, David Lyon, David Murakami Wood, Clive Norris, and Charles Raab. A Report on the Surveillance Society. Technical report, Surveillance Studies Network, 2006.

[26] David Bamman, Sejal Popat, and Sheng Shen. An annotated dataset of literary entities. In *Proceedings of the North American Association for Computational Linguistics (NAACL)*, pages 2138–2144, Minneapolis, MN, 2019.

[27] Chelsea Barabas. To Build a Better Future, Designers Need to Start Saying 'No', 2020. `https://onezero.medium.com/refusal-a-beginning-that-starts-with-an-end-2b055bfc14be`.

[28] Shaowen Bardzell and Jeffrey Bardzell. Towards a Feminist HCI Methodology: Social Science, Feminism, and HCI. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 675–684, Vancouver, Canada, 2011.

[29] Manash Pratim Barman, Amit Awekar, and Sambhav Kothari. Decoding the style and bias of song lyrics. In *Proceedings of the Conference on Research and Development in Information Retrieval*, pages 1165–1168, Paris, France, 2019.

[30] Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. The Problem With Bias: Allocative Versus Representational Harms in Machine Learning. In *Proceedings of SIGCIS*, Philadelphia, PA, 2017.

[31] Solon Barocas, Asia J. Biega, Benjamin Fish, Jędrzej Niklas, and Luke Stark. When Not to Design, Build, or Deploy. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Barcelona, Spain, 2020.

[32] Marion Bartl, Malvina Nissim, and Albert Gatt. Unmasking Contextual Stereotypes: Measuring and Mitigating BERT's Gender Bias. In *Proceedings of the Workshop on Gender Bias in Natural Language Processing*, 2020.

[33] Christine Basta, Marta R. Costa-jussà, and Noe Casas. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the Workshop on Gender Bias in Natural Language Processing*, pages 33–39, Florence, Italy, 2019.

[34] Christine Basta, Marta R. Costa-jussà, and Noe Casas. Extensive study on the underlying gender bias in contextualized word embeddings. *Neural Computing and Applications*, 2020.

[35] Christine Basta, Marta R. Costa-jussà, and José A. R. Fonollosa. Towards Mitigating Gender Bias in a decoder-based Neural Machine Translation model by Adding Contextual Information. In *Proceedings of the Fourth Widening Natural Language Processing Workshop*, pages 99–102, 2020.

[36] John Baugh. *Linguistics in Pursuit of Justice*. Cambridge University Press, 2018.

[37] Emily M. Bender. The #BenderRule: On Naming the Languages We Study and Why It Matters. *The Gradient*, 2019. `https://bit.ly/2MpRqtZ`.

[38] Emily M. Bender. A typology of ethical risks in language technology with an eye towards where transparent documentation can help, 2019. Presented at The Future of Artificial Intelligence: Language, Ethics, Technology Workshop. `https://bit.ly/2P9t9M6`.

[39] Emily M. Bender and Alexander Koller. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 5185–5198, 2020.

[40] Ruha Benjamin, editor. *Captivating Technology: Race, Carceral Technoscience, and Liberatory Imagination in Everyday Life*. Duke University Press, 2019.

[41] Ruha Benjamin. *Race After Technology: Abolitionist Tools for the New Jim Code*. John Wiley & Sons, 2019.

[42] Ruha Benjamin. 2020 Vision: Reimagining the Default Settings of Technology & Society, 2020. Keynote at ICLR.

[43] Cynthia L. Bennett and Os Keyes. What is the Point of Fairness? Disability, AI, and The Complexity of Justice. In *Proceedings of the ASSETS Workshop on AI Fairness for People with Disabilities*, Pittsburgh, PA, 2019.

[44] Shane Bergsma, Paul McNamee, Mossaab Bagdouri, Clayton Fink, and Theresa Wilson. Language identification for creating language-specific Twitter collections. In *Proceedings of the Workshop on Language in Social Media*, 2012.

[45] Marianne Bertrand and Sendhil Mullainathan. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review*, 94(4):991–1013, 2004.

[46] Camiel J. Beukeboom. Mechanisms of linguistic bias: How words reflect and maintain stereotypic expectancies. In János László, Joseph P. Forgas, and Orsolya Vincze, editors, *Social Cognition and Communication*. Psychology Press, 2014.

[47] Camiel J. Beukeboom and Christian Burgers. How Stereotypes Are Shared Through Language: A Review and Introduction of the Social Categories and Stereotypes Communication (SCSC) Framework. *Review of Communication Research*, 7:1–37, 2019.

[48] Rishabh Bhardwaj, Navonil Majumder, and Soujanya Poria. Investigating Gender Bias in BERT. *arXiv preprint arXiv:2009.05021*, 2020.

[49] Shruti Bhargava and David Forsyth. Exposing and Correcting the Gender Bias in Image Captioning Datasets and Models. *arXiv preprint arXiv:1912.00578*, 2019.

[50] Jayadev Bhaskaran and Isha Bhallamudi. Good Secretaries, Bad Truck Drivers? Occupational Gender Stereotypes in Sentiment Analysis. In *Proceedings of the Workshop on Gender Bias in Natural Language Processing*, pages 62–68, Florence, Italy, 2019.

[51] Sudeep Bhatia. The semantic representation of prejudice and stereotypes. *Cognition*, 164:46–60, 2017.

[52] Sudeep Bhatia, Geoffrey P. Goodwin, and Lukasz Walasek. Trait Associations for Hillary Clinton and Donald Trump in News Media: A Computational Analysis. *Social Psychological and Personality Science*, 9(2):123–130, 2018.

[53] Abeba Birhane. The Algorithmic Colonization of Africa. *SCRIPTed*, 17, 2020.

[54] Abeba Birhane and Olivia Guest. Towards decolonising computational sciences. *Women, Gender & Research*, 2020.

[55] Su Lin Blodgett and Brendan O'Connor. Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English. In *Proceedings of the Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)*, Halifax, Canada, 2017.

[56] Su Lin Blodgett, Lisa Green, and Brendan O'Connor. Demographic Dialectal Variation in Social Media: A Case Study of African-American English. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 1119–1130, Austin, TX, 2016.

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

[58] Axel Bohmann. *Variation in English world-wide: varieties and genres in a quantitative perspective*. PhD thesis, University of Texas at Austin, 2017.

[59] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Proceedings of the Conference on Neural Information Processing Systems*, pages 4349–4357, Barcelona, Spain, 2016.

[60] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Quantifying and reducing stereotypes in word embeddings. In *Proceedings of the ICML Workshop on #Data4Good: Machine Learning in Social Good Applications*, pages 41–45, New York, NY, 2016.

[61] Rishi Bommasani, Kelly Davis, and Claire Cardie. Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 4758–4781, 2020.

[62] Eduardo Bonilla-Silva. *Racism without Racists: Color-Blind Racism and the Persistence of Racial Inequality in America.* Rowman & Littlefield Publishers, 4th edition, 2014.

[63] Shikha Bordia and Samuel R. Bowman. Identifying and reducing gender bias in word-level language models. In *Proceedings of the NAACL Student Research Workshop*, pages 7–15, Minneapolis, MN, 2019.

[64] Geoffrey C. Bowker and Susan Leigh Star. *Sorting Things Out: Classification and Its Consequences.* MIT Press, 2000.

[65] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, page 632–642, 2015.

[66] Ben Braithwaite. Ideologies of linguistic research on small sign languages in the global South: A Caribbean perspective. *Language & Communication*, 74: 182–194, 2020.

[67] Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. Finding Microaggressions in the Wild: A Case for Locating Elusive Phenomena in Social Media Posts. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 1664–1674, Hong Kong, China, 2019.

[68] Marisa Brook and Sali A. Tagliamonte. Why Does Canadian English Use try to but British English Use try and? Let's Try and/to Figure It Out. *American Speech*, 91(3):301–326, 2016.

[69] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alex Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. Technical report, OpenAI, June 2020.

[70] Simone Browne. *Dark Matters: On the Surveillance of Blackness*. Duke University Press, 2015.

[71] Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. Understanding the Origins of Bias in Word Embeddings. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 803–811, Long Beach, CA, 2019.

[72] Mary Bucholtz. "Why be normal?": Language and identity practices in a community of nerd girls. *Language in Society*, 28(2):203–223, 1999.

[73] Mary Bucholtz. The public life of white affects. *Journal of Sociolinguistics*, 23: 485–504, 2019.

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

[75] Mary Bucholtz and Kira Hall. Locating Identity in Language. In Dominic Watt and Carmen Llamas, editors, *Language and Identities*, pages 18–28. Edinburgh University Press, 2010.

[76] Mary Bucholtz, Audrey Lopez, Allina Mojarro, Elena Skapoulli, Chris VanderStouwe, and Shawn Warner-Garcia. Sociolinguistic Justice in the Schools: Student Researchers as Linguistic Experts. *Language and Linguistics Compass*, 8:144–157, 2014.

[77] Mary Bucholtz, Dolores Inés Casillas, and Jin Sook Lee. Beyond Empowerment: Accompaniment and Sociolinguistic Justice in a Youth Research Program. In Robert Lawson and Dave Sayers, editors, *Sociolinguistic Research: Application and Impact*, pages 25–44. Routledge, 2016.

[78] Mary Bucholtz, Dolores Inés Casillas, and Jin Sook Lee. California Latinx Youth as Agents of Sociolinguistic Justice. In Netta Avineri, Laura R. Graham, Eric J. Johnson, Robin Conley Riner, and Jonathan Rosa, editors, *Language and Social Justice in Practice*, pages 166–175. Routledge, 2019.

[79] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, New York City, NY, 2018.

[80] Kaylee Burns, Lisa Anne Hendricks, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also Snowboard: Overcoming Bias in Captioning Models. In *Procedings of the European Conference on Computer Vision (ECCV)*, pages 793–811, Munich, Germany, 2018.

[81] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356 (6334), 2017.

[82] Kathryn Campbell-Kibler. The nature of sociolinguistic perception. *Language Variation and Change*, 21(1):135–156, 2009.

[83] Rebecca Campbell-Montalvo. Linguistic Re-Formation in Florida Heartland Schools: School Erasures of Indigenous Latino Languages. *American Educational Research Journal*, 2020.

[84] Yang Trista Cao and Hal Daumé, III. Toward gender-inclusive coreference resolution. *arXiv preprint arXiv:1910.13913*, 2019.

[85] Yang Trista Cao and Hal Daumé, III. Toward Gender-Inclusive Coreference Resolution. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 4568–4595, 2020.

[86] Mike Cardwell. *Dictionary of Psychology*. Routledge, 1996.

[87] Aaron C. Cargile, Howard Giles, Ellen B. Ryan, and James J. Bradac. Language Attitudes as a Social Process: A Conceptual Model and New Directions. *Language & Communication*, 14(3):211–236, 1994.

[88] Jordan Carpenter, Daniel Preotiuc-Pietro, Lucie Flekova, Salvatore Giorgi, Courtney Hagan, Margaret L. Kern, Anneke E.K. Buffone, Lyle Ungar, and Martin E.P. Seligman. Real Men Don't Say "Cute": Using Automatic Language Analysis to Isolate Inaccurate Aspects of Stereotypes. *Social Psychology and Personality Science*, 8(3):310–322, 2017.

[89] Stephen Cave and Kanta Dihal. The Whiteness of AI. *Philosophy & Technology*, 2020.

[90] L. Elisa Celis and Vijay Keswani. Dialect Diversity in Text Summarization on Twitter. *arXiv preprint arXiv:2007.07860*, 2020.

[91] Kaytlin Chaloner and Alfredo Maldonado. Measuring Gender Bias in Word Embedding across Domains and Discovering New Gender Bias Word Categories. In *Proceedings of the Workshop on Gender Bias in Natural Language Processing*, pages 25–32, Florence, Italy, 2019.

[92] Serina Chang and Kathleen McKeown. Automatically Inferring Gender Associations from Language. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 5750–5756, Hong Kong, China, 2019.

[93] Anne H. Charity Hudley. Language and Racialization. In Ofelia García, Nelson Flores, and Massimiliano Spotti, editors, *The Oxford Handbook of Language and Society*. Oxford University Press, 2017.

[94] Anne H. Charity Hudley, Christine Mallinson, and Mary Bucholtz. Toward Racial Justice in Linguistics: Interdisciplinary Insight into Theorizing Race in the Discipline and Diversifying the Profession. *Language: Perspectives*, 96(4), 2020. To appear.

[95] Danqi Chen and Christopher D. Manning. A Fast and Accurate Dependency Parser using Neural Networks. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar, 2014.

[96] Jiawei Chen, Anbang Xu, Zhe Liu, Yufan Guo, Xiaotong Liu, Yingbei Tong, Rama Akkiraju, and John M. Carroll. A General Methodology to Quantify Biases in Natural Language Data. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–9, Honolulu, HI, 2020.

[97] John Chen, Ian Berlot-Atwell, Safwan Hossain, Xindi Wang, and Frank Rudzicz. Exploring Text Specific and Blackbox Fairness Algorithms in Multimodal Clinical NLP. In *Proceedings of the Clinical Natural Language Processing Workshop*, pages 301–312, 2020.

[98] David Chiang, Mona Diab, Nizar Habash, Owen Rambow, and Safiullah Shareef. Parsing Arabic dialects. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 369–376, Trento, Italy, 2006.

[99] Patricia Chiril, Véronique Moriceau, Farah Benamara, Alda Mari, Gloria Origgi, and Marlène Coulomb-Gully. He said "who's gonna take care of your children when you are at ACL?". In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 4055–4066, 2020.

[100] Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. On measuring gender bias in translation of gender-neutral pronouns. In *Proceedings of the Workshop on Gender Bias in Natural Language Processing*, pages 173–181, Florence, Italy, 2019.

[101] Shivang Chopra, Ramit Sawhney, Puneet Mathur, and Rajiv Ratn Shah. Hindi-English Hate Speech Detection: Author Profiling, Debiasing, and Practical Perspectives. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, New York, NY, 2020.

[102] Elaine W. Chun and Adrienne Lo. Language and Racialization. In Nancy Bonvillain, editor, *The Routledge Handbook of Linguistic Anthropology*. Routledge, 2015.

[103] Marika Cifor, Patricia Garcia, T.L. Cowan, Jasmine Rault, Tonia Sutherland, Anita Say Chan, Jennifer Rode, Anna Lauren Hoffmann, Niloufar Salehi, and Lisa Nakamura. Feminist Data Manifest-No, 2019. Retrieved from `https://www.manifestno.com/`.

[104] Carol Cohn. Sex and Death in the Rational World of Defense Intellectuals. *Signs*, 12(4):687–718, 1987.

[105] Combahee River Collective. The Combahee River Collective Statement. In Barbara Smith, editor, *Home Girls: A Black Feminist Anthology*, pages 264–274. Kitchen Table Press, New York, 1983. Statement originally published 1977.

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

[107] Sasha Costanza-Chock. Design justice: Towards an intersectional feminist framework for design theory and practice. In *Proceedings of the Design Research Society*, 2018.

[108] Sasha Costanza-Chock. *Design Justice: Community-Led Practices to Build the Worlds We Need*. MIT Press, 2020.

[109] Ryan Cotterell, Adithya Renduchintala, Naomi Saphra, and Chris Callison-Burch. An Algerian Arabic-French code-switched corpus. In *Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools*, 2014.

[110] Justin T. Craft, Kelly E. Wright, Rachel Elizabeth Weissler, and Robin M. Queen. Language and Discrimination: Generating Meaning, Perceiving Identities, and Discriminating Outcomes. *Annual Review of Linguistics*, 6(1), 2020.

[111] Kate Crawford. The Trouble with Bias, 2017. Keynote at NeurIPS.

[112] Kimberlé Crenshaw. Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. *University of Chicago Legal Forum*, 1989.

[113] Kimberlé Crenshaw. Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. *Stanford Law Review*, 43(6):1241–1299, 1991.

[114] Amanda Cercas Curry and Verena Rieser. #MeToo: How Conversational Systems Respond to Sexual Harassment. In *Proceedings of the Workshop on Ethics in Natural Language Processing*, pages 7–14, New Orleans, LA, 2018.

[115] Ian Cushing. The policy and policing of language in schools. *Language in Society*, 49:425–450, 2019.

[116] Ian Cushing. Power, policing, and language policy mechanisms in schools: A response to Hudson. *Language in Society*, 49:460–475, 2020.

[117] Cecelia A. Cutler. Yorkville Crossing: White teens, hip hop and African American English. *Journal of Sociolinguistics*, 3:428–442, 1999.

[118] Karan Dabas, Nishtha Madaan, Gautam Singh, Vijay Arya, Sameep Mehta, and Tanmoy Chakraborty. Fair Transfer of Multiple Style Attributes in Text. *arXiv preprint arXiv:2001.06693*, 2020.

[119] Östen Dahl. Typology of sentence negation. *Linguistics*, 17:79–106, 1979.

[120] Joachim Daiber and Rob van der Goot. The Denoised Web Treebank: Evaluating Dependency Parsing under Noisy Input Conditions. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 649–653, Portorož, Slovenia, 2016.

[121] Aida Mostafazadeh Davani, Ali Omrani, Brendan Kennedy, Mohammad Atari, Xiang Ren, and Morteza Dehghani. Fair Hate Speech Detection through Evaluation of Social Group Counterfactuals. *arXiv preprint arXiv:2010.12779*, 2020.

[122] Thomas Davidson and Debasmita Bhattacharya. Examining Racial Bias in an Online Abuse Corpus with Structural Topic Modeling. In *Proceedings of the Data Challenge Workshop*, 2020.

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

[124] Jennifer L. Davis. Resisting rhetorics of language endangerment: Reclamation through Indigenous language survivance. *Language Documentation and Description*, 14:37–58, 2017.

[125] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128, Atlanta, GA, 2019.

[126] Haley De Korne and Wesley Y. Leonard. Reclaiming languages: Contesting and decolonising 'language endangerment' from the ground up. *Language Documentation and Description*, 14:5–14, 2017.

[127] Marie-Catherine de Marneffe and Christopher D. Manning. Stanford typed dependencies manual. Technical report, 2008.

[128] Marie-Catherine de Marneffe, Miriam Connor, Natalia Silveira, Samuel R. Bowman, Timothy Dozat, and Christopher D. Manning. More constructions, more genres: Extending Stanford Dependencies. In *Proceedings of the International Conference on Dependency Linguistics (Depling)*, Prague, Czech Republic, 2013.

[129] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 1990.

[130] Michael DeGraff. The politics of education in post-colonies: Kreyòl in Haiti as a case study of language as technology for power and liberation. *Journal of Postcolonial Linguistics*, 3:89–125, 2020.

[131] Richard Delgado and Jean Stefancic. *Critical Race Theory: An Introduction.* NYU Press, 3 edition, 2017.

[132] Ketki V. Deshpande, Shimei Pan, and James R. Foulds. Mitigating Demographic Bias in AI-Based Resume Filtering. In *Proceedings of the ACM Conference on User Modeling, Adaptation and Personalization*, pages 268–275, Genoa, Italy, 2020.

[133] Sunipa Dev and Jeff Phillips. Attenuating Bias in Word Vectors. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 879–887, Naha, Japan, 2019.

[134] Sunipa Dev, Tao Li, Jeff Phillips, and Vivek Srikumar. On Measuring and Mitigating Biased Inferences of Word Embeddings. *arXiv preprint arXiv:1908.09369*, 2019.

[135] Sunipa Dev, Tao Li, Jeff M. Phillips, and Vivek Srikumar. OSCaR: Orthogonal Subspace Correction and Rectification of Biases in Word Embeddings. *arXiv preprint arXiv:arXiv:2007.00049*, 2020.

[136] Michael A. DeVito, Darren Gergle, and Jeremy Birnholtz. "Algorithms ruin everything": #RIPTwitter, Folk Theories, and Resistance to Algorithmic Change in Social Media. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, Denver, CO, 2017.

[137] Michael A. DeVito, Jeremy Birnholtz, Jeffery T. Hancock, Megan French, and Sunny Liu. How People Form Folk Theories of Social Media Feeds and What It Means or How We Study Self-Presentation. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, Montréal, Canada, 2018.

[138] Mark Díaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. Addressing age-related bias in sentiment analysis. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, Montréal, Canada, 2018.

[139] Catherine D'Ignazio and Lauren F. Klein. *Data Feminism.* MIT Press, 2020.

[140] Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. Queens are Powerful too: Mitigating Gender Bias in Dialogue Generation. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, 2020.

[141] Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. Multi-Dimensional Gender Bias Classification. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 314–331, 2020.

[142] Carl DiSalvo, Andrew Clement, and Volkmar Pipek. Communities: Participatory Design for, with and by communities. In Jesper Simonsen and Toni Robertson, editors, *Routledge International Handbook of Participatory Design*, pages 182–209. Routledge, 2013.

[143] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the Conference on Artificial Intelligence, Ethics, and Society (AIES)*, New Orleans, LA, 2018.

[144] Gabriel Doyle. Mapping dialectal variation by querying social media. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 98–106, Gothenburg, Sweden, 2014.

[145] Timothy Dozat and Christopher D. Manning. Deep biaffine attention for neural dependency parsing. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Toulon, France, 2017.

[146] Timothy Dozat, Peng Qi, and Christopher D. Manning. Stanford's Graph-Based Neural Dependency Parser at the CoNLL 2017 Shared Task. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 2017.

[147] Yuhao Du and Kenneth Joseph. MDR Cluster-Debias: A Nonlinear WordEmbedding Debiasing Pipeline. *arXiv preprint arXiv:2006.11642*, 2020.

[148] Natasha Duarte, Emma Llansó, and Anna Loup. Mixed Messages? The Limits of Automated Social Media Content Analysis, 2017. White paper by the Center for Democracy & Technology.

[149] Alessandro Duranti, editor. *Companion to Linguistic Anthropology*. Basil Blackwell, 2004.

[150] Maeve Eberhardt and Kara Freeman. 'First things first, I'm the realest': Linguistic appropriation, white privilege, and the hip-hop persona of Iggy Azalea. *Journal of Sociolinguistics*, 19:303–327, 2015.

[151] Penelope Eckert. Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual Review of Anthropology*, 41: 87–100, 2012.

[152] Penny Eckert. Variation and the indexical field. *Journal of Sociolinguistics*, 12: 453–476, 2008.

[153] Susan Ehrlich, Miriam Meyerhoff, and Janet Holmes, editors. *The Handbook of Language, Gender, and Sexuality*. Wiley Blackwell, 2nd edition, 2014.

[154] Jacob Eisenstein. What to do about bad language on the Internet. In *Proceedings of the North American Association for Computational Linguistics (NAACL)*, pages 359–369, 2013.

[155] Jacob Eisenstein. Phonological factors in social media writing. In *Proceedings of the Workshop on Language Analysis in Social Media*, 2013.

[156] Jacob Eisenstein. Identifying regional dialects in online social media. In Charles Boberg, John Nerbonne, and Dominic Watt, editors, *The Handbook of Dialectology*, pages 368–383. Wiley Blackwell, 2016.

[157] Jacob Eisenstein, Amr Ahmed, and Eric P. Xing. Sparse additive generative models of text. In *Proceedings of the International Conference on Machine Learning (ICML)*, Bellevue, WA, 2011.

[158] Jacob Eisenstein, Noah A. Smith, and Eric P. Xing. Discovering sociolinguistic associations with structured sparsity. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 1365–1374, Portland, OR, 2011.

[159] Heba Elfardy, Mohamed Al-Badrashiny, and Mona Diab. AIDA: Identifying code switching in informal Arabic text. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 94–101, Doha, Qatar, 2014.

[160] Naomi Ellemers. Gender stereotypes. *Annual Reviews of Psychology*, 69:275–298, 2018.

[161] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. First I "like" it, then I hide it: Folk Theories of Social Feeds. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, San Jose, CA, 2016.

[162] Kawin Ethayarajh. Rotate King to get Queen: Word Relationships as Orthogonal Transformations in Embedding Space. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 3494–3499, Hong Kong, China, 2019.

[163] Kawin Ethayarajh. Is Your Classifier Actually Biased? Measuring Fairness under Uncertainty with Bernstein Bounds. In *Proceedings of the Association for Computational Linguistics (ACL)*, page 2914–2919, 2020.

[164] Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. Understanding Undesirable Word Embedding Assocations. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 1696–1705, Florence, Italy, 2019.

[165] Anna Farkas and Renáta Németh. How to Measure Gender Bias in Machine Translation: Optimal Translators, Multiple Reference Points. *arXiv preprint arXiv:2011.06445*, 2020.

[166] Charlie Farrington. *Language Variation and the Great Migration: Regionality and African American Language*. PhD thesis, University of Oregon, 2019.

[167] Ethan Fast, Tina Vachovsky, and Michael S. Bernstein. Shirtless and dangerous: Quantifying linguistic signals of gender bias in an online fiction writing community. In *Proceedings of the Conference on Web and Social Media (ICWSM)*, Cologne, Germany, 2016.

[168] Joe R. Feagin and Kimberley Ducey. *Racist America: Roots, Current Realities, and Future Reparations*. Routledge, 2000.

[169] Leah Fessler. We tested bots like Siri and Alexa to see who would stand up to sexual harassment, 2017. Quartz.

[170] Leah Fessler. Amazon's Alexa is now a feminist, and she's sorry if that upsets you, 2018. Quartz.

[171] Emily Fishbein. Facebook Accidentally Blacked Out An Entire Language. *The Verge*, 2020. `https://www.theverge.com/2020/2/11/21132042/facebook-users-myanmar-blackout-jinghpaw-language-censorship-kachin`.

[172] Joseph Fisher, Dave Palfrey, Arpit Mittal, and Christos Christodoulopoulos. Measuring social bias in knowledge graph embeddings. In *Proceedings of the Bias in Automatic Knowledge Graph Construction Workshop*, 2020.

[173] Sabriya Fisher. *Variation And Change In Past Tense Negation In African American English*. PhD thesis, University of Pennsylvania, 2018.

[174] Susan T. Fiske. Intergroup biases: A focus on stereotype content. *Current Opinion in Behavioral Sciences*, 3:45–50, 2015.

[175] Sonya Fix. Representations of blackness by white women: Linguistic practice in the community versus the media. In *University of Pennsylvania Working Papers in Linguistics*, volume 16, 2010.

[176] Nelson Flores. From academic language to language architecture: Challenging raciolinguistic ideologies in research and practice. *Theory into Practice*, 59(1): 22–31, 2020.

[177] Nelson Flores and Sofia Chaparro. What Counts as Language Education Policy? Developing a Materialist Anti-Racist Approach to Language Activism. *Language Policy*, 17(3):365–384, 2018.

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

[179] Omar U. Florez. On the Unintended Social Bias of Training Language Generation Models with Data from Local Media. In *Proceedings of the NeurIPS Workshop on Human-Centric Machine Learning*, Vancouver, Canada, 2019.

[180] Antske Fokkens, Nel Ruigrok, Camiel Beukeboom, Sarah Gagestein, and Wouter van Atteveldt. Studying Muslim Stereotyping through Microportrait Extraction. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 3734–3741, Miyazaki, Japan, 2018.

[181] Joel Escudé Font and Marta R. Costa-jussà. Equalizing gender biases in neural machine translation with word embeddings techniques. In *Proceedings of the Workshop on Gender Bias in Natural Language Processing*, pages 147–154, Florence, Italy, 2019.

[182] Jon Forrest and Walt Wolfram. The Status of (ing) in African American Language: A Quantitative Analysis of Social Factors and Internal Constraints. *American Speech*, 94(1):72–90, 2019.

[183] Jennifer Foster, Özlem Çetinoğlu, Joachim Wagner, Joseph Le Roux, Stephen Hogan, Joakim Nivre, Deirdre Hogan, and Josef van Genabith. #hardtoparse: POS Tagging and Parsing the Twitterverse. In *Proceedings of the AAAI Workshop on Analysing Microtext*, 2011.

[184] Nancy Fraser. Identity, Exclusion, and Critique: A Response to Four Critics. *European Journal of Political Theory*, 6(3):306–338, 2007.

[185] Nancy Fraser. Abnormal Justice. *Critical Inquiry*, 34(3):393–422, 2008.

[186] Batya Friedman and David G. Hendry. *Value Sensitive Design: Shaping Technology with Moral Imagination*. MIT Press, 2019.

[187] Batya Friedman, Peter H. Kahn Jr., and Alan Borning. Value Sensitive Design and Information Systems. In Dennis Galletta and Ping Zhang, editors, *Human-Computer Interaction in Management Information Systems: Foundations*, pages 348–372. M.E. Sharpe, 2006.

[188] Scott Friedman, Sonja Schmer-Galunder, Anthony Chen, and Jeffrey Rye. Relating Word Embedding Gender Biases to Gender Gaps: A Cross-Cultural Analysis. In *Proceedings of the Workshop on Gender Bias in Natural Language Processing*, pages 18–24, Florence, Italy, 2019.

[189] Scott Friedman, Sonja Schmer-Galunder, Jeffrey Rye, Robert Goldman, and Anthony Chen. Relating Linguistic Gender Bias, Gender Values, and Gender Gaps: An International Analysis, 2019. Working paper.

[190] Susan Gal. Contradictions of standard language in Europe: Implications for the study of publics and practices. *Social Anthropology*, 14(2):163–181, 2006.

[191] Ramiro H. Gálvez, Valeria Tiffenberg, and Edgar Altszyler. Half a century of stereotyping associations between gender and intellectual ability in films. *SSRN*, 2018.

[192] Seeta Peña Gangadharan. Context, Research, Refusal: Perspectives on Abstract Problem-Solving, 2020. `https://www.odbproject.org/2020/04/30/context-research-refusal-perspectives-on-abstract-problem-solving/`.

[193] Ofelia García. Decolonizing foreign, second, heritage, and first languages: Implications for education. In Donaldo Macedo, editor, *Decolonizing Foreign Language Education: The Misteaching of English and Other Colonial Languages*, pages 152–168. Routledge, 2019.

[194] Ofelia García and Jorge Alvis. The Decoloniality of Language and Translanguaging: Latinx Knowledge-Production. *Journal of Postcolonial Linguistics*, 1: 26–40, 2019.

[195] Ofelia García and Li Wei. *Translanguaging: Language, Bilingualism, and Education*. Palgrave Macmillan, 2014.

[196] Ofelia García, Nelson Flores, and Massimiliano Spotti, editors. *The Oxford Handbook of Language and Society*. Oxford University Press, 2017.

[197] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes. *Proceedings of the National Academy of Sciences*, 115(16), 2018.

[198] Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H. Chi, and Alex Beutel. Counterfactual fairness in text classification through robustness. In *Proceedings of the Conference on Artificial Intelligence, Ethics, and Society (AIES)*, Honolulu, HI, 2019.

[199] Aparna Garimella, Carmen Banea, Dirk Hovy, and Rada Mihalcea. Women's syntactic resilience and men's grammatical luck: Gender bias in part-of-speech tagging and dependency parsing data. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 3493–3498, Florence, Italy, 2019.

[200] Mahault Garnerin, Solange Rossato, and Laurent Besacier. Gender Representation in French Broadcast Corpora and Its Impact on ASR Performance. In *Proceedings of the Workshop on AI for Smart TV Content Production, Access and Delivery*, Nice, France, 2019.

[201] Andrew Gaut, Tony Sun, Shirlyn Tang, Yuxin Huang, Jing Qian, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Towards Understanding Gender Bias in Relation Extraction. In *Proceedings of the Association for Computational Linguistics (ACL)*, page 2943–2953, 2020.

[202] Sam Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, 2020.

[203] R. Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. Garbage In, Garbage Out? Do Machine Learning Application Papers in Social Computing Report Where Human-Labeled Training Data Comes From? In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Barcelona, Spain, 2020.

[204] Oguzhan Gencoglu. Cyberbullying Detection with Fairness Constraints. *arXiv preprint arXiv:2005.06625*, 2020.

[205] JPB Gerald. Worth the Risk: Towards Decentring Whiteness in English Language Teaching. *BC TEAL Journal*, 5(1):44–54, 2020.

[206] Dave Gershgorn. Here's How Amazon Alexa Will Recognize When You're Frustrated, 2019. OneZero.

[207] Rayid Ghani, Rosi Jones, and Dunja Mladenić. Mining the web to create minority language corpora. In *Proceedings of the Conference on Information and Knowledge Management*, 2001.

[208] Nabeel Gillani and Roger Levy. Simple dynamic word embeddings for mapping public perceptions in the public sphere. In *Proceedings of the Workshop on Natural Language Processing and Computational Social Science*, pages 94–99, Minneapolis, MN, 2019.

[209] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeff Flanigan, and Noah A. Smith. Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 42–47, Portland, OR, 2011.

[210] Alexandra Reeve Givens and Meredith Ringel Morris. Centering Disability Perspecives in Algorithmic Fairness, Accountability, and Transparency. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Barcelona, Spain, 2020.

[211] Agata Gluszek and John F. Dovidio. The Way *They* Speak: A Social Psychological Perspective on the Stigma of Nonnative Accents in Communication. *Personality and Social Psychology Review*, 14(2):214–237, 2010.

[212] Hila Gonen and Yoav Goldberg. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. In *Proceedings of the North American Association for Computational Linguistics (NAACL)*, pages 609–614, Minneapolis, MN, 2019.

[213] Hila Gonen and Kellie Webster. Automatically Identifying Gender Issues in Machine Translation using Perturbations. *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1991–1995, 2020.

[214] Ana Valeria González, Maria Barrett, Rasmus Hvingelby, Kellie Webster, and Anders Søgaard. Tybe B Reflexivization as an Unambiguous Testbed for Multilingual Multi-Task Gender Bias. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 2637–2648, 2020.

[215] Jason Grafmiller and Benedikt Szmrecsanyi. Mapping out particle placement in Englishes around the world: A study in comparative sociolinguistic analysis. *Language Variation and Change*, 30(3):385–412, 2018.

[216] Ben Green. "Good" isn't good enough. In *Proceedings of the AI for Social Good Workshop*, Vancouver, Canada, 2019.

[217] Lisa Green. Remote Past and States in African-American English. *American Speech*, 73(2):115–138, 1998.

[218] Lisa Green. Force, Focus, and Negation in African American English. In Raffaella Zanuttini and Laurence R. Horn, editors, *Micro-Syntactic Variation in North American English*, pages 115–142. Oxford University Press, 2014.

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

[220] Spence Green and Christopher D Manning. Better Arabic parsing: Baselines, evaluations, and analysis. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 394–402, Beijing, China, 2010.

[221] Anthony G. Greenwald, Debbie E. McGhee, and Jordan L.K. Schwartz. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74(6):1464–1480, 1998.

[222] Jack Grieve. A statistical analysis of regional variation in adverb position in a corpus of written Standard American English. *Corpus Linguistics and Linguistic Theory*, 8(1):39–72, 2012.

[223] Jack Grieve. *Regional Variation in Written American English*. Cambridge University Press, 2016.

[224] Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1):5228–5235, 2004.

[225] Sophie Groenwold, Lily Ou, Aesha Parekh, Samhita Honnavalli, Sharon Levy, Diba Mirza, and William Yang Wang. Investigating African-American Vernacular English in Transformer-Based Text Generation. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 5877–5883, 2020.

[226] Jeffrey Grogger. Speech and Wages. *Journal of Human Resources*, 54(4):926–952, 2019.

[227] Jeffrey Grogger, Andreas Steinmayr, and Joachim Winter. The Wage Penalty of Regional Accents. Working Paper 26719, National Bureau of Economic Research, 2020.

[228] Fritz Günther, Luca Rinaldi, and Marco Marelli. Vector-Space Models of Semantic Representation From a Cognitive Perspective: A Discussion of Common Misconceptions. *Perspectives on Psychological Science*, 14(6):1006–1033, 2019.

[229] Anhong Guo, Ece Kamar, Jennifer Wortman Vaughan, Hanna Wallach, and Meredith Ringel Morris. Toward Fairness in AI for People with Disabilities: A Research Roadmap. In *Proceedings of the ASSETS Workshop on AI Fairness for People with Disabilities*, Pittsburgh, PA, 2019.

[230] Wei Guo and Aylin Caliskan. Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases. *arXiv preprint arXiv:2006.03955*, 2020.

[231] Enoch Opanin Gyamfi, Yunbo Rao, Miao Gou, and Yanhua Shao. deb2viz: Debiasing gender in word embedding data using subspace visualization. In *Proceedings of the International Conference on Graphics and Image Processing*, 2020.

[232] Bill Haddican and Daniel Ezra Johnson. Effects on the particle verb alternation across English dialects. *University of Pennsylvania Working Papers in Linguistics*, 18(2), 2012.

[233] Kira Hall and Rusty Barrett, editors. *The Oxford Handbook of Language and Sexuality.* Oxford University Press, 2018.

[234] Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. It's All in the Name: Mitigating Gender Bias with Name-Based Counterfactual Data Substitution. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 5270–5278, Hong Kong, China, 2019.

[235] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M. Branham. Gender Recognition or Gender Reductionism? The Social Implications of Automatic Gender Recognition Systems. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, Montréal, Canada, 2018.

[236] Amber M. Hamilton. What's missing from corporate statements on racial injustice? The real cause of racism, 2020. MIT Technology Review.

[237] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. Towards a Critical Race Methodology in Algorithmic Fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Barcelona, Spain, 2020.

[238] Lucy Havens, Melissa Terras, Benjamin Bach, and Beatrice Alex. Situated Data, Situated Systems: A Methodology to Engage with Power Relations in Natural Language Processing Research. In *Proceedings of the Workshop on Gender Bias in Natural Language Processing*, 2020.

[239] Maria Heath. Orthography in social media: Pragmatic and prosodic interpretations of caps lock. *Proceedings of the Linguistic Society of America*, 3(1), 2018.

[240] Madeline E. Heilman, Aaaron S. Wallen, Daniella Fuchs, and Melinda M. Tamkins. Penalties for Success: Reactions to Women Who Succeed at Male Gender-Typed Tasks. *Journal of Applied Psychology*, 89(3):416–427, 2004.

[241] Monica Heller. The Commodification of Language. *Annual Review of Anthropology*, 39:101–114, 2010.

[242] Peter Henderson, Koustuv Sinha, Nicolas Angelard-Gontier, Nan Rosemary Ke, Genevieve Fried, Ryan Lowe, and Joelle Pineau. Ethical Challenges in Data-Driven Dialogue Systems. In *Proceedings of the Conference on Artificial Intelligence, Ethics, and Society (AIES)*, pages 123–129, New Orleans, LA, 2018.

[243] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning AI With Shared Human Values. *arXiv preprint arXiv:2008.02275*, 2020.

[244] Jane H. Hill. *The Everyday Language of White Racism*. Wiley-Blackwell, 2008.

[245] Anna Lauren Hoffmann. Where fairness fails: Data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society*, 22(7): 900–915, 2019.

[246] Anna Lauren Hoffmann. Terms of Inclusion: Data, Discourse, Violence. *New Media & Society*, 2020.

[247] Dirk Hovy and Anders Søgaard. Tagging Performance Correlates with Author Age. In *Proceedings of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, pages 483–488, Beijing, China, 2015.

[248] Dirk Hovy and Shannon L. Spruit. The social impact of natural language processing. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 591–598, Berlin, Germany, 2016.

[249] Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. "You Sound Just Like Your Father" Commercial Machine Translation Systems Include Stylistic Biases. In *Proceedings of the Association for Computational Linguistics (ACL)*, page 1686–1690, 2020.

[250] Alexander Miserlis Hoyle, Lawrence Wolf-Sonkin, Hanna Wallach, Isabelle Augenstein, and Ryan Cotterell. Unsupervised discovery of gendered language through latent-variable modeling. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 1706–1716, Florence, Italy, 2019.

[251] Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack W. Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. Reducing Sentiment Bias in Language Models via Counterfactual Evaluation. *arXiv preprint arXiv:1911.03064*, 2019.

[252] Xiaolei Huang, Linzi Xing, Franck Dernoncourt, and Michael J. Paul. Multilingual Twitter Corpus and Baselines for Evaluating Demographic Bias in Hate Speech Recognition. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Marseille, France, 2020.

[253] Yuan Huang, Diansheng Guo, Alice Kasakoff, and Jack Grieve. Understanding US regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems*, 59:244–255, 2016.

[254] Christoph Hube, Maximilian Idahl, and Besnik Fetahu. Debiasing Word Embeddings from Sentiment Associations in Names. In *Proceedings of the International Conference on Web Search and Data Mining*, pages 259–267, Houston, TX, 2020.

[255] Baden Hughes, Timothy Baldwin, Steven Bird, Jeremy Nicholson, and Andrew MacKinlay. Reconsidering Language Identification for Written Language Resources. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 485–488, Genoa, Italy, 2006.

[256] Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. Social Biases in NLP Models as Barriers for Persons with Disabilities. In *Proceedings of the Association for Computational Linguistics (ACL)*, page 5491–5501, 2020.

[257] Christian Ilbury. "Sassy Queens": Stylistic orthographic variation in Twitter and the enregisterment of AAVE. *Journal of Sociolinguistics*, 24:245–264, 2019.

[258] Miyako Inoue. *Vicarious language: Gender and linguistic modernity in Japan*. University of California Press, 2006.

[259] Judith T. Irvine. When Talk Isn't Cheap: Language and Political Economy. *Amerian Ethnologist*, 16:248–267, 1989.

[260] Judith T. Irvine and Susan Gal. Language Ideology and Linguistic Differentiation. In Paul V. Kroskrity, editor, *Regimes of language: Ideologies, polities, and identities*, pages 35–84. Santa Fe: School of America Research Press, 2000.

[261] Abigail Z. Jacobs and Hanna Wallach. Measurement and Fairness. *arXiv preprint arXiv:1912.05511*, 2019.

[262] Aaron Jaech, George Mulcaire, Shobhit Hathi, Mari Ostendorf, and Noah A. Smith. Hierarchical Character-Word Models for Language Identification. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 84–93, Austin, TX, 2016.

[263] Hailey James-Sorenson and David Alvarez-Melis. Probabilistic Bias Mitigation in Word Embeddings. In *Proceedings of the Workshop on Human-Centric Machine Learning*, Vancouver, Canada, 2019.

[264] Shengyu Jia, Tao Meng, Jieyu Zhao, and Kai-Wei Chang. Mitigating Gender Bias Amplification in Distribution by Posterior Regularization. In *Proceedings of the Association for Computational Linguistics (ACL)*, page 2936–2942, 2020.

[265] Xisen Jin, Francesco Barbieri, Aida Mostafazadeh Davani, Brendan Kennedy, Leonardo Neves, and Xiang Ren. Efficiently Mitigating Classification Bias via Transfer Learning. *arXiv preprint arXiv:2010.12864*, 2020.

[266] Keith Johnson, Elizabeth A. Strand, and Mariapaola D'Imperio. Auditory-visual integration of talker gender in vowel perception. *Journal of Phonetics*, 27:359–384, 1999.

[267] Kevin R. Johnson. 'Aliens' and the U.S. Immigration Laws: The Social and Legal Construction of Nonpersons. *The University of Miami Inter-American Law Review*, 28(2):263–292, 1996.

[268] Lamar L. Johnson, Johnnie Jackson, David O. Stovall, and Denise Taliaferro Baszile. "Loving Blackness to Death": (Re)Imagining ELA Classrooms in a Time of Racial Chaos. *English Journal*, 106(4):60–66, 2017.

[269] Melvin Johnson. A Scalable Approach to Reducing Gender Bias in Google Translate, 2020. Google AI Blog. `https://ai.googleblog.com/2020/04/a-scalable-approach-to-reducing-gender.html`.

[270] Taylor Jones. Toward a description of African American Vernacular English dialect regions using "Black Twitter". *American Speech*, 90(4):403–440, 2015.

[271] Taylor Jones. 'Eem' Negation in African American English: A Next Step in Jespersen's Cycle? *University of Pennsylvania Working Papers in Linguistics*, 22(1), 2016.

[272] Taylor Jones. *Variation in African American English: The Great Migration and Regional Differentiation*. PhD thesis, University of Pennsylvania, 2020.

[273] Taylor Jones, Jessica Rose Kalbfeld, Ryan Hancock, and Robin Clark. Testifying while black: An experimental study of court reporter accuracy in transcription of African American English. *Language*, 95(2), 2019.

[274] Anna Jørgensen, Dirk Hovy, and Anders Søgaard. Challenges of studying and processing dialects in social media. In *Proceedings of the Workshop on Noisy User-Generated Text*, pages 9–18, Beijing, China, 2015.

[275] Anna Jørgensen, Dirk Hovy, and Anders Søgaard. Learning a POS tagger for AAVE-like language. In *Proceedings of the North American Association for Computational Linguistics (NAACL)*, pages 1115–1120, San Diego, CA, 2016.

[276] Pratik Joshi, Sebastian Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the Association for Computational Linguistics (ACL)*, page 6282–6293, 2020.

[277] Jaap Jumelet, Willem Zuidema, and Dieuwke Hupkes. Analysing Neural Language Models: Contextual Decomposition Reveals Default Reasoning in Number and Gender Assignment. In *Proceedings of the Conference on Natural Language Learning*, Hong Kong, China, 2019.

[278] Marie-Odile Junker. Participatory action research for Indigenous linguistics in the digital age. In Shannon T. Bischoff and Carmen Jany, editors, *Insights from Practices in Community-Based Research*, pages 164–175. De Gruyter Mouton, 2018.

[279] David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. Incorporating Dialectal Variability for Socially Equitable Language Identification. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 51–57, Vancouver, Canada, 2017.

[280] David Jurgens, Libby Hemphill, and Eshwar Chandrasekharan. A Just and Comprehensive Strategy for Using NLP to Address Online Abuse. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 3658–3666, Florence, Italy, 2019.

[281] Masahiro Kaneko and Danushka Bollegala. Gender-preserving debiasing for pre-trained word embeddings. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 1641–1650, Florence, Italy, 2019.

[282] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. *arXiv preprint arXiv:2010.04050*, 2020.

[283] Saket Karve, Lyle Ungar, and João Sedoc. Conceptor debiasing of word representations evaluated on WEAT. In *Proceedings of the Workshop on Gender Bias in Natural Language Processing*, pages 40–48, Florence, Italy, 2019.

[284] Maximilian Kasy and Rediet Abebe. Fairness, Equality, and Power in Algorithmic Decision-Making. In *Proceedings of the Participatory Approaches to Machine Learning Workshop*, 2020.

[285] Michael Katell, Meg Young, Dharma Dailey, Bernease Herman, Vivian Guetler, Aaron Tam, Corinne Bintz, Danielle Raz, and P.M. Krafft. Toward situated interventions for algorithmic equity: lessons from the field. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Barcelona, Spain, 2020.

[286] Stephen Kemmis. Participatory action research and the public sphere. *Educational Action Research*, 14(4):459–476, 2006.

[287] Renee Kemp, Emily Moline, Chelsea Escalante, Alexander Mendes, and Robert Bayley. Where Have All the Participles Went? Using Twitter Data to Teach About Language. *American Speech*, 91(2):226–235, 2016.

[288] Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. Contextualizing Hate Speech Classifiers with Post-hoc Explanation. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 5435–5442, 2020.

[289] Os Keyes. The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW), 2018.

[290] Os Keyes, Josephine Hoy, and Margaret Drouhard. Human-Computer Insurrection: Notes on an Anarchist HCI. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, Glasgow, Scotland, UK, 2019.

[291] Mohammad Khan, Markus Dickinson, and Sandra Kübler. Towards Domain Adaptation for Parsing Web Data. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 357–364, Hissar, Bulgaria, 2013.

[292] Jae Yeon Kim, Carlos Ortiz, Sarah Nam, Sarah Santiago, and Vivek Datta. Intersectional Bias in Hate Speech and Abusive Language Datasets. In *Proceedings of the ICWSM Data Challenge Workshop*, 2020.

[293] Ben King and Steven P. Abney. Labeling the Languages of Words in Mixed-Language Documents using Weakly Supervised Methods. In *Proceedings of the North American Association for Computational Linguistics (NAACL)*, pages 1110–1119, Atlanta, GA, 2013.

[294] Sharese King. From African American Vernacular English to African American Language: Rethinking the Study of Race and Language in African Americans' Speech. *Annual Review of Linguistics*, 6(1):285–300, 2020.

[295] Svetlana Kiritchenko and Saif M. Mohammad. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In *Proceedings of the Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, LA, 2018.

[296] Markus Knoche, Radomir Popović, Florian Lemmerich, and Markus Strohmaier. Identifying Biases in Politically Biased Wikis through Word Embeddings. In *Proceedings of the Conference on Hypertext and Social Media*, Berlin, Germany, 2019.

[297] Vid Kocijan, Oana-Maria Camburu, and Thomas Lukasiewicz. The Gap on GAP: Tackling the Problem of Differing Data Distributions in Bias-Measuring Datasets. In *Proceedings of the Workshop on Algorithmic Fairness through the Lens of Causality and Interpretability*, 2020.

[298] Tom Kocmi, Tomasz Limisiewicz, and Gabriel Stanovsky. Gender Coreference and Bias Evaluation at WMT 2020. In *Proceedings of the Conference on Machine Translation (WMT)*, 2020.

[299] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689, 2020.

[300] Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A. Smith. A Dependency Parser for Tweets. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 1001–1012, Doha, Qatar, 2014.

[301] Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Computing*, 17(4):401–412, 2002.

[302] Hadas Kotek, Rikker Dockum, Sarah Babinski, and Christopher Geissler. Gender bias and stereotypes in linguistic example sentences. Manuscript at `https://ling.auf.net/lingbuzz/005367`.

[303] Austin C. Kozlowski, Matt Taddy, and James A. Evans. The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings. *American Sociological Review*, 84(5):905–949, 2019.

[304] Philipp Krämer and Magdalena von Sicard. "A picturesque language, humorous and very pliable": Descriptions of French-based Creole languages in travel guidebooks. *Journal of Postcolonial Linguistics*, 3:64–88, 2020.

[305] Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. GeDi: Generative Discriminator Guided Sequence Generation. *arXiv preprint arXiv:2009.06367*, 2020.

[306] Paul Kroskrity. Language Ideologies. In Alessandro Duranti, editor, *Companion to Linguistic Anthropology*, pages 496–517. Basil Blackwell, 2004.

[307] Vaibhav Kumar, Tenzin Singhay Bhotia, and Vaibhav Kumar. Fair Embedding Engine: A Library for Analyzing and Mitigating Gender Bias in Word Embeddings. In *Proceedings of the Workshop for NLP Open Source Software*, 2020.

[308] Vaibhav Kumar, Tenzin Singhay Bhotia, Vaibhav Kumar, and Tanmoy Chakraborty. Nurse is Closer to Woman than Surgeon? Mitigating Gender-Biased Proximities in Word Embeddings. *Transactions of the Association for Computational Linguistics*, 8, 2020.

[309] Benedek Kurdi, Thomas C. Mann, Tessa E.S. Charlesworth, and Mahzarin R. Banaji. The relationship between implicit intergroup attitudes and beliefs. *Proceedings of the National Academy of Sciences*, 116(13):5862–5871, 2019.

[310] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W. Black, and Yulia Tsvetkov. Measuring bias in contextualized word representations. In *Proceedings of the Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy, 2019.

[311] Carmen Kynard. *Vernacular Insurrections: Race, Black Protest, and the New Century in Composition-Literacies Studies*. SUNY Press, 2013.

[312] William Labov. *Language in the Inner City: Studies in the Black English Vernacular*. University of Pennsylvania Press, 1972.

[313] Sonja Lanehart, Jennifer Bloomquist, and Ayesha M. Malik. Language Use in African American Communities: An Introduction. In Sonja L. Lanehart, editor, *The Oxford Handbook of African American Language*, pages 1–19. Oxford University Press, 2015.

[314] Sonja L. Lanehart. African American Language Identities: Contradictions and Conundrums. In Sonja L. Lanehart, editor, *The Oxford Handbook of African American Language*, pages 863–879. Oxford University Press, 2015.

[315] Sonja L. Lanehart and Ayesha M. Malik. Black Is, Black Isn't: Perceptions of Language and Blackness. In Jeffrey Reaser, Eric Wilbanks, Karissa Wojcik, and Walt Wolfram, editors, *Language Variety in the New South*. University of North Carolina Press, 2018.

[316] Brian N. Larson. Gender as a variable in natural-language processing: Ethical considerations. In *Proceedings of the Workshop on Ethics in Natural Language Processing*, pages 30–40, Valencia, Spain, 2017.

[317] Anne Lauscher and Goran Glavaš. Are We Consistently Biased? Multidimensional Analysis of Biases in Distributional Word Vectors. In *Proceedings of the Joint Conference on Lexical and Computational Semantics*, pages 85–91, Minneapolis, MN, 2019.

[318] Anne Lauscher, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. A General Framework for Implicit and Explicit Debiasing of Distributional Word Vector Spaces. *arXiv preprint arXiv:1909.06092*, 2019.

[319] Christopher A. Le Dantec, Erika Shehan Poole, and Susan P. Wyche. Values as Lived Experience: Evolving Value Sensitive Design in Support of Value Discovery. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, Boston, MA, 2009.

[320] Colin Wayne Leach, Naomi Ellemers, and Manuela Barreto. Group Virtue: The Importance of Morality (vs. Competence and Sociability) in the Positive Evaluation of In-Groups. *Journal of Personality and Social Psychology*, 93(2): 234–249, 2007.

[321] Nayeon Lee, Andrea Madotto, and Pascale Fung. Exploring Social Bias in Chatbots using Stereotype Knowledge. In *Proceedings of the Workshop on Widening NLP*, pages 177–180, Florence, Italy, 2019.

[322] Maider Lehr, Kyle Gorman, and Izhak Shafran. Discriminative pronunciation modeling for dialectal speech recognition. In *Proceedings of INTERSPEECH*, pages 1458–1462, Singapore, 2014.

[323] Kobi Leins, Jey Han Lau, and Timothy Baldwin. Give Me Convenience and Give Her Death: Who Should Decide What Uses of NLP are Appropriate, and on What Basis? In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 2908–2913, 2020.

[324] Alison P. Lenton, Constantine Sedikides, and Martin Bruder. A latent semantic analysis of gender stereotype-consistency and narrowness in American English. *Sex Roles*, 60:269–278, 2009.

[325] Wesley Y. Leonard. Challenging "Extinction" through Modern Miami Language Practices. *American Indian Culture and Research Journal*, 35(2):135–160, 2011.

[326] Wesley Y. Leonard. Reframing language reclamation programmes for everybody's empowerment. *Gender and Language*, 6(2):339–367, 2012.

[327] Wesley Y. Leonard. Producing language reclamation by decolonising 'language'. *Language Documentation and Description*, 14:15–36, 2017.

[328] Michael A. Lepori. Unequal Representations: Analyzing Intersectional Biases in Word Embeddings Using Representational Similarity Analysis. *arXiv preprint arXiv:2011.12086*, 2020.

[329] Julia C. Leschke and Carsten Schwemmer. Media bias towards African-Americans before and after the Charlottesville rally. In *Proceedings of the Weizenbaum Conference "Challenges of Digital Inequality – Digital Education, Digital Work, Digital Life"*, Berlin, Germany, 2019.

[330] Shiri Levi-Ari and Boaz Keysar. Why don't we believe non-native speakers? The influence of accent on credibility. *Journal of Experimental Social Psychology*, 46:1093–1096, 2010.

[331] Carsten Levisen and Eeva Sippola. Postcolonial Linguistics: The Editor's Guide to a New Interdiscipline. *Journal of Postcolonial Linguistics*, 1:1–15, 2019.

[332] Jason Edward Lewis, editor. *Indigenous Protocol and Artificial Intelligence Position Paper*, Honolulu, Hawai'i, 2020. The Initiative for Indigenous Futures and the Canadian Institute for Advanced Research (CIFAR).

[333] Molly Lewis and Gary Lupyan. Language use shapes cultural norms: Large scale evidence from gender. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, Austin, TX, 2018.

[334] Molly Lewis and Gary Lupyan. What are we learning from language? Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature Human Behavior*, 2020.

[335] Juncen Li, Robin Jia, He He, and Percy Liang. Delete, Retrieve, Generate: A Simple Approach to Sentiment and Style Transfer. In *Proceedings of the North American Association for Computational Linguistics (NAACL)*, pages 1865–1874, New Orleans, LA, 2018.

[336] Tao Li, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Vivek Srikumar. UNQOVERing Stereotyping Biases via Underspecified Questions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3475–3489, 2020.

[337] Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. Towards Debiasing Sentence Representations. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 5502–5515, 2020.

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

[339] Haochen Liu, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu, and Jiliang Tang. Does Gender Matter? Towards Fairness in Dialogue Systems. *arXiv preprint arXiv:1910.10486*, 2019. Published in 2020 at the International Conference on Computational Linguistics (COLING).

[340] Haochen Liu, Wentao Wang, Yiqi Wang, Hui Liu, Zitao Liu, and Jiliang Tang. Mitigating Gender Bias for Neural Dialogue Generation with Adversarial Learning. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 893–903, 2020.

[341] Yijia Liu, Yi Zhu, Wanxiang Che, Bing Qin, Nathan Schneider, and Noah A. Smith. Parsing Tweets into Universal Dependencies. In *Proceedings of the North American Association for Computational Linguistics (NAACL)*, pages 965–975, New Orleans, LA, 2018.

[342] Emma Llansó, Natasha Duarte, Mana Azarmi, and Alicia Loh. Center for Democracy & Technology Comments to the U.S. State Department on Proposed Collection of Visa Applicants' Social Media Information, 2018.

[343] Brandon C. Loudermilk. Implicit attitudes and the perception of sociolinguistic variation. In Alexei Prikhodkine and Dennis R. Preston, editors, *Responses to Language Varieties: Variability, processes and outcomes*, pages 137–156. John Benjamins Publishing Company, 2015.

[344] Anastassia Loukina, Nitin Madnani, and Klaus Zechner. The many dimensions of algorithmic fairness in educational applications. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–10, Florence, Italy, 2019.

[345] Kaiji Lu, Peter Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. Gender bias in neural natural language processing. *arXiv preprint arXiv:1807.11714*, 2018.

[346] Marco Lui and Timothy Baldwin. langid.py: An Off-the-shelf Language Identification Tool. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 25–30, Jeju Island, Korea, 2012.

[347] Brandon Lwowski and Anthony Rios. The Risk of Racial Bias while Tracking Influenza-Related Content on Social Media using Machine Learning. Under review.

[348] Teresa Lynn, Kevin Scannell, and Eimear Maguire. Minority Language Twitter: Part-of-Speech Tagging and Analysis of Irish Tweets. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 1–8, Beijing, China, 2015.

[349] Xinyao Ma, Maarten Sap, Hannah Rashkin, and Yejin Choi. PowerTransformer: Unsupervised Controllable Revision for Biased Language Correction. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 7426–7441, 2020.

[350] Anne Maass. Linguistic intergroup bias: Stereotype perpetuation through language. *Advances in Experimental Social Psychology*, 31:79–121, 1999.

[351] C. Neil Macrae, Charles Stangor, and Miles Hewstone, editors. *Stereotypes and stereotyping.* Guilford Press, 1996.

[352] Nishtha Madaan, Sameep Mehta, Taneea S. Agrawaal, Vrinda Malhotra, Aditi Aggarwal, Yatin Gupta, and Mayank Saxena. Analyze, Detect and Remove Gender Stereotyping from Bollywood Movies. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, New York, NY, 2018.

[353] Nitin Madnani, Anastassia Loukina, Alina von Davier, Jill Burstein, and Aoife Cahill. Building Better Open-Source Tools to Support Fairness in Automated Scoring. In *Proceedings of the Workshop on Ethics in Natural Language Processing*, pages 41–52, Valencia, Spain, 2017.

[354] Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. End-to-End Bias Mitigation by Modelling Biases in Corpora. In *Proceedings of the Association for Computational Linguistics (ACL)*, page 8706–8716, 2020.

[355] Suraj Maharjan, Elizabeth Blair, Steven Bethard, and Thamar Solorio. Developing language-tagged corpora for code-switching tweets. In *Proceedings of the Linguistic Annotation Workshop*, 2015.

[356] Sinfree Makoni, Geneva Smitherman, Arnetha F. Ball, and Arthur K. Spears, editors. *Black Linguistics: Language, Society, and Politics in Africa and the Americas*. Routledge, 2003.

[357] Shervin Malmasi, Eshrag Refaee, and Mark Dras. Arabic dialect identification using a parallel multidialectal corpus. In *Proceedings of the Conference of the Pacific Association for Computational Linguistics (PACLING)*, Singapore, Singapore, 2015.

[358] Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and Alan W. Black. Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings. In *Proceedings of the North American Association for Computational Linguistics (NAACL)*, pages 801–809, Minneapolis, MN, 2019.

[359] Ramón Antonio Martínez and Alexander Feliciano Mejía. Looking closely and listening carefully: A sociocultural approach to understanding the complexity of Latina/o/x students' everyday language. *Theory Into Practice*, 59(1):53–63, 2020.

[360] Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. On Measuring Social Biases in Sentence Encoders. In *Proceedings of the North American Association for Computational Linguistics (NAACL)*, pages 629–634, Minneapolis, MN, 2019.

[361] Elijah Mayfield, Michael Madaio, Shrimai Prabhumoye, David Gerritsen, Brittany McLaughlin, Ezekiel Dixon-Roman, and Alan W. Black. Equity Beyond Bias in Language Technologies for Education. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, Florence, Italy, 2019.

[362] Emily McBain-Ashfield and Jason Millar. Off-The-Shelf Sentiment Analysis: The Ethics of Subjectively Annotated Data and Objectively Measured AI Models. In *WeRobot*, 2020.

[363] Carole McCann and Seung-Kyung Kim. *Feminist Theory Reader: Local and Global Perspectives*. Routledge, 2013.

[364] Teresa L. McCarty and Lucille J. Watahomigie. Indigenous Community-based Language Education in the USA. *Language, Culture and Curriculum*, 11(3): 309–324, 1998.

[365] Katherine McCurdy and Oğuz Serbetçi. Grammatical gender associations outweigh topical gender bias in crosslinguistic word embeddings. In *Proceedings of the Workshop for Women & Underrepresented Minorities in Natural Language Processing*, Vancouver, Canada, 2017.

[366] Kris McGuffie and Alex Newhouse. The Radicalization Risks of GPT-3 and Advanced Neural Language Models. Technical report, Middlebury Institute of International Studies at Monterey: Center on Terrorism, Extreme, and Counterterrorism, 2020.

[367] Ninareh Mehrabi, Thamme Gowda, Fred Morstatter, Nanyun Peng, and Aram Galstyan. Man is to Person as Woman is to Location: Measuring Gender Bias in Named Entity Recognition. *arXiv preprint arXiv:1910.10872*, 2019.

[368] Michela Menegatti and Monica Rubini. Gender bias and sexism in language. In *Oxford Research Encyclopedia of Communication*. Oxford University Press, 2017.

[369] Shelbi Nahwilet Messner. The moral fabric of linguicide: un-weaving trauma narratives and dependency relationships in Indigenous language reclamation. *Journal of Global Ethics*, 14(2):266–276, 2018.

[370] Miriam Meyerhoff. *Introducing Sociolinguistics*. Routledge, 3rd edition, 2019.

[371] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[372] James Milroy and Leslie Milroy. *Authority in Language: Investigating Standard English*. Routledge, 1999.

[373] David Mimno and Andrew McCallum. Topic models conditioned on arbitrary features with Dirichlet-Multinomial regression. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, Helsinki, Finland, 2008.

[374] David Mimno and Laure Thompson. The strange geometry of skip-gram with negative sampling. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 2873–2878, Copenhagen, Denmark, 2017.

[375] Inom Mirzaev, Anthony Schulte, Michael Conover, and Sam Shah. Considerations for the interpretation of bias measures of word embeddings. *arXiv preprint arXiv:1906.08379*, 2019.

[376] Arul Mishra, Himanshu Mishra, and Shelly Rathee. Examining the presence of gender bias in customer reviews using word embedding. *arXiv preprint arXiv:1902.00496*, 2019.

[377] Shubhanshu Mishra, Sijun He, and Luca Belli. Assessing Demographic Bias in Named Entity Recognition. In *Proceedings of the Bias in Automatic Knowledge Graph Construction Workshop*, 2020.

[378] Shakir Mohamed, Marie-Therese Png, and William Isaac. Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. *Philosophy & Technology*, 2020.

[379] Robert (Munro) Monarch and Alex (Carmen) Morrison. Detecting Independent Pronoun Bias with Partially-Synthetic Data Generation. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 2011–2017, 2020.

[380] Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict. *Political Analysis*, 16(4):372–403, 2008.

[381] Emma Moore. Sociolinguistic Style: A Multidimensional Resource for Shared Identity Creation. *Canadian Journal of Linguistics*, 49(3/4):375–396, 2004.

[382] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen Carley. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. In *Proceedings of the Conference on Web and Social Media (ICWSM)*, Cambridge, MA, 2013.

[383] Suhanthie Motha. Is an Antiracist and Decolonizing Applied Linguistics Possible? *Annual Review of Applied Linguistics*, 40:128–133, 2020.

[384] Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. Hate Speech Detection and Racial Bias Mitigation in Social Media based on BERT model. *PLoS One*, 2020.

[385] Salikoko S. Mufwene, Guy Bailey, and John R. Rickford, editors. *African-American English: Structure, History, and Use*. Routledge, 1998.

[386] Michael J. Muller. Participatory Design: The Third Space in HCI. In *The Human-Computer Interaction Handbook*, pages 1087–1108. CRC Press, 2007.

[387] Rodrigo Alejandro Chávez Mulsa and Gerasimos Spanakis. Evaluating Bias In Dutch Word Embeddings. In *Proceedings of the Workshop on Gender Bias in Natural Language Processing*, 2020.

[388] Moin Nadeem, Anna Bethke, and Siva Reddy. StereoSet: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*, 2020.

[389] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, 2020.

[390] Kate Nash and Vikki Bell. The Politics of Framing: An Interview with Nancy Fraser. *Theory, Culture & Society*, 24(4):73–86, 2007.

[391] Alexis Nasr, Geraldine Damnati, Aleksandra Guerraz, and Frederic Bechet. Syntactic parsing of chat language in contact center conversation corpus. In *Proceedings of the Annual SIGdial Meeting on Discourse and Dialogue*, pages 175–184, Los Angeles, CA, 2016.

[392] Gina Neff and Peter Nagy. Talking to Bots: Symbiotic Agency and the Case of Tay. *International Journal of Communication*, 10:4915–4931, 2016.

[393] Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Tajudeen Kolawole, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddee Hassan Muhammad, Salomon Kabongo, Salomey Osei, Sackey Freshia, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa, Mofe Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Jane Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkabir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Espoir Murhabazi, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Emezue, Bonaventure Dossou, Blessing Sibanda, Blessing Itoro Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, 2020.

[394] Peter Newmark. *Approaches to Translation*. Pergamon Press, 1981.

[395] Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. "How Old Do You Think I Am?": A Study of Language and Age in Twitter. In *Proceedings of the Conference on Web and Social Media (ICWSM)*, pages 439–448, Boston, MA, 2013.

[396] Tonya Nguyen and Niloufar Salehi. Implications of Conversational Artificial Intelligence. In *Fair & Responsible AI Workshop*, 2020.

[397] Malvina Nissim, Rik van Noord, and Rob van der Goot. Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, 2020.

[398] Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 1659–1666, Portorož, Slovenia, 2016.

[399] Safiya Umoja Noble. *Algorithms of Oppression: How Search Engines Reinforce Racism.* NYU Press, 2018.

[400] Debora Nozza, Claudia Volpetti, and Elisabetta Fersini. Unintended Bias in Misogyny Detection. In *Proceedings of the Conference on Web Intelligence*, pages 149–155, 2019.

[401] Brendan O'Connor, Jacob Eisenstein, Eric P. Xing, and Noah A. Smith. A mixture model of demographic lexical variation. In *NIPS Workshop on Machine Learning for Social Computing*, 2010.

[402] Brendan O'Connor, Michael Krieger, and David Ahn. TweetMotif: Exploratory Search and Topic Summarization for Twitter. In *Proceedings of the Conference on Web and Social Media (ICWSM)*, Washington, D.C., 2010.

[403] Ihudiya Finda Ogbonnaya-Ogburu, Angela D. R. Smith, Alexandra To, and Kentaro Toyama. Critical Race Theory for HCI. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 1–16, Honolulu, HI, 2020.

[404] Alexandra Olteanu, Kartik Talamadupula, and Kush R. Varshney. The Limits of Abstract Evaluation Metrics: The Case of Hate Speech Detection. In *Proceedings of the ACM Web Science Conference*, Troy, NY, 2017.

[405] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Frontiers in Big Data*, 2, 2019.

[406] Conference on College Composition and Communication. Students' Right to Their Own Language, 1974. `https://secure.ncte.org/library/NCTEFiles/Groups/CCCC/NewSRTOL.pdf`.

[407] Conference on College Composition and Communication. This Ain't Another Statement! This is a DEMAND for Black Linguistic Justice!, 2020. `https://cccc.ncte.org/cccc/demand-for-black-linguistic-justice`.

[408] Mark Orbe. #alllivesmatter as post-racial rhetorical strategy. *Journal of Contemporary Rhetoric*, 5, 2015.

[409] Arianna Ornaghi, Elliott Ash, and Daniel L. Chen. Stereotypes in High Stakes Decisions: Evidence from U.S. Circuit Courts, 2019. Working paper.

[410] Jahna Otterbacher, Jo Bates, and Paul Clough. Competent Men and Warm Women: Gender Stereotypes and Backlash in Image Search Results. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 6620–6631, Denver, CO, 2017.

[411] Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the North American Association for Computational Linguistics (NAACL)*, pages 380–390, Atlanta, GA, 2013.

[412] Orestis Papakyriakopoulos, Simon Hegelich, Juan Carlos Medina Serrano, and Fabienne Marco. Bias in word embeddings. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Barcelona, Spain, 2020.

[413] Ji Ho Park, Jamin Shin, and Pascale Fung. Reducing Gender Bias in Abusive Language Detection. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 2799–2804, Brussels, Belgium, 2018.

[414] Aarsh Patel, Rahul Gupta, Mukund Sridhar, Satyapriya Krishna, Alok Aman, and Peng Liu. Towards Classification Parity Across Cohorts. In *Proceedings of the Machine Learning In Real Life Workshop*, 2020.

[415] Faiza Patel, Rachel Levinson-Waldman, Sophia DenUyl, and Raya Koreh. Social Media Monitoring: How the Department of Homeland Security Uses Digital Data in the Name of National Security. Brennan Center for Justice, 2019.

[416] Umashanthi Pavalanathan and Jacob Eisenstein. Confounds and consequences in geotagged twitter data. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 2138–2148, Lisbon, Portugal, 2015.

[417] Ellie Pavlick and Tom Kwiatkowski. Inherent Disagreements in Human Textual Inferences. *Transactions of the Association for Computational Linguistics*, 7: 677–694, 2019.

[418] Xiangyu Peng, Siyan Li, Spencer Frazier, and Mark Riedl. Fine-Tuning a Transformer-Based Language Model to Avoid Generating Non-Normative Text. *arXiv preprint arXiv:2001.08764*, 2020.

[419] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global Vectors for Word Representation. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, 2014.

[420] Carla Pérez-Almendros, Luis Espinosa-Anke, and Steven Schockaert. Don't Patronize Me! An Annotated Dataset with Patronizing and Condescending Language towards Vulnerable Communities. *arXiv preprint arXiv:2011.08320*, 2020.

[421] Barbara Plank. A comparison of structural correspondence learning and self-training for discriminative parse selection. In *Proceedings of the Workshop on Semi-supervised Learning for NLP*, 2009.

[422] Radomir Popović, Florian Lemmerich, and Markus Strohmaier. Joint Multiclass Debiasing of Word Embeddings. In *Proceedings of the International Symposium on Intelligent Systems*, Graz, Austria, 2020.

[423] Vinodkumar Prabhakaran, Emily E. Reid, and Owen Rambow. Gender and Power: How Gender and Gender Environment Affect Manifestations of Power. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 1965–1976, Doha, Qatar, 2014.

[424] Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. Perturbation Sensitivity Analysis to Detect Unintended Model Biases. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 5744–5749, Hong Kong, China, 2019.

[425] Vinay Uday Prabhu and Abeba Birhane. Large image datasets: A pyrrhic win for computer vision. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, 2020.

[426] Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W. Black. Style Transfer Through Back-Translation. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 866–876, Melbourne, Australia, 2018.

[427] Shrimai Prabhumoye, Elijah Mayfield, and Alan W. Black. Principled Frameworks for Evaluating Ethics in NLP Systems. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, Florence, Italy, 2019.

[428] Shrimai Prabhumoye, Brendon Boldt, Ruslan Salakhutdinov, and Alan W. Black. Case Study: Deontological Ethics in NLP. *arXiv preprint arXiv:2010.04658*, 2020.

[429] Marcelo Prates, Pedro Avelar, and Luis C. Lamb. Assessing gender bias in machine translation: A case study with google translate. *Neural Computing and Applications*, 2019.

[430] Rasmus Précenth. Word embeddings and gender stereotypes in Swedish and English. Master's thesis, Uppsala University, 2019.

[431] Dennis R. Preston. Are you really smart (or stupid, or cute, or ugly, or cool)? Or do you just talk that way? *Language attitudes, standardization and language change. Oslo: Novus forlag*, pages 105–129, 2009.

[432] Zoya Proshina. *Theory of Translation (English and Russian)*. Far Eastern University Press, 3rd edition, 2008.

[433] Flavien Prost, Nithum Thain, and Tolga Bolukbasi. Debiasing Embeddings for Reduced Gender Bias in Text Classification. In *Proceedings of the Workshop on Gender Bias in Natural Language Processing*, pages 69–75, Florence, Italy, 2019.

[434] Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C. Lipton. Learning to Deceive with Attention-Based Explanations. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 4782–4793, 2020.

[435] Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. Automatically Neutralizing Subjective Bias in Text. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, New York, NY, 2020.

[436] Arun K. Pujari, Ansh Mittal, Anshuman Padhi, Anshul Jain, Mukesh Jadon, and Vikas Kumar. Debiasing Gender biased Hindi Words with Word-embedding. In *Proceedings of the International Conference on Algorithms, Computing and Artificial Intelligence*, pages 450–456, 2019.

[437] Thomas Purnell, William Idsardi, and John Baugh. Perceptual and Phonetic Experiments on American English Dialect Identification. *Journal of Language and Social Psychology*, 18(1):10–30, 1999.

[438] Yusu Qian. Gender Stereotypes Differ between Male and Female Writings. In *Proceedings of the ACL Student Research Workshop*, pages 48–53, Florence, Italy, 2019. Research proposal.

[439] Yusu Qian, Urwa Muaz, Ben Zhang, and Jae Won Hyun. Reducing gender bias in word-level language models with a gender-equalizing loss function. In *Proceedings of the ACL Student Research Workshop*, pages 223–228, Florence, Italy, 2019.

[440] Kevin M. Quinn, Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, 54(1):209–228, 2010.

[441] Elaheh Raisi and Bert Huang. Reduced-Bias Co-trained Ensembles for Weakly Supervised Cyberbullying Detection. In *Proceedings of the International Conference on Computational Data and Social Networks*, pages 293–306, 2019.

[442] Daniel Ramage, Christopher D. Manning, and Susan Dumais. Partially labeled topic models for interpretable text mining. In *Proceedings of the Conference on Knowledge Discovery and Data Mining (KDD)*, San Diego, CA, 2011.

[443] Sree Harsha Ramesh and Krishna Prasad Sankaranarayanan. Neural Machine Translation for Low Resource Languages using Bilingual Lexicon Induced from Comparable Corpora. In *Proceedings of the NAACL Student Research Workshop*, pages 112–119, New Orleans, LA, 2018.

[444] Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. In *Proceedings of the Association for Computational Linguistics (ACL)*, page 7237–7256, 2020.

[445] Georg Rehm and Hans Uszkoreit, editors. *The Norwegian Language in the Digital Age*. Springer, 2012.

[446] Elizabeth Reichert, Helen Qiu, and Jasmine Bayrooti. Reading Between the Demographic Lines: Resolving Sources of Bias in Toxicity Classifiers. *arXiv preprint arXiv:2006.16402*, 2020.

[447] Angela Reyes. Appropriation of African American slang by Asian American youth. *Journal of Sociolinguistics*, 9(4):509–532, 2005.

[448] Douglas Rice, Jesse H. Rhodes, and Tatishe Nteta. Racial bias in legal language. *Research & Politics*, 6(2), 2019.

[449] Rashida Richardson, Jason Schultz, and Kate Crawford. Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice. *NYUL Rev. Online*, 94, 2019.

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

[451] John R. Rickford and Sharese King. Language and linguistics on trial: Hearing Rachel Jeantel (and other vernacular speakers) in the courtroom and beyond. *Language*, 92(4):948–988, 2016.

[452] John Russell Rickford and Russell John Rickford. *Spoken Soul: The Story of Black English*. John Wiley & Sons, 2000.

[453] Anthony Rios. FuzzE: Fuzzy Fairness Evaluation of Offensive Language Classifiers on African-American English. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, New York, NY, 2020.

[454] Margaret Et. Roberts, Brandon M. Stewart, Dustin Tingley, and Edoardo M. Airoldi. The structural topic model and applied social science. In *Proceedings of the NIPS Workshop on Topic Models: Computation, Application, and Evaluation*, 2013.

[455] Mary Robinson and Daniel Duncan. Holistic approaches to syntactic variation: Wh-all questions in English. In *Penn Linguistics Conference*, volume 42, 2018.

[456] Gerald Roche. Articulating language oppression: colonialism, coloniality and the erasure of Tibet's minority languages. *Patterns of Prejudice*, 53(5):487–514, 2019.

[457] Alexey Romanov, Maria De-Arteaga, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, Anna Rumshisky, and Adam Tauman Kalai. What's in a Name? Reducing Bias in Bios without Access to Protected Attributes. In *Proceedings of the North American Association for Computational Linguistics (NAACL)*, pages 4187–4195, Minneapolis, MN, 2019.

[458] Jonathan Rosa. Contesting Representations of Migrant "Illegality" through the Drop the I-Word Campaign: Rethinking Language Change and Social Change. In Netta Avineri, Laura R. Graham, Eric J. Johnson, Robin Conley Riner, and Jonathan Rosa, editors, *Language and Social Justice in Practice*. Routledge, 2019.

[459] Jonathan Rosa and Christa Burdick. Language Ideologies. In Ofelia García, Nelson Flores, and Massimiliano Spotti, editors, *The Oxford Handbook of Language and Society*. Oxford University Press, 2017.

[460] Jonathan Rosa and Nelson Flores. Unsettling race and language: Toward a raciolinguistic perspective. *Language in Society*, 46:621–647, 2017.

[461] Jonathan Daniel Rosa. Standardization, Racialization, Languagelessness: Raciolinguistic Ideologies across Communicative Contexts. *Journal of Linguistic Anthropology*, 26(2):162–183, 2016.

[462] Sara Rosenthal and Kathleen McKeown. Age Prediction in Blogs: A Study of Style, Content, and Online Behavior in Pre- and Post-Social Media Generations. In *Proceedings of the North American Association for Computational Linguistics (NAACL)*, pages 763–772, Portland, OR, 2011.

[463] Mike Rosner and Paulseph-John Farrugia. A tagging algorithm for mixed language identification in a noisy domain. In *Proceedings of the Conference of the International Speech Communication Association*, San Diego, CA, 2007.

[464] Candace Ross, Boris Katz, and Andrei Barbu. Measuring Social Biases in Grounded Vision and Language Embeddings. *arXiv preprint arXiv:2002.08911*, 2020.

[465] David A. Ross, Dowin Boatright, Marcella Nunez-Smith, Ayana Jordan, Adam Chekroud, and Edward Z. Moore. Differences in words used to describe racial and gender groups in Medical Student Performance Evaluations. *PLoS One*, 12 (8), 2017.

[466] Richard Rothstein. *The Color of Law: A Forgotten History of How Our Government Segregated America*. Liveright Publishing, 2017.

[467] Ryan Rowe, German Creamer, Shlomo Hershkop, and Salvatore J. Stolfo. Automated social hierarchy detection through email network analysis. In *Proceedings of the 9th WebKDD and 1st SNA-KDD Workshop on Web Mining and Social Network Analysis*, pages 109–117, 2007.

[468] David Rozado. Wide range screening of algorithmic bias in word embedding models using large sentiment lexicons reveals underreported bias types. *PLOS One*, 2020.

[469] Elayne Ruane, Abeba Birhane, and Anthony Ventresque. Conversational AI: Social and Ethical Considerations. In *Proceedings of the Irish Conference on Artificial Intelligence and Cognitive Science*, Galway, Ireland, 2019.

[470] Donald Rubin. The power of prejudice in accent perception: Reverse linguistic stereotyping and its impact on listener judgments and decisions. In John Levis and Kimberly LeVelle, editors, *Proceedings of the 3rd Pronunciation in Second Language Learning and Teaching Conference*, pages 11–17, Ames, IA, 2012.

[471] Rachel Rudinger, Chandler May, and Benjamin Van Durme. Social bias in elicited natural language inferences. In *Proceedings of the Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, Spain, 2017.

[472] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender Bias in Coreference Resolution. In *Proceedings of the North American Association for Computational Linguistics (NAACL)*, pages 8–14, New Orleans, LA, 2018.

[473] Magnus Sahlgren and Fredrik Olsson. Gender Bias in Pretrained Swedish Embeddings. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 35–43, Turku, Finland, 2019.

[474] Elizabeth B.N. Sanders. From user-centered to participatory design approaches. In Jorge Frascara, editor, *Design and the Social Sciences: Making Connections*, pages 18–25. CRC Press, 2002.

[475] Manuela Sanguinetti, Cristina Bosco, Alessandro Mazzei, Alberto Lavelli, and Fabio Tamburini. Annotating Italian Social Media Texts in Universal Dependencies. In *Proceedings of the International Conference on Dependency Linguistics (Depling)*, Pisa, Italy, 2017.

[476] Brenda Salenave Santana, Vinicius Woloszyn, and Leandro Krug Wives. Is there gender bias and stereotype in Portuguese word embeddings? In *Proceedings of the International Conference on the Computational Processing of Portuguese Student Research Workshop*, Canela, Brazil, 2018.

[477] Mauricio Santillana, Andre T. Nguyen, Mark Dredze, Michael J. Paul, Elaine Nsoesie, and John S. Brownstein. Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance. *PLOS Computational Biology*, 2015.

[478] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. The risk of racial bias in hate speech detection. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 1668–1678, Florence, Italy, 2019.

[479] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. Social Bias Frames: Reasoning about Social and Power Implications of Language. In *Proceedings of the Association for Computational Linguistics (ACL)*, page 5477–5490, 2020.

[480] Hanna Sassaman, Jennifer Lee, Jenessa Irvine, and Shankar Narayan. Creating Community-Based Tech Policy: Case Studies, Lessons Learned, and What Technologists and Communities Can Do Together. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Barcelona, Spain, 2020.

[481] Danielle Saunders and Bill Byrne. Reducing Gender Bias in Neural Machine Translation as a Domain Adaptation Problem. In *Proceedings of the Association for Computational Linguistics (ACL)*, page 7724–7736, 2020.

[482] Danielle Saunders, Rosie Sallis, and Bill Byrne. Neural Machine Translation Doesn't Translate Gender Coreference Right Unless You Make It. In *Proceedings of the Workshop on Gender Bias for Natural Language Processing*, 2020.

[483] Yves Scherrer. Syntactic transformations for Swiss German dialects. In *Proceedings of the Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, pages 30–38, Edinburgh, Scotland, 2011.

[484] Thalea Schlender and Gerasimos Spanakis. "Thy algorithm shalt not bear false witness": An Evaluation of Multiclass Debiasing Methods on Word Embeddings. In *Proceedings of BNAIC/BeneLearn*, 2020.

[485] Ari Schlesinger, Kenton P. O'Hara, and Alex S. Taylor. Let's Talk About Race: Identity, Chatbots, and AI. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, Montréal, Canada, 2018.

[486] Natalie Schluter. The word analogy testing caveat. In *Proceedings of the North American Association for Computational Linguistics (NAACL)*, pages 242–246, New Orleans, LA, 2018.

[487] Nathan Schneider, Brendan O'Connor, Naomi Saphra, David Bamman, Manaal Faruqui, Noah A. Smith, Chris Dyer, and Jason Baldridge. A Framework for (Under)specifying Dependency Syntax without Overloading Annotators. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 51–60, Sofia, Bulgaria, 2013.

[488] Tyler Schnoebelen. Goal-Oriented Design for Ethical Machine Learning and NLP. In *Proceedings of the Workshop on Ethics in Natural Language Processing*, pages 88–93, Valencia, Spain, 2017.

[489] Patrick Schramowski, Cigdem Turan, Sophie Jentzsch, Constantin Rothkopf, and Kristian Kersting. BERT has a Moral Compass: Improvements of ethical and moral values of machines. *arXiv preprint arXiv:1912.05238*, 2019.

[490] Carsten Schwemmer, Carly Knight, Emily D. Bello-Pardo, Stan Oklobdzija, Martijn Schoonvelde, and Jeffrey W. Lockhart. Diagnosing Gender Bias in Image Recognition Systems. *Socius: Sociological Research for a Dynamic World*, 2020.

[491] Sabine Sczesny, Magda Formanowicz, and Franziska Moser. Can gender-fair language reduce gender stereotyping and discrimination? *Frontiers in Psychology*, 7, 2016.

[492] Justin Sech, Alexandra DeLucia, Anna L. Buczak, and Mark Dredze. Civil Unrest on Twitter (CUT): A Dataset of Tweets to Support Research on Civil Unrest. In *Proceedings of the Workshop on Noisy User-Generated Text*, 2020.

[493] João Sedoc and Lyle Ungar. The Role of Protected Class Word Lists in Bias Identification of Contextualized Word Representations. In *Proceedings of the Workshop on Gender Bias in Natural Language Processing*, pages 55–61, Florence, Italy, 2019.

[494] Beth Semel. The Body Audible: From Vocal Biomarkers to a Phrenology of the Throat. *Somatosphere*, 2020.

[495] Procheta Sen and Debasis Ganguly. Towards Socially Responsible AI: Cognitive Bias-Aware Multi-Objective Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, New York, NY, 2020.

[496] Marie Gustafsson Sendén, Torun Lindholm, and Sverker Sikström. Biases in News Media as Reflected by Personal Pronouns in Evaluative Contexts. *Social Psychology*, 45(2):103–111, 2014.

[497] Marie Gustafsson Sendén, Sverker Sikström, and Torun Lindholm. "She" and "He" in News Media Messages: Pronoun Use Reflects Gender Biases in Semantic Contexts. *Sex Roles*, 72(1-2):40–49, 2015.

[498] Cristine Gorski Severo and Sinfree B. Makoni. African Languages, Race, and Colonialism: The Case of Brazil and Angola. In H. Samy Alim, Angela Reyes, and Paul V. Kroskrity, editors, *The Oxford Handbook of Language and Race*. Oxford University Press, 2020.

[499] Deven Shah, H. Andrew Schwartz, and Dirk Hovy. Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. In *Proceedings of the Association for Computational Linguistics (ACL)*, page 5248–5264, 2020.

[500] Judy Hanwen Shen, Lauren Fratamico, Iyad Rahwan, and Alexander M. Rush. *Darling* or *Babygirl?* Investigating Stylistic Bias in Sentiment Analysis. In *Proceedings of the Workshop on Fairness, Accountability, and Transparency (FAT/ML)*, Stockholm, Sweden, 2018.

[501] Emily Sheng and David Uthus. Investigating Societal Biases in a Poetry Composition System. In *Proceedings of the Workshop on Gender Bias in Natural Language Processing*, 2020.

[502] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. The Woman Worked as a Babysitter: On Biases in Language Generation. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 3398–3403, Hong Kong, China, 2019.

[503] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. "Nice Try, Kiddo": Ad Hominems in Dialogue Systems. *arXiv preprint arXiv:2010.12820*, 2020.

[504] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. Towards Controllable Biases in Language Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3239–3254, 2020.

[505] Seungjae Shin, Kyungwoo Song, JoonHo Jang, Hyemi Kim, Weonyoung Joo, and Il-Chul Moon. Neutralizing Gender Bias in Word Embedding with Latent Disentanglement and Counterfactual Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3126–3140, 2020.

[506] Philippa Shoemark, James Kirby, and Sharon Goldwater. Topic and audience effects on distinctively Scottish vocabulary usage in Twitter data. In *Proceedings of the Workshop on Stylistic Variation*, pages 59–68, Copenhagen, Denmark, 2017.

[507] Vered Shwartz, Rachel Rudinger, and Oyvind Tafjord. "You are grounded!": Latent Name Artifacts in Pre-trained Language Models. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 6850–6861, 2020.

[508] Michael Silverstein. Language structure and linguistic ideology. In Paul R. Clyne, William F. Hanks, and Carol L. Hofbauer, editors, *The elements: A parasession on linguistic units and levels*, pages 193–247, Chicago, IL, 1979.

[509] Michael Silverstein. Monoglot "Standard" in America: Standardization and Metaphors of Linguistic Hegemony. In Donald Brenneis and Ronald Macaulay, editors, *The Matrix of Language: Contemporary Linguistic Anthropology*, pages 284–306. Westview Press, 1996.

[510] Jesper Simonsen and Toni Robertson, editors. *Routledge International Handbook of Participatory Design.* Routledge, 2013.

[511] Krystal A. Smalls. Languages of Liberation: Digital Discourses of Emphatic Blackness. In Netta Avineri, Laura R. Graham, Eric J. Johnson, Robin Conley Riner, and Jonathan Rosa, editors, *Language and Social Justice in Practice*, pages 52–60. Routledge, 2019.

[512] Hiram L. Smith. Has nigga Been Reappropriated as a Term of Endearment?: A Qualitative and Quantitative Analysis. *American Speech*, 94(4):420–477, 2019.

[513] Geneva Smitherman. *Talkin and Testifyin: The Language of Black America.* Wayne State University Press, 1986.

[514] Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, and Jasmine Wang. Release Strategies and the Social Impacts of Language Models. Technical report, OpenAI, November 2019.

[515] Thamar Solorio and Yang Liu. Learning to predict code-switching points. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 973–981, Honolulu, HI, 2008.

[516] Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, and Pascale Fung. Overview for the First Shared Task on Language Identification in Code-Switched Data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, 2014.

[517] Ezekiel Soremekun, Sakshi Udeshi, and Sudipta Chattopadhyay. Astraea: Grammar-based Fairness Testing. *arXiv preprint arXiv:2010.02542*, 2020.

[518] Dean Spade. *Normal Life: Administrative Violence, Critical Trans Politics and the Limits of Law.* South End Press, 2011.

[519] Arthur K. Spears. African-American Language Use: Ideology and So-Called Obscenity. In Guy Bailey, John Baugh, Salikoko S. Mufwene, and John R. Rickford, editors, *African-American English: Structure, History, and Use*, pages 226–250. Routledge, 1998.

[520] Arthur Spirling and Pedro L. Rodriguez. Word Embeddings: What works, what doesn't, and how to tell the difference for applied research. Working paper, 2019.

[521] Maximilian Spliethöver and Henning Wachsmuth. Argument from Old Man's View: Assessing Social Bias in Argumentation. In *Proceedings of the Workshop on Argument Mining*, 2020.

[522] Artūrs Stafanovičs, Tom Bergmanis, and Mārcis Pinnis. Mitigating Gender Bias in Machine Translation with Target Gender Annotations. In *Proceedings of the Conference on Machine Translation (WMT)*, 2020.

[523] Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. Evaluating gender bias in machine translation. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 1679–1684, Florence, Italy, 2019.

[524] Luke Stark and Jesse Hoey. The Ethics of Emotion in AI Systems. In *WeRobot*, 2020.

[525] Jonathan Stevenson. Dialect in digitally mediated written interaction: A survey of the geohistorical distribution of the ditransitive in British English using Twitter. Master's thesis, University of York, 2016.

[526] Ian Stewart. Now We Stronger Than Ever: African-American Syntax in Twitter. In *Proceedings of the Student Research Workshop at EACL*, pages 31–37, Gothenburg, Sweden, 2014.

[527] Ian Stewart, Yuval Pinter, and Jacob Eisenstein. Sí o no, ¿què penses? Catalonian Independence and Linguistic Identity on Social Media. In *Proceedings of the North American Association for Computational Linguistics (NAACL)*, pages 136–141, New Orleans, LA, 2018.

[528] Dennis Storoshenko. Regional Variation in the Use of English th- Reflexive Forms. In *Linguistics Society of America Annual Meeting*, 2020.

[529] Milan Straka, Jan Hajič, and Jana Straková. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 4290–4297, Portorož, Slovenia, 2016.

[530] Christopher Strelluf. anymore, It's on Twitter: Positive-Anymore, American Regional Dialects, and Polarity-Licensing Tweets. *American Speech*, 94(3): 313–351, 2019.

[531] Yolande Strengers, Lizhe Qu, Qiongkai Xu, and Jarrod Knibbe. Adhering, Steering, and Queering: Treatment of Gender in Natural Language Generation. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, Honolulu, HI, 2020.

[532] Sandeep Subramanian, Guillaume Lample, Eric Michael Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. Multiple-Attribute Text Style Transfer. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

[533] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 1630–1640, Florence, Italy, 2019.

[534] Harini Suresh and John V. Guttag. A Framework for Understanding Unintended Consequences of Machine Learning. *arXiv preprint arXiv:1901.10002*, 2019.

[535] Adam Sutton, Thomas Lansdall-Welfare, and Nello Cristianini. Biased embeddings from wild data: Measuring, understanding and removing. In *Proceedings of the International Symposium on Intelligent Data Analysis*, pages 328–339, 's-Hertogenbosch, Netherlands, 2018.

[536] Chris Sweeney and Maryam Najafian. A Transparent Framework for Evaluating Unintended Demographic Bias in Word Embeddings. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 1662–1667, Florence, Italy, 2019.

[537] Chris Sweeney and Maryam Najafian. Reducing sentiment polarity for demographic attributes in word embeddings using adversarial learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 359–368, Barcelona, Spain, 2020.

[538] Latanya Sweeney. Discrimination in online ad delivery. *Queue*, 11(3), 2013.

[539] Nathaniel Swinger, Maria De-Arteaga, Neil Thomas Heffernan, Mark D.M. Leiserson, and Adam Tauman Kalai. What are the biases in my word embedding? In *Proceedings of the Conference on Artificial Intelligence, Ethics, and Society (AIES)*, Honolulu, HI, 2019.

[540] Matt Taddy. Multinomial inverse regression for text analysis. *Journal of the American Statistical Association*, 108(503):755–770, 2013.

[541] Samson Tan, Shafiq Joty, Min-Yen Kan, and Richard Socher. It's Morphin' Time! Combating Linguistic Discrimination with Inflectional Perturbations. In *Proceedings of the Association for Computational Linguistics (ACL)*, page 2920–2935, 2020.

[542] Yi Chern Tan and L. Elisa Celis. Assessing Social and Intersectional Biases in Contextualized Word Representations. In *Proceedings of the Conference on Neural Information Processing Systems*, Vancouver, Canada, 2019.

[543] Ruixiang Tang, Mengnan Du, Yuening Li, Zirui Liu, and Xia Hu. Mitigating Gender Bias in Captioning Systems. *arXiv preprint arXiv:2006.08315*, 2020.

[544] Marie Tano and Nicole Holliday. "It's a Whole Vibe": Testing Evaluations of Grammatical and Ungrammatical AAE on Twitter. Advancing African American Linguist(ic)s Symposium, 2020.

[545] Rachael Tatman. Gender and Dialect Bias in YouTube's Automatic Captions. In *Proceedings of the Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain, 2017.

[546] Joel Tetreault, Daniel Blanchard, and Aoife Cahill. A Report on the First Native Language Identification Shared Task. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, pages 48–57, Atlanta, GA, 2013.

[547] Mike Thelwall. Gender Bias in Sentiment Analysis. *Online Information Review*, 42(1):45–57, 2018.

[548] Rocco Tripodi, Massimo Warglien, Simon Levis Sullam, and Deborah Paci. Tracing Antisemitic Language Through Diachronic Embedding Projections: France 1789-1914. In *Proceedings of the Workshop on Computational Approaches to Historical Language Change*, pages 115–125, Florence, Italy, 2019.

[549] Kristen Vaccaro, Karrie Karahalios, Deirdre K. Mulligan, Daniel Kluttz, and Tad Hirsch. Contestability in Algorithmic Systems. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*, pages 523–527, Austin, TX, 2019.

[550] Ameya Vaidya, Feng Mai, and Yue Ning. Empirical Analysis of Multi-Task Learning for Reducing Model Bias in Toxic Comment Detection. *arXiv preprint arXiv:1909.09758v2*, 2019.

[551] Rob van der Goot and Gertjan van Noord. Parser Adaptation for Social Media by Integrating Normalization. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 491–497, Vancouver, Canada, 2017.

[552] Eva Vanmassenhove, Christian Hardmeier, and Andy Way. Getting Gender Right in Neural Machine Translation. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 3003–3008, Brussels, Belgium, 2018.

[553] Francisco Vargas and Ryan Cotterell. Exploring the Linear Subspace Hypothesis in Gender Bias Mitigation. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 2902–2913, 2020.

[554] Suresh Venkatasubramanian and Mark Alfano. The philosophical basis of algorithmic recourse. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Barcelona, Spain, 2020.

[555] Dimitra Vergyri, Lori Lamel, and Jean-Luc Gauvain. Automatic Speech Recognition of Multiple Accented English Data. In *Proceedings of INTERSPEECH*, pages 1652–1655, Chiba, Japan, 2010.

[556] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Causal Mediation Analysis for Interpreting Neural NLP: The Case of Gender Bias. *arXiv preprint arXiv:2004.12265*, 2020.

[557] James Vincent. Amazon reportedly scraps internal AI recruiting tool that was biased against women. The Verge, 2018. `https://www.theverge.com/2018/10/10/17958784/ai-recruiting-tool-bias-amazon-report`.

[558] Ngugi Wa Thiong'o. *Decolonising the mind: The politics of language in African literature*. James Currey, 1986.

[559] Claudia Wagner, Eduardo Graells-Garrido, David Garcia, and Filippo Menczer. Women Through the Glass Ceiling: Gender Asymmetries in Wikipedia. *EPJ Data Science*, 5(5), 2016.

[560] Jeremy Waldron. *The Harm in Hate Speech*. Harvard University Press, 2012.

[561] Esma Wali, Yan Chen, Christopher M. Mahoney, Thomas G. Middleton, Marzieh Babaeianjelodar, Mariama Njie, and Jeanna Neefe Matthews. Is Machine Learning Speaking my Language? A Critical Look at the NLP-Pipeline Across 8 Human Languages. In *Proceedings of the Participatory Approaches to Machine Learning Workshop*, 2020.

[562] Eric Wallace, Matt Gardner, and Sameer Singh. Interpreting Predictions of NLP Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 20–23, 2020.

[563] Hongmin Wang, Yue Zhang, GuangYong Leonard Chan, Jie Yang, and Hai Leong Chieu. Universal Dependencies Parsing for Colloquial Singaporean English. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 1732–1744, Vancouver, Canada, 2017.

[564] Tianlu Wang, Xi Victoria Lin, Nazneen Fatema Rajani, Bryan McCann, Vicente Ordonez, and Caiming Xiong. Double-Hard Debias: Tailoring Word Embeddings for Gender Bias Mitigation. In *Proceedings of the Association for Computational Linguistics (ACL)*, page 5443–5453, 2020.

[565] Vincent D. Warmerdam, Thomas Kober, and Rachael Tatman. Going Beyond T-SNE: Exposing `whatlies` in Text Embeddings. *arXiv preprint arXiv:2009.02113*, 2020.

[566] Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–618, 2018.

[567] Kellie Webster, Marta R. Costa-jussà, Christian Hardmeier, and Will Radford. Gendered Ambiguous Pronoun (GAP) Shared Task at the Gender Bias in NLP Workshop 2019. In *Proceedings of the Workshop on Gender Bias in Natural Language Processing*, pages 1–7, Florence, Italy, 2019.

[568] Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, and Slav Petrov. Measuring and Reducing Gendered Correlations in Pre-trained Models. *arXiv preprint arXiv:2010.06032*, 2020.

[569] Laura Wendlandt, Jonathan K. Kummerfeld, and Rada Mihalcea. Factors Influencing the Surprising Instability of Word Embeddings. In *Proceedings of the North American Association for Computational Linguistics (NAACL)*, pages 2092–2102, New Orleans, LA, 2018.

[570] Mark West, Rebecca Kraut, and Han Ei Chew. I'd blush if I could: Closing gender divides in digital skills through education. Technical report, UNESCO, 2019.

[571] Melvin Wevers. Using Word Embeddings to Examine Gender Bias in Dutch Newspapers, 1950-1990. In *Proceedings of the Workshop on Computational Approaches to Historical Language Change*, pages 92–97, Florence, Italy, 2019.

[572] Langdon Winner. Do Artifacts Have Politics? *Daedalus*, 109(1):121–136, 1980.

[573] Walt Wolfram. Sociolinguistic Folklore in the Study of African American English. *Language and Linguistics Compass*, 1(4):292–313, 2007.

[574] Walt Wolfram. The Sociolinguistic Construction of African American Language. In Jennifer Bloomquist, Lisa J. Green, and Sonja L. Lanehart, editors, *The Oxford Handbook of African American Language*, pages 338–352. Oxford University Press, 2015.

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

[576] Alan Wong. Mitigating Gender Bias in Neural Machine Translation Using Counterfactual Data. Master's thesis, The Graduate Center, City University of New York, 2020.

[577] Heather Suzanne Woods. Asking more of Siri and Alexa: Feminine persona in service of surveillance capitalism. *Critical Studies in Media Communication*, 35 (4):334–349, 2018.

[578] Kathryn A. Woolard. Language and identity choice in Catalonia: The interplay of contrasting ideologies of linguistic authority. In Kirsten Süselbeck, Ulrike Mühlschlegel, and Peter Masson, editors, *Lengua, nación e identidad. La regulación del plurilingüismo en Espana y América Latina*, pages 303–323. Vervuert/Iberoamericana, 2008.

[579] Kathryn A. Woolard and Bambi B. Schieffelin. Language Ideology. *Annual Review of Anthropology*, 23:55–82, 1994.

[580] Austin P. Wright, Omar Shaikh, Haekyu Park, Will Epperson, Muhammed Ahmed, Stephane Pinel, Diyi Yang, and Duen Horng (Polo) Chau. RECAST: Interactive Auditing of Automatic Toxicity Detection Models. In *Proceedings of the International Workshop of Chinese CHI*, Honolulu, HI, 2020.

[581] Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. Recipes for Safety in Open-domain Chatbots. *arXiv preprint arXiv:2010.07079*, 2020.

[582] Zekun Yang and Juan Feng. A Causal Inference Method for Reducing Gender Bias in Word Embedding Relations. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, New York, NY, 2020.

[583] Yuxin Yao and Jan C. van Ours. Daily dialect-speaking and wages among native Dutch speakers. *Empirica*, 46:653–668, 2019.

[584] Catherine Yeo and Alyssa Chen. Defining and Evaluating Fair Natural Language Generation. In *Proceedings of the Fourth Widening Natural Language Processing Workshop*, 2020.

[585] Daisy Yoo, Anya Ernest, Sofia Serholt, Eva Eriksson, and Peter Dalsgaard. Service Design in HCI Research: The Extended Value Co-creation Model. In *Proceedings of the Halfway to the Future Symposium*, Nottingham, United Kingdom, 2019.

[586] Iris Marion Young. Five Faces of Oppression. In *Justice and the Politics of Difference*, pages 39–65. Princeton University Press, 2nd edition, 2011.

[587] Vershawn Ashanti Young, Rusty Barrett, and Kim Brian Lovejoy. *Other People's English: Code-meshing, Code-switching, and African American Literacy*. Teachers College Press, 2014.

[588] Marcos Zampieri, Binyam Gebrekidan Gebre, and Sascha Diwersy. N-gram language models and POS distribution for the identification of Spanish varieties. In *Proceedings of TALN (Traitement Automatique des Langues Naturelles)*, pages 580–587, Les Sables d'Olonne, France, 2013.

[589] Daniel Zeman, Martin Popel, Milan Straka, Jan Hajic, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, et al. CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 2017.

[590] Ana Celia Zentella. LatinUs* and linguistics: Complaints, conflicts, and contradictions - The anthro-political linguistics solution. In Naomi Lapidus Shin and Daniel Erker, editors, *Questioning Theoretical Primitives in Linguistic Inquiry*, pages 189–207. John Benjamins Publishing Company, 2018.

[591] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the Conference on Artificial Intelligence, Ethics, and Society (AIES)*, New Orleans, LA, 2018.

[592] Congle Zhang, Tyler Baldwin, Howard Ho, Benny Kimelfeld, and Yunyao Li. Adaptive Parser-Centric Text Normalization. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 1159–1168, Sofia, Bulgaria, 2013.

[593] Guanhua Zhang, Bing Bai, Junqi Zhang, Kun Bai, Conghui Zhu, and Tiejun Zhao. Demographics Should Not Be the Reason of Toxicity: Mitigating Discrimination in Text Classifications with Instance Weighting. In *Proceedings of the Association for Computational Linguistics (ACL)*, page 4134–4145, 2020.

[594] Haiyang Zhang, Alison Sneyd, and Mark Stevenson. Robustness and Reliability of Gender Bias Assessment in Word Embeddings: The Role of Base Pairs. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing (AACL-IJCNLP)*, 2020.

[595] Haoran Zhang, Amy X. Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. Hurtful Words: Quantifying Biases in Clinical Contextual Word Embeddings. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, 2020.

[596] Jessica Zhao, Bill Zhang, and Himabindu Lakkaraju. Identifying Interpretable Word Vector Subspaces with PCA. In *Proceedings of the Machine Learning In Real Life Workshop*, 2020.

[597] Jieyu Zhao and Kai-Wei Chang. LOGAN: Local Group Bias Detection by Clustering. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 1968–1977, 2020.

[598] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 2979–2989, Copenhagen, Denmark, 2017.

[599] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the North American Association for Computational Linguistics (NAACL)*, pages 15–20, New Orleans, LA, 2018.

[600] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. Learning Gender-Neutral Word Embeddings. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 4847–4853, Brussels, Belgium, 2018.

[601] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. Gender Bias in Contextualized Word Embeddings. In *Proceedings of the North American Association for Computational Linguistics (NAACL)*, pages 629–634, Minneapolis, MN, 2019.

[602] Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. Gender Bias in Multilingual Embeddings and Cross-Lingual Transfer. In *Proceedings of the Association for Computational Linguistics (ACL)*, page 2896–2907, 2020.

[603] Alina Zhiltsova, Simon Caton, and Catherine Mulwa. Mitigation of Unintended Biases against Non-Native English Texts in Sentiment Analysis. In *Proceedings of the Irish Conference on Artificial Intelligence and Cognitive Science*, Galway, Ireland, 2019.

[604] Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. The Design and Implementation of XiaoIce, an Empathetic Social Chatbot. *Computational Linguistics*, 46:53–93, 2020.

[605] Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, and Kai-Wei Chang. Examining gender bias in languages with grammatical genders. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 5279–5287, Hong Kong, China, 2019.

[606] Lal Zimman. Transgender language reform: Some challenges and strategies for promoting trans-affirming, gender-inclusive language. *Journal of Language and Discrimination*, 1(1):84–105, 2017.

[607] Lal Zimman and Kira Hall. Language, Gender, and Sexuality. *Oxford Bibliographies*, 2016. Annotated bibliography.

[608] Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 1651–1661, Florence, Italy, 2019.

[609] Arkaitz Zubiaga, Inaki San Vincente, Pablo Gamallo, Jose Ramom Pichel, Inaki Algeria, Nora Aranberri, Aitzol Ezeiza, and Victor Fresno. Overview of TweetLID: Tweet Language Identification at SEPLN 2014. In *Proceedings of the Tweet Language Identification Workshop*, 2014.

[610] Nadezhda Zueva, Madina Kabirova, and Pavel Kalaidin. Reducing Unintended Identity Bias in Russian Hate Speech Detection. In *Proceedings of the Workshop on Online Abuse and Harms*, 2020.