

Situated Data, Situated Systems: A Methodology to Engage with Power Relations in Natural Language Processing Research

Lucy Havens[†] Melissa Terras[‡] Benjamin Bach[†] Beatrice Alex^{§†}

[†]School of Informatics

[‡]College of Arts, Humanities and Social Sciences

[§]Edinburgh Futures Institute; School of Literatures, Languages and Cultures
University of Edinburgh

lucy.havens@ed.ac.uk, m.terras@ed.ac.uk
bbach@inf.ed.ac.uk, balex@ed.ac.uk

Abstract

We propose a bias-aware methodology to engage with power relations in natural language processing (NLP) research. NLP research rarely engages with bias in social contexts, limiting its ability to mitigate bias. While researchers have recommended actions, technical methods, and documentation practices, no methodology exists to integrate critical reflections on bias with technical NLP methods. In this paper, after an extensive and interdisciplinary literature review, we contribute a bias-aware methodology for NLP research. We also contribute a definition of biased text, a discussion of the implications of biased NLP systems, and a case study demonstrating how we are executing the bias-aware methodology in research on archival metadata descriptions.

1 Introduction

Analysis of computer systems has raised awareness of their biases, prompting researchers to make recommendations to mitigate harms that biased computer systems cause. Analysis has shown computer systems exhibiting biases through racism¹ (Noble, 2018), sexism² (Perez, 2019), and classism³ (D’Ignazio and Klein, 2020). This list of harms is not exhaustive; biased computer systems may also harm people based on ability, citizenship, and any other identity characteristic. To mitigate harms from biased computer systems, researchers have recommended actions, methods, and practices. However, none of the recommendations comprehensively address the complexity of the problems bias causes.

Considering the numerous *types* of bias that may enter a natural language processing (NLP) system, *places* that bias may enter, and *harms* that bias may cause, we propose a bias-aware methodology to comprehensively address the consequences of bias for NLP research. Our methodology integrates critical reflection on social influences on and implications of NLP research with technical NLP methods. To scope our research direction and inform our methodology, we draw on an interdisciplinary selection of literature that includes work from the humanities, arts, and social sciences. We intend the methodology to (a) support the reproducibility of NLP research, enabling researchers to better understand which perspectives were considered in the research; and (b) diversify perspectives in NLP systems, guiding researchers in explicitly communicating the social context their research so others can situate future research in contexts that have yet to be investigated.

We begin with our bias statement (§2) and motivations for proposing a bias-aware NLP research methodology (§3). Next, we summarize the interdisciplinary literature informing our methodology (§4), explain the methodology (§5), and demonstrate it with a case study of our ongoing research with bias in archival metadata descriptions (§6). We end with a summary and vision for future NLP research (§7).

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: <http://creativecommons.org/licenses/by/4.0/>.

¹“A belief that one’s own racial or ethnic group is superior” (Oxford English Dictionary, 2013c).

²“[P]rejudice, stereotyping, or discrimination, typically against women, on the basis of sex” (Oxford English Dictionary, 2013d).

³“The belief that people can be distinguished or characterized, esp. as inferior, on the basis of their social class” (Oxford English Dictionary, 2013a).

2 Bias Statement

We situate this paper in the United Kingdom (UK) in the 21st century, writing as authors who primarily work as academic researchers. We identify as three females and one male; and as American, German, and Scots. Together we have experience in natural language processing, human-computer interaction, data visualization, digital humanities, and digital cultural heritage. In this paper, we propose a bias-aware methodology for NLP researchers. We define **biased language** as *written or spoken language that creates or reinforces inequitable power relations among people, harming certain people through simplified, dehumanizing, or judgmental words or phrases that restrict their identity; and privileging other people through words or phrases that favor their identity*. Biased language causes representational harms (Vainapel et al., 2015; Sweeney, 2013), or the restriction of a person’s identity through the use of hyperbolic or simplistic language (Blodgett et al., 2020; Talbot, 2003). NLP systems built on biased language become biased computer systems, which “*systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others*” (Friedman and Nissenbaum, 1996, p. 332). Representational harms may cause inequitable system performance for different groups of people, leading to allocative harms (Zhang et al., 2020; Noble, 2018), or the denial of a resource or opportunity (Blodgett et al., 2020). The people who experience harms from biased NLP systems varies with the context in which people use the system and with the language source on which the system relies. Moreover, people may not be aware they are being harmed given the black-box nature of many systems (Koene et al., 2017). That being said, whether or not people realize they are being prejudiced against, the people harmed will be those excluded from the most powerful social group.

3 Why does NLP need a Bias-Aware Methodology?

Statistics report a homogeneity of perspectives among students in computer-related disciplines that do not reflect the diversity of people affected by computer systems, risking a homogeneity of perspectives in the technology workforce and the computer systems that workforce develops. For academic year 2018/19, statistics on students in the UK⁴ report that the dominant group of people studying computer-related subjects overwhelmingly are white males without a disability.⁵ Moreover, differences in total numbers of surveyed students across identity characteristics (e.g. sex, ethnicity, disability) skew the statistics in favor of those reported as white, male, and without a disability. Lack of diverse perspectives among students in computer-related disciplines may limit the diversity of perspectives in the workforce, where the development of NLP and other computer systems occurs. As of 2019, the Wise Campaign reported that women comprise 24% of the core-STEM workforce in the UK.⁶ Lack of diverse perspectives in the development of NLP and other computer systems risks technological decisions that exclude groups of people (“technical bias”), as well as applications of computer systems that oppress groups of people (“emergent bias”) (Friedman and Nissenbaum, 1996).

That being said, even if student demographics in NLP and computer-related disciplines become more balanced, the data underlying NLP systems will still cause bias. Theories of discourse state that language (written or spoken) reflects and reinforces “society, culture and power” (Bucholtz, 2003, p. 45). In turn, NLP systems built on human language reflect and reinforce power relations in society, inheriting biases in language (Caliskan et al., 2017) such as stereotypical expectations of genders (Haines et al., 2016) and ethnicities (Garg et al., 2018). Drawing on feminist theory, we argue that all language is biased, because language records human interpretations that are situated in a specific time, place, and worldview (Haraway, 1988). Consequently, all NLP systems are subject to biases originating in the social contexts in which the systems are built (“preexisting bias”) (Friedman and Nissenbaum, 1996). Psychology research suggests that biased language causes representational harms: Vainapel et al. (2015) studied how masculine-generic language (e.g. “he”) versus gender-neutral language (e.g. “he or she”)

⁴Situating our research in the UK, we reference statistics from the UK’s Higher Education Statistical Agency (HESA).

⁵www.hesa.ac.uk/news/16-01-2020/sb255-higher-education-student-statistics/subjects.

⁶<http://www.wisecampaign.org.uk/statistics/2019-workforce-statistics-one-million-women-in-stem-in-the-uk/>

affected participants' responses to questionnaires. The authors report that women gave themselves lower scores on intrinsic goal orientation and task value in questionnaires using masculine-generic language in contrast to questionnaires using gender-neutral language.⁷ The study provides an example of how biased language may harm select groups of people, because the participants reported as women experienced a restriction of their identity, influencing their behavior to conform to stereotypes.

Acknowledging the harms of biased language and biased NLP systems, researchers have proposed approaches mitigating bias, though no approach has fully removed bias from an NLP dataset or algorithm. To mitigate bias in datasets, Webster et al. (2018) produced a dataset of gendered ambiguous pronouns (GAP) to provide an unbiased text source on which to train NLP algorithms. However, the GAP dataset reverses gender roles, assuming that gender is a binary rather than a spectrum.⁸ Any NLP system that uses the GAP dataset thus adopts its preexisting gender bias. Efforts to mitigate bias in algorithms are similarly limited, focusing on technical performance rather than performance in social contexts. Zhao et al. (2018) describe an approach to debias word embeddings, writing, "Finally we show that given sufficiently strong alternative cues, systems can ignore their bias" (p. 16). However, the paper does not explain the intended social context in which to apply the authors' approach, risking emergent bias.⁹ Additionally, Gonen and Goldberg (2019) demonstrate how this debiasing approach hides, rather than removes, bias. In our bias-aware methodology, we describe documentation and user research practices that facilitate transparent communication of biases that may be present in NLP systems, facilitating reflection on how to include more diverse perspectives and empower underrepresented people.

4 Interdisciplinary Literature Review

To inform our proposed bias-aware NLP research methodology, we draw on an interdisciplinary corpus of literature from computer science, data science, the humanities, the arts, and the social sciences.

NLP and ML scholars have recommended actions to diversify perspectives in technological research, recognizing the value of diversity to bias mitigation. Blodgett et al. (2020) and Crawford (2017) recommend interdisciplinary collaboration so researchers can learn from humanistic, artistic, and sociological disciplines regarding human behavior, helping researchers to more effectively anticipate harms that computer systems may cause, in addition to benefits they may bring, addressing risks of emergent bias. They also recommend engaging with the people affected by NLP and other computer systems, testing on more diverse populations to address the risk of technical bias, and rethinking power relations between those who create and those who are affected by computer systems to address the risk of preexisting bias. Though these recommendations address the three types of bias that may enter an NLP system, they do not articulate how to identify relevant people to include in the development and testing of NLP systems. Our bias-aware methodology builds on recommendations from Blodgett et al. (2020) and Crawford (2017) by outlining how to identify and include stakeholders in NLP research (§5.1).

D'Ignazio and Klein (2020) propose data feminism as an approach to addressing bias in data science. They define data feminism as, "a way of thinking about data, both their uses and their limits, that is informed by direct experience, by a commitment to action, and by intersectional feminist thought" (p. 8).¹⁰ Data feminism has seven principles: examine power, challenge power, elevate emotion and embodiment, rethink binaries and hierarchies, embrace pluralism, consider context, and make labor visible. These principles facilitate critical reflection on the impacts of data's collection and use in social contexts. Our bias-aware methodology tailors these principles to NLP research, outlining activities that encourage researchers to consider influences on and implications of their work beyond the NLP community (§5.1).

⁷The authors report that men showed no difference in their intrinsic goal orientation and task value scores with masculine-generic versus gender-neutral language in the questionnaires; impacts on people who do not identify as either a man or a woman are unknown as the study groups participants into these two gender categories (Vainapel et al., 2015).

⁸See HCI Guidelines for Gender Equity and Inclusivity at www.morgan-klaus.com/gender-guidelines.html.

⁹While earlier paragraphs in the paper indicate a focus on gender bias and stereotypes related to professional occupations, the authors do not define *bias* or *gender bias*, nor do they identify the types of *systems* to which they refer.

¹⁰Intersectionality refers to the way in which different combinations of identity characteristics from one individual to another result in different experiences of privilege and oppression (Crenshaw, 1991). In feminist thought, multiple viewpoints are needed to understand reality; viewpoints that claim to be objective are, in fact, subjective, because knowledge is the result of human interpretation (Haraway, 1988).

Within the NLP research community, Bender and Friedman (2018) recommend improved documentation practices to mitigate emergent, technical, and preexisting biases. They recommend all NLP research includes a “data statement,” which they describe as, “a characterization of a dataset that provides context to allow developers and users to better understand how experimental results might generalize, how software might be appropriately deployed, and what biases might be reflected in systems built on the software” (p. 587). Aimed at developers and users of NLP systems, data statements reduce the risk of emergent bias. The authors also note: “As systems are being built, data statements enable developers and researchers to make informed choices about training sets and to flag potential underrepresented populations who may be overlooked or treated unfairly” (p. 599), helping authors of data statements reduce the risk of technical and preexisting biases. A data statement serves as guiding documentation for the case study approach we propose in our bias-aware methodology (§5.2), documenting the specific context in which NLP researchers work. Our bias-aware methodology guides research activities before, during, and after the writing of a data statement: for researchers reading data statements to find a dataset for an NLP system, our methodology guides their evaluation of a dataset’s suitability for research; for researchers writing data statements, our methodology guides their documentation of the data collection process.

In addition to technological disciplines, our methodology draws on critical discourse analysis (van Leeuwen, 2009), participatory action research (Reid and Frisby, 2008; Swantz, 2008), intersectionality (Crenshaw, 1991; D’Ignazio and Klein, 2020), feminism (Haraway, 1988; Harding, 1995; Moore, 2018), and design (Martin and Hanington, 2012). Participatory action research provides a way for NLP researchers to diversify perspectives in their research, engaging with the social context that influences and is affected by NLP systems. Intersectionality reminds researchers of the multitude of experiences of privilege and oppression that bias causes, because no single identity characteristic determines whether a person is “dominant” (favored) or “minoritized” (harmed) (D’Ignazio and Klein, 2020). The case study approach common to design methods enables a researcher to make progress on addressing bias through explicitly situating research in a specific time and place, and conducting user research with people to understand their power relations in that time and place. Feminist theory values perspectives at the margins, encouraging researchers to engage with people who are excluded from the dominant group in a social context. Feminist theorist Harding (1995) writes, “In order to gain a causal critical view of the interests and values that constitute the dominant conceptual projects...one must start from the lives excluded as origins of their design - from ‘marginal’ lives” (p. 341). Our bias-aware research methodology includes collaboration with people at the margins of NLP research in an effort to empower minoritized people.

5 A Bias-aware Methodology

Our bias-aware methodology has three main activities: examining power relations (§5.1), explaining the bias of focus (§5.2), and applying NLP methods (§5.3). Though we discuss the activities individually, we recommend researchers execute them in parallel because each activity informs the others. We aim for the methodology to include activities that researchers may adapt to their own research context, be their focus on algorithm development, adaptation, or application; or on dataset creation. We hope for this paper to begin a dialogue on tailoring a bias-aware methodology to different types of NLP research.

5.1 Examining Power Relations

Stakeholder Identification

An NLP researcher executing the bias-aware methodology will document the distribution of power in the social context relevant to their research and language source. In the bias-aware methodology, a researcher considers language to be a partial record that provides knowledge situated in a specific time, place, and perspective. To understand which people’s perspectives their language source (“the data”) includes and excludes, an NLP researcher will identify **stakeholders**, or those who are represented in, use, manage, or provide the data. Specifically, NLP research stakeholders are (1) the researcher(s), (2) producers of the data, (3) institutions providing access to the data, (4) people represented in the data, and (5) people who use the data. To investigate their stakeholders’ power relations, an NLP researcher will observe who dominates the social setting(s) relevant to their research, and who experiences minoritization in the same

setting(s). After identifying the stakeholders, the researcher will document their roles as dominant or minoritized, along with any limitations to their identification.

Stakeholder Collaboration

To understand how privilege and oppression are experienced among stakeholders, an NLP researcher will conduct **participatory action research** (PAR) (Reid and Frisby, 2008; Swantz, 2008) with representative individuals from all five stakeholder groups. Researchers who conduct PAR attempt to establish collaborative relationships with representatives from their groups of stakeholders. Researchers are not experts bringing NLP systems to stakeholders; rather, researchers and stakeholders collaboratively study a social context to understand how NLP systems could empower people, particularly minoritized people. Instead of seeking an objective perspective, researchers foreground individual stakeholder perspectives, recording them as situated in a specific time and place, and using their multiplicity to gain insight into the complexity of the research’s social context. To understand how NLP research can empower people in a specific social context, we propose four **power relations questions**¹¹ for NLP researchers to answer: (1) who or what is included in the research, (2) who or what is excluded from the research, (3) how will the research define knowledge, and (4) who has agency and who can be empowered?

To understand the impacts of dominant people’s interests and values, research following a bias-aware methodology will begin from the perspective of minoritized people, those who are typically excluded as a result (even if unintentional) of the interests and values of dominant people. The research will define knowledge as situated in specific times, places, and perspectives. The widespread availability of language as digital data may give the illusion of universal representation. However, critical discourse analysis reminds the NLP researcher that their data, composed of discourses,¹² are “socially constructed ways of knowing some aspect of reality” (van Leeuwen, 2009, p. 141). Social hierarchies influence the data that becomes widely available, rendering minoritized groups of people invisible due to their exclusion from the data, or misrepresenting them due to their exclusion from the data collection process.

An NLP researcher will weigh insights gathered from different stakeholder groups equally, making the research’s knowledge multi-faceted. Explicit documentation of the time, place, and perspective that produced the knowledge will inform future NLP research. Should a future researcher wish to reproduce the research, the documentation will guide the future researcher in seeking the proper social context. Should a future researcher wish to build upon the research, they will be able to compare and contrast the research’s social setting with their own, guiding them in determining potential contributions.

Unavailable Stakeholders

In situations where the researcher cannot conduct PAR with stakeholders, the researcher will write a data biography.¹³ A data biography documents where data were collected and stored, who collected and owns the data, and why, when, and how the data were collected (Krause, 2019). Writing a data biography facilitates critical reflection on the social influences on and social implications of a dataset, informing technical decisions when applying NLP methods. Datasets may circulate oppression of minoritized groups through inclusion and through omission. The key to recognizing who is dominant and minoritized is understanding that an individual may be both; power relations vary with the context of research.

5.2 Explaining the Bias of Focus

When explaining the type of bias on which NLP research focuses, a researcher will provide a definition and explain how this type of bias relates to other types of bias. For example, AllSides.com’s ratings may guide the classification of political bias in news,¹⁴ Hanson et al.’s (2015) Accessible Writing Guide may inform research with stakeholders who include people with disabilities, and Hitti et al. (2019) provide a model for how to clearly define and classify gender bias in collaboration with interdisciplinary experts. Table 1 provides examples of gender biased language organized into their gender bias taxonomy. When

¹¹We adapted these questions from Moore’s work on feminist community archiving (Moore, 2018).

¹²“A connected series of utterances by which meaning is communicated” (Oxford English Dictionary, 2013b).

¹³We All Count has a free, interactive data biography tool at wac-survey-rails.herokuapp.com.

¹⁴See the Media Bias Ratings at www.allsides.com/media-bias/media-bias-ratings.

Structural Bias		Contextual Bias	
Gender Generalization	<i>A lawyer must always carry his phone.</i>	Societal Stereotype	<i>The event was sports-themed for all the fathers volunteering.</i>
Explicit Marking of Sex	<i>The role of a waitress is overlooked by the restaurant owners.</i>	Behavioral Stereotype	<i>All girls are sensitive.</i>

Table 1: Biased text examples classified into the gender bias taxonomy of Hitti et al. (2019).

following the bias-aware methodology, NLP research to create annotated datasets for other types of bias will similarly include collaboration with relevant disciplinary experts (i.e. racial bias with critical race theory experts) to define and categorize types of bias relevant to the research. When writing a data statement’s *curation rationale*, an NLP researcher will include a definition of their bias of focus. In the answers to the power relations questions, an NLP researcher will describe how they consider intragroup differences within their stakeholder groups, in addition to differences between dominating and minoritized stakeholder groups, because the intersection of identity characteristics, rather than one identity characteristic in isolation, determines how people experience oppression (Crenshaw, 1991). Due to the complexity that intersecting identity characteristics add to evaluations of bias, in the bias-aware methodology, an NLP researcher will use case studies. Case studies gather information in a clearly-defined context and present the resulting knowledge as connected to a specific time, place, and people. To conduct a case study, an NLP researcher will “determine a problem, make initial hypotheses, conduct research through interviews, observations, and other forms of information gathering [such as PAR], revise hypotheses and theory, and tell a story” (Martin and Hanington, 2012, p. 28). Feminist theory’s focus on agency and lived experience as situated in a specific context adds value to PAR by helping a researcher anticipate and critically examine the implications of PAR’s drive towards action (Reid and Frisby, 2008). When documenting their case study in blogs, presentations, or publications, an NLP researcher will discuss potential applications of the research beyond the case study’s context, anticipating potential benefits and harms. Potential harms may outweigh potential benefits, making the best decision not to build an NLP system (Crawford, 2017).

5.3 Applying NLP Methods

When applying NLP methods in the bias-aware methodology, an NLP researcher should acknowledge biases found with any algorithms they use in their data statement. For example, when applying word embeddings, an NLP researcher could look to Bolukbasi et al. (2016), Caliskan et al. (2017), and Kurita et al. (2019) on gender bias; Swinger et al. (2019) on racial bias; Diaz et al. (2018) on age bias; Papakyr-iakopoulos (2020) on sexuality and nationality bias; and Gonen and Goldberg (2019) on the inadequacy of debiasing word embeddings. When applying part-of-speech tagging, dependency parsing, or machine translation, an NLP researcher could look to Garimella et al. (2019) and Stanovsky et al. (2019) for understanding how these methods have been shown to exhibit gender bias. If an NLP researcher will train an algorithm on their language source, research documentation will describe the training process and results. If the research includes annotation, documentation will include instructions given to annotators.

For NLP research on algorithms, we recommend considering approaches to making bias transparent, in addition to reducing the biased behavior of algorithms. Research from Kaneko et al. (2019) and Zhao et al. (2018) on mitigating bias in word embeddings provide starting points for algorithmic bias research, as their methods have yet to be evaluated in diverse contexts. However, Gonen and Goldberg (2019) have shown the limits of debiasing word embeddings. We argue that the situated nature of data, and thus the situated nature of knowledge drawn from data, makes the elimination of bias impossible. Investigating how to make bias transparent provides an alternative direction for NLP researchers interested in mitigating bias in NLP systems. Whether making bias transparent or reducing biased behavior of algorithms,

NLP researchers following the bias-aware methodology will collaborate with relevant disciplinary experts and minoritized stakeholders in determining how to evaluate an algorithm for bias.

To support the training of algorithms in diverse contexts, NLP research on datasets will define the context of its language source’s collection and annotation. An NLP researcher will provide data statements to inform algorithms’ training and evaluation, ensuring reproducibility and avoiding unintended harms from misapplications of algorithms (Bender and Friedman, 2018). Similarly, dataset research will include disciplinary experts and minoritized stakeholders in datasets’ creation, annotation, and evaluation.

6 Case Study

In this section we describe how we are implementing the bias-aware NLP research methodology in a case study on bias in metadata descriptions from the online archival catalog of the Centre for Research Collections at the University of Edinburgh (“the Archive”).¹⁵ For consistency with the outline of a bias-aware methodology (§5), we group our case study into the same three activities, explaining our examination of power relations (§6.1), our bias of focus (§6.2), and then our application of NLP methods (§6.3). Each subsection includes accomplished, ongoing, and planned future work. To demonstrate how we execute the three activities in parallel, as proposed in §5, we first provide a chronological overview.

Initially, our research began with information gathering linked to a participatory action research (PAR) methodology. We reviewed literature on bias in NLP and archives, and on digital humanities research (collaborations between technologists and humanists that often analyze data sources with historical language). We also met with employees at the Archive to better understand the Archive’s policies, which guide the writing of metadata descriptions and documentation practices, such as the metadata standards used. The employees described how they are proactively challenging the inherited metadata and inherited practices of the Archive, which date back to the 16th century. After the literature review and meeting we began writing data statements for the Archive’s metadata descriptions and for our research. Due to the limited research on NLP methods applied to archival metadata, and limited large-scale analysis of metadata descriptions, we undertook a pilot data project,¹⁶ walking through the process of extracting metadata descriptions from a single archival collection, adding historical context to our documentation of the extracted descriptions, and calculating corpus analytics (using ElementTree¹⁷ and NLTK¹⁸ in a Jupyter Notebook¹⁹). After establishing a workflow to extract metadata descriptions from the Archive’s online catalog, we again met employees at the Archive to discuss the challenges that biased language poses to their work and to their visitors. This meeting helped us add to our data statements, identify stakeholders in our research, and begin describing the stakeholders’ power relations. Moreover, the meeting confirmed the value of an NLP system that detects and classifies bias, as the Archive does not currently have a systematic approach to measuring bias in its catalog’s metadata descriptions.

6.1 Researcher and Archive Power Relations

Stakeholder Identification

In our execution of the bias-aware methodology, we study power relations among five stakeholders: (1) us (the authors) as researchers, (2) the Archive’s employees, (3) the Archive (as an institution), (4) people represented in metadata descriptions, and (5) the Archive’s visitors. Literature on power relations in archives and the wider gallery, library, archive, and museum (GLAM) sector (Adler, 2017; Caswell and Cifor, 2019; Hauswedell et al., 2020; McPherson, 2012; Risam, 2015) informed our identification of these stakeholders. We recorded our understanding of their power relations in our data statement (Appendix A) and power relations document (Appendix C), and will continue expanding and revising these documents until our research ends.

¹⁵Metadata documents information about collections of cultural heritage records. Archival catalogs have numerous metadata fields that contain descriptions written by people who archives hire to document their collection items. These descriptions are the language source we refer to as *archival metadata descriptions* (Angel, Christine M., and Caroline Fuchs, 2018).

¹⁶View the pilot in a Jupyter Notebook at github.com/thegoose20/eula41.

¹⁷docs.python.org/3/library/xml.etree.elementtree.html

¹⁸www.nltk.org

¹⁹jupyter.org

Stakeholder Collaboration

In line with PAR, we collaborate with stakeholders at the Archive to learn about their perception of biased language in metadata descriptions, as well as challenges and potential approaches to addressing the bias. Thus far, we facilitated a group discussion with stakeholders who had a range of roles, including technical, curatorial, administrative, servicing, and documenting responsibilities; and a range of GLAM work experience, from one year to over 20 years. The group discussion informs our understanding of the range of attitudes towards bias and neutrality in archival documentation. We are preparing a survey to study how the Archive’s attitudes about bias and neutrality relate to those of other UK archives. Results of the group discussion enabled us to draft answers to the power relations questions.

Unavailable Stakeholders

To fully answer the power relations questions, we are researching historical changes in the structure of metadata standards used at the Archive. Our stakeholders include people who documented the Archive’s collections but no longer work there, and people who are written about in the Archive’s metadata, which document material dating back to the 1st century AD. To study power relations among these unavailable stakeholders, we are writing a data biography (Appendix B) for the metadata descriptions with the Archive. The data biography informs our understanding of the power relations at play in our research, which in turn informs our data statement and technical decisions about NLP methods to apply.

6.2 Contextual Gender Bias as a Focus

Our NLP research focuses on identifying types of contextual gender bias from archival metadata descriptions, complementing Hitti et al.’s (2019) focus on identifying structural gender bias. We adopt the their taxonomy of gender bias (illustrated in Table 1). The taxonomy has two subtypes of contextual bias: behavioral stereotypes and societal stereotypes. We may expand on definitions and subtypes of contextual bias during our research into simplistic, hyperbolic language in metadata descriptions that indicates the presence of stereotypes, because historical text often contains spellings and syntax (among other linguistic characteristics) different to the modern text on which NLP tools have been developed (Casey et al., 2020). In the context of the Archive, gender biased metadata descriptions may cause representational harms, because the Archive supports information access, circulating ideas documented in its metadata when users search its online catalog. Societal and behavioral stereotypes present in the Archive’s metadata descriptions may negatively impact perceptions of people represented in the descriptions. We are researching the types of gender bias in the descriptions, and ways to measure such biases, in an effort to support the Archive in mitigating harms from biased metadata descriptions.

6.3 Information Extraction for Classification

Information Extraction Methods

The archival metadata descriptions we use as this case study’s language source are from the Archive’s public, online catalog. We obtained descriptive metadata fields as Extensible Markup Language (XML) data using the Open Archives Initiative - Protocol for Metadata Harvesting (OAI-PMH),²⁰ filtered the metadata for descriptive fields relevant to our research, and then removed duplicate descriptions. Table 2 summarizes the resulting corpus. The Archive organizes metadata hierarchically, creating metadata for collections, subcollections, and items; we group subcollection and item descriptions within their overarching collection. Currently, we are exploring how to further filter our extracted descriptions through a combination of historical research on archival metadata standards and corpus analytics of terms surrounding gender-related words (as in the third use case from Casey et al. (2020)). For example, the Archive uses Library of Congress Subject Headings (LCSH), which use terms offensive to certain social groups: Adler (2017) discusses how LCSH represents people who do not identify with binary genders or do not conform to heterosexuality as “deviations.” To further filter our extracted metadata descriptions, we can associate the descriptions with the dates they were written and look for offensive terms that were used in metadata standards during those dates. Our data statement further details this process.

²⁰www.openarchives.org/OAI/2.0/openarchivesprotocol.htm

By Metadata Field	Biographical/ Historical	Scope and Contents	Processing Information	Total (sum of the metadata fields)
Sentences	11,323	55,434	1,691	68,448
Words	801,893	208,190	11,016	966,763

By Collection	Minimum	Maximum	Mean	Standard Deviation
Words	7	156,747	1,036.2	7,784.5

Table 2: Words and sentences in the extracted metadata descriptions from the Archive’s 1,231 collections, calculated using Punkt tokenizers in the Natural Language Toolkit Python library (Loper and Bird, 2002).

Annotations to Inform Classification

With our case study, we aim to create and annotate a gold standard dataset on which we will train a classification algorithm to identify types of gender bias in text. We will perform the annotations as part of the research for a Doctor of Philosophy project. Due to ethical concerns regarding the use of crowdsourcing platforms (Gleibs, 2017), anyone employed to contribute to the annotation work will be paid at least minimum wage. To guide the annotation process and ensure the reproducibility of our research, we will document instructions we follow to annotate contextual gender bias. We will collaborate with the Archive and a gender studies expert to write these instructions; we are in the process of finding a language expert with whom to collaborate. When we publish the results of our research, we will provide documentation of the annotation instructions, data statements, data biography, and power relations questions for our NLP research. After creating a gold standard dataset annotated for contextual gender bias, we plan to train a discriminative classifier on the dataset using supervised learning. We will then experiment with and evaluate how the classifier differentiates between types of contextual gender bias in archival metadata descriptions, and report openly on the results of this research.

7 Conclusion

In this paper we propose a bias-aware methodology for NLP research to mitigate harms from biased NLP systems. The methodology integrates practices and methods from NLP, ML, data science, gender and feminist studies, linguistics, and design. Due to the numerous types of bias, the intersectional nature of oppression, and the possibility of direct and indirect harms from bias, detecting and measuring bias is a complex process. Our methodology encourages NLP researchers to situate their work in case studies, explicitly describing the context of and stakeholders in their research. We advise NLP researchers to build the time and resources needed to undertake such work into project plans, and to put eliminating bias at the center of their research. Documenting instances of bias and their associated power relations will enable the NLP community to look for patterns across different contexts that use NLP systems. Amassing case studies in order to look for such patterns will guide NLP research towards generalizable approaches to bias mitigation, approaches that do not unintentionally minoritize people whose perspectives were unknowingly excluded.

Acknowledgments

This paper describes work conducted in collaboration with Rachel Hosker and her team at the Centre for Research Collections (CRC) at the University of Edinburgh. Hosker and her team are activists seeking to change archives’ descriptive language and practices to more accurately and inclusively represent the diverse populations for whom their collections are intended. Before we joined them as collaborators, they were discussing and making changes to the Archive’s descriptive language and practices. We are grateful for the willingness of Hosker and her team at the CRC to collaborate with us, bringing together the knowledge and practices of the archival and NLP communities to mitigate harms from biased language.

Appendix A Data Statement for Metadata Descriptions Extracted from the Archive's Online Catalog (version 1)

A.1 Curation Rationale

We (the research team) will use the extracted metadata descriptions to create a gold standard dataset annotated for contextual gender bias. We adopt Hitti et al.'s definition of contextual gender bias in text: written language that connotes or implies an inclination or prejudice against a gender through the use of gender-marked keywords and their context (2019, p. 10-11).

A member of our research team has extracted text from three descriptive metadata fields for all collections, subcollections, and items in the Archive's online catalog. One of these fields provide information about the people, time period, and places associated with the collection, subcollection, or item to which the field belongs. Another field summarizes the contents of the collection, subcollection, or item to which the field belongs. The last field records the person who wrote the text for the collection, subcollection, or item's descriptive metadata fields, and the date the person wrote the text.

Using the dataset of extracted text, we will experiment with training a discriminative classification algorithm to identify types of contextual gender bias. Additionally, the dataset will serve as a source of annotated, historical text to complement datasets composed of contemporary texts (i.e. from social media, Wikipedia, news articles).

To Do: We will group the metadata descriptions based on the collection to which they're associated, rather than segmenting by sentence or paragraph for annotation. Prior to making annotations for contextual gender bias, a member of our research team will review a subset of the metadata descriptions to determine whether all the descriptions should be annotated or whether the dataset should be filtered to include only a portion of the extracted metadata descriptions. Section B. in our data biography describes our plans for filtering.

We chose to use archival metadata descriptions as a data source because:

1. Metadata descriptions in the Archive's catalog (and most GLAM catalogs) are freely, publicly available online
2. GLAM metadata descriptions have yet to be analyzed at large scale using natural language processing (NLP) methods and, as records of cultural heritage, the descriptions have the potential to provide historical insights on changes in language and society (Welsh, 2016)
3. GLAM metadata standards are freely, publicly available, often online, meaning we can use historical changes in metadata standards used in the Archive to guide large-scale text analysis of changes in the language of the metadata descriptions over time
4. The Archive's policy acknowledges its responsibility to address legacy descriptions in its catalogs that use language considered biased or otherwise inappropriate today²¹

A.2 Language Variety

The metadata descriptions extracted from the Archive's catalog are written in British English.

A.3 Producer Demographic

We (the research team) are of American, German, and Scots nationalities, and are three females and one male. We all work primarily as academic researchers in the disciplines of natural language processing, data science, data visualization, human-computer interaction, digital humanities, and digital cultural heritage. Additionally, one of us is auditing an online course on feminist and social justice studies.

²¹The Archive is not alone; across the GLAM sector, institutions acknowledge and are exploring ways to address legacy language in their catalogs' descriptions. The "Note" in We Are What We Steal provides one example: <https://dxlab.sl.nsw.gov.au/we-are-what-we-steal/notes/>.

A.4 Annotator Demographic

For the research team who will write the annotation rule book, please refer to the previous section.

A gender, sexuality, and social justice studies expert based at a North American university will collaborate with us (the research team) on writing the annotation rule book. One member of our research team will annotate the metadata in collaboration with a second annotator.

Ongoing: we are seeking a second annotator with a background in gender studies, linguistics, or the information sciences; or with GLAM work experience.

A.5 Speech or Publication Situation

The metadata descriptions extracted from the Archive's online catalog using Open Access Initiative - Protocol for Metadata Harvesting (OAI-PMH). For OAI-PMH, an institution (in this case, the Archive) provides a URL to its catalog that displays its catalog metadata in XML format. A member of our research team wrote scripts in Python to extract three descriptive metadata fields for every collection, subcollection, and item in the Archive's online catalog (the metadata is organized hierarchically). Using Python and its Natural Language Toolkit (NLTK) library, the researcher removed duplicate sentences and calculated that the extracted metadata descriptions consist of a total of 966,763 words and 68,448 sentences across 1,231 collections. The minimum number of words in a collection is 7 and the maximum, 156,747, with an average of 1,306 words per collection and standard deviation of 7,784 words.

Please refer to the Provenance Appendix for information on the Speech or Publication Situation of all of the Archive's metadata descriptions.

A.6 Data Characteristics

Upon extracting the metadata descriptions using OAI-PMH, the XML tags were removed so that the total words and sentences of the metadata descriptions could be calculated to ensure the text source provided a sufficiently large dataset. A member of our research team has grouped all the extracted metadata descriptions by their collection (the "fonds" level in the XML data), preserving the context in which the metadata descriptions were written and will be read by visitors to the Archive's online catalog.

A.7 Data Quality

As a member of our research team extracts and filters metadata descriptions from the Archive's online catalog, they write assertions and tests to ensure as best as possible that metadata isn't being lost or unintentionally changed.

Please refer to the Provenance Appendix for information on the Data Quality of all of the Archive's metadata descriptions.

A.8 Other

Not applicable

A.9 Provenance Appendix

Data Statement for Metadata Descriptions from the Archive's Online Catalog (version 1)

Curation Rationale

The Archive's policy describes a commitment to develop collections that are as inclusive and diverse as possible, keeping up with social changes and looking for opportunities to better represent communities of people. Additionally, the Archive's policy states that the Archive aims to make its collections accessible to as many people as possible.

To Do: If available, review historical policy documents to understand how the Archive's curation rationale has evolved since its founding.

Language Variety

The Archive's metadata descriptions are written in British English.

Producer Demographic

People who write metadata descriptions to document the Archive's collections include employees, interns, and volunteers. Employees have received professional training in archival documentation, in addition to training at the Archive. Interns and volunteers are typically students studying information sciences, museology, history, or related disciplines who have also received training at the Archive. The Archive began in the 16th century, so the metadata descriptions in its online catalog date from that time period up through the present day (the Archive continues to collect and document cultural heritage records).

Additional demographic information on all those who have written the Archive's metadata descriptions is limited, however the Archive is based in the United Kingdom, meaning the perspectives of those who wrote the descriptions is most likely English, Irish, Scottish, British, or European. The Archive is closely associated with a research university, so interns and volunteers who write the Archive's metadata descriptions are likely to have received, or be in the process of receiving, higher education degrees.

Annotator Demographic

Not applicable

Speech or Publication Situation

The metadata descriptions in the Archive's online catalog document collections created by a university associated with the Archive and acquired or donated from other people and organizations. The Archive's earliest metadata descriptions were written in the 16th century; metadata descriptions continue to be written today.

The goal of the metadata descriptions is to help people find primary source material in the Archives. At the time most of the Archive's metadata descriptions were written, the descriptions were intended for employees of the Archive, who would help visitors locate primary source material. Circa 2015, employees of the Archive began writing metadata descriptions with visitors included in their intended audience.

Current employees at the Archives have stated that they would be happy for the metadata descriptions they write to be viewed as works in progress, because the Archive could never have enough time to document all its collection items completely. Moreover, often information about collections items is impossible to know due to their historical nature and lack of accompanying documentation, so the metadata descriptions will always be incomplete.

The metadata descriptions include information available from the cultural heritage records they describe, from any available documentation that accompanied those records when the Archive acquired them, from authorities such as the Library of Congress Subject Headings, and from other documentation resources considered trustworthy among archives (a more extensive list is provided here).

Data Characteristics

Beginning circa 2017, people documenting collections in the Archive have written metadata descriptions according to the General International Standard Archival Description (ISAD(G)). Past metadata descriptions were written according to library metadata standards. Metadata descriptions may include contextual information about the people, places, and time periods relevant to the collection items, as well as the date a description was written and who wrote the description. Though all of this descriptive information ideally exists for a collection item, some collection items do not have this complete of a description.

To Do: If possible, determine which library metadata standards were used for documentation prior to 2017.

Data Quality

The metadata descriptions in the Archive's online catalog consists of manually entered data, some of which was initially written in digital form, and some of which was initially written on paper and has since been manually typed into digital form.

To Do: Determine how much the metadata descriptions are born-digital versus re-written digitally, and when the Archive transitioned from writing metadata descriptions on paper to writing metadata descriptions digitally (typing manually).

Other

Not applicable

Provenance Appendix

None

Appendix B Data Biography for Metadata Descriptions Extracted from the Archive's Online Catalog (version 1)

B.1 Dataset

Metadata descriptions from the Archive's online catalog

B.2 Where was the Data Collected or Created?

We (the research team) collected the data using the Open Access Initiative - Protocol for Metadata Harvesting (OAI-PMH).

Employees, interns, and volunteers at the Archive who wrote the metadata descriptions collected information to include in the descriptions from documentation accompanying the cultural heritage record(s) they were describing, from the cultural heritage records themselves, from authorities such as Library of Congress Subject Headings, and from other trusted sources for archival documentation. Examples of other trusted sources are available here.

Where possible, we will use dates associated with the descriptions to contextualize their text in relation to historical changes in metadata structures. For example, the metadata standard Library of Congress Subject Headings (LCSH) once used the term "Jewish Question" instead of the current term "Jews," so GLAM who use LCSH may have descriptions in their catalogs that use the historical term now considered biased. After historical analysis of metadata standards the Archive uses, we will filter our collected text to include those that reference groups of people who have historically been described stereotypically.

B.3 Who Collected or Created the Data?

The Archive and the university to which it is associated collected some of the cultural heritage records and the accompanying documentation that informs the records' metadata descriptions. For other cultural heritage records and their accompanying documentation, individual collectors gathered the records and wrote their documentation, which employees, interns, and volunteers used to write descriptive metadata for the records in the Archive's catalog.

The Archive has existed since the 16th century, so its directors will each have established different policies and goals for acquiring and documenting cultural heritage records. The latest policy document for the Archive includes a statement about diversity, inclusion and accessibility that describes the Archive's commitment to providing representative collections for local, national, and international audiences.

B.4 Why was the Data Collected or Created?

The Archive's policy explains that it documents cultural heritage records in its catalog so that researchers can find the records and use them as primary source material to guide their work. Current employees of the Archive reiterated the goal of discoverability as the main reason for writing metadata descriptions.

Individuals and institutions who have donated their collections to the Archive had personal reasons motivating their choices of records to save. A directory of the Archive's collections contains information

about select individuals and institutions that suggest their reasons for saving the records they did. Information in the metadata descriptions themselves may also provide insight on why their associated records were collected.

B.5 When was the Data Collected or Created?

Among the metadata descriptions we extracted that include a year documenting when they were written, the years show that the descriptions were written from the 19th century up through the 21st century. Further research is needed to determine how early the extracted metadata descriptions without a year were written.

Appendix C Stakeholder Power Relations in NLP Research on Bias in Archival Metadata Descriptions (version 1)

C.1 The Stakeholders

Identification:

1. Us as the research team
2. Employees of the Archive (current and former) who wrote the metadata descriptions that serve as this research's text source
3. The Archive and its associated university as institutions that provide access to the metadata descriptions
4. People represented in the metadata descriptions
5. Visitors to the Archive, as they will read the metadata descriptions used as this research's text source when using the Archive's online catalog

Limitations: Due to the length of the text and the historical nature of the metadata descriptions we use from the Archive's catalog, we do not have access to every person represented in the metadata descriptions. However, the Archive does have a take-down policy that we will follow with our text source to respect the people represented in metadata descriptions as best as possible: if a person requests that information about them or someone they are connected to be removed from or anonymized in the catalog, the Archive will comply. To the best of our ability, we will make sure that the metadata descriptions we use as the text source for our research do not include information that a visitor has requested the Archive take down.

C.2 Power Relations Questions

Who or what is included in the research?

Who:

- Current employees of the Archive: To account for intragroup differences, we include employees with different years of experience and employees working in several positions within the hierarchy of job roles in the Archive.
- Us (the research team): The size of the team is small enough that all members are included, meaning intragroup differences are accounted for by default.

To Do: Find visitors to the Archive who I can speak to about their experience reading its catalog's metadata descriptions. To account for intragroup differences among visitors, we will seek out a selection of visitors with as diverse of identity characteristics as possible.

What: Ongoing work includes conducting historical research to understand the context in which the metadata descriptions were written. For example, employees at the Archive stated that for many

years, people wrote metadata descriptions with the aim of being as neutral and objective as possible, however the latest generation of archivists is challenging this, arguing that neutrality isn't possible and encouraging transparency instead.

Who or what is excluded from the research?

Who:

- Past employees of the Archive
- People represented in the Archive's cultural heritage records
- The majority of the Archive's visitors (the research only has the capacity to include a selection of visitors in user research and participatory action research activities)

What: The historical context of metadata descriptions written before my lifetime

To Do: Determine if policy guidelines for the Archive since its beginnings in the 16th century are available to understand how it perceived itself and what drove its collection and documentation practices. Otherwise, the historical existence of the Archive is also excluded from the research.

How will the research define knowledge?

The research will define knowledge as multifaceted. We (the research team) will draw on the disciplines of gender studies and linguistics to manually identify and annotate types of contextual gender bias in metadata descriptions. The research will share the annotated dataset as one interpretation of gender bias, recognizing that different people have different experiences of oppression that cause variations in attitudes towards words or phrases.

We will use the annotated dataset to train a discriminative classification algorithm. The types of gender bias that the algorithm identifies will be presented as potentially biased text, requiring verification from a person working with the text to decide whether the text should be considered biased.

Who has agency and who can be empowered?

We (the research team) have agency as the people applying NLP methods to the Archive's metadata descriptions.

The employees of the Archive can be empowered through participatory action research, with collaborative activities in which we situate the employees as partners in the research and as experts on archival practices and metadata.

The employees of the Archive have determined that people who do not identify as male are under-represented in the Archive's collections and thus those collections' metadata descriptions. We focus our bias identification and classification efforts on gender bias to explore how we can empower people who do not identify as male through the process and outputs of our NLP research.

To Do: Provide examples of how our research process and outputs empowers people who do not identify as male.

References

- Melissa Adler. 2017. Introduction: A Book is Being Cataloged. In *Cruising the Library: Perversities in the Organization of Knowledge*, pages 1–26. Fordham University Press.
- Angel, Christine M., and Caroline Fuchs, editor. 2018. *Organization, representation and description through the digital age: information in libraries, archives and museums*. Walter de Gruyter GmbH, Berlin; Boston.
- Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604, December.

- Su Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of “Bias” in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.
- Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 4356–4364.
- Mary Bucholtz. 2003. Theories of Discourse as Theories of Gender: Discourse Analysis in Language and Gender Studies. In *The Handbook of Language and Gender*, pages 43–68, Oxford, GB, January. Blackwell Publishing Ltd.
- Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, April.
- Arlene Casey, Mike Bennett, Richard Tobin, Claire Grover, Iona Walker, Lukas Engelmann, and Beatrice Alex. 2020. Plague Dot Text: Text mining and annotation of outbreak reports of the Third Plague Pandemic (1894–1952). *Computing Research Repository*, arXiv:2002.01415:23.
- Michelle Caswell and Marika Cifor. 2019. Neither a Beginning Nor an End: Applying an Ethics of Care to Digital Archival Collections. In *The Routledge International Handbook of New Digital Practices in Galleries, Libraries, Archives, Museums and Heritage Sites*, pages 159–168. Routledge, November.
- Kate Crawford. 2017. The Trouble with Bias. In *Neural Information Processing Systems Conference Keynote*. [Online; accessed 10-July-2020].
- Kimberlé Crenshaw. 1991. Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. *Stanford Law Review*, 43(6):1241–1299.
- Mark Diaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing Age-Related Bias in Sentiment Analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, pages 1–14, Montréal, CA. ACM Press.
- Catherine D’Ignazio and Lauren F. Klein. 2020. *Data Feminism*. Strong ideas series. The MIT Press, Cambridge, US.
- Batya Friedman and Helen Nissenbaum. 1996. Bias in Computer Systems. *ACM Transactions on Information Systems*, 14(3):330–347, June.
- Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, April.
- Aparna Garimella, Carmen Banea, Dirk Hovy, and Rada Mihalcea. 2019. Women’s Syntactic Resilience and Men’s Grammatical Luck: Gender-Bias in Part-of-Speech Tagging and Dependency Parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3493–3498, Florence, IT. Association for Computational Linguistics.
- Ilka H. Gleibs. 2017. Are all “research fields” equal? Rethinking practice for the use of data from crowdsourcing market addresss. *Behavior Research Methods*, 49(4):1333–1342, August.
- Hila Gonen and Yoav Goldberg. 2019. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. *NAACL 2019*, arXiv:1903.03862v2, September.
- Elizabeth L. Haines, Kay Deaux, and Nicole Lofaro. 2016. The Times They Are a-Changing . . . or Are They Not? A Comparison of Gender Stereotypes, 1983–2014. *Psychology of Women Quarterly*, 40(3):353–363, September.
- Vicki L. Hanson, Anna Cavender, and Shari Trewin. 2015. Writing about accessibility. *Interactions*, 22(6):62–65, October.
- Donna Haraway. 1988. Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies*, 14(3):575.
- Sandra Harding. 1995. “Strong objectivity”: A response to the new objectivity question. *Synthese*, 104(3), September.
- Tessa Hauswedell, Julianne Nyhan, Melodee H. Beals, Melissa Terras, and Emily Bell. 2020. Of global reach yet of situated contexts: an examination of the implicit and explicit selection criteria that shape digital archives of historical newspapers. *Archival Science*, 20(2):139–165, June.

- Yasmeen Hitti, Eunbee Jang, Ines Moreno, and Carlyne Pelletier. 2019. Proposed Taxonomy for Gender Bias in Text: A Filtering Methodology for the Gender Generalization Subtype. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 8–17, Florence, IT. Association for Computational Linguistics.
- Masahiro Kaneko and Danushka Bollegala. 2019. Gender-preserving Debiasing for Pre-trained Word Embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1641–1650, Florence, IT. Association for Computational Linguistics.
- Ansgar Koene, Elvira Perez, Sofia Ceppi, Michael Rovatsos, Helena Webb, Menisha Patel, Marina Jirotko, and Giles Lane. 2017. Algorithmic Fairness in Online Information Mediating Systems. In *Proceedings of the 2017 ACM on Web Science Conference, WebSci '17*, page 391–392, New York, US. Association for Computing Machinery.
- Heather Krause. 2019. An Introduction to the Data Biography. *We All Count*. [Online; accessed 17-October-2020].
- Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring Bias in Contextualized Word Representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, IT. Association for Computational Linguistics.
- Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1, ETMTNLP '02*, pages 63–70, US. Association for Computational Linguistics.
- Bella Martin and Bruce Hanington. 2012. 11 Case studies. In *Universal Methods of Design: 100 Ways to Research Complex Problems, Develop Innovative Ideas, and Design Effective Solutions*, Beverly, US. Rockport Publishers.
- Tara McPherson. 2012. Why are the Digital Humanities So White? or Thinking the Histories of Race and Computation. In *Debates in the Digital Humanities*, pages 139–160, Minneapolis, US. University of Minnesota Press.
- Niamh Moore. 2018. A cat’s cradle of feminist and other critical approaches to participatory research. In *Connected Communities Foundation Series*, Bristol, UK, September. University of Bristol/AHRC Connected Communities Programme. [Online; accessed 24-July-2020].
- Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press, New York, US.
- Oxford English Dictionary. 2013a. Classism. In *OED Online*. Oxford University Press, June. [Online; accessed 21-August-2020].
- Oxford English Dictionary. 2013b. Discourse. In *OED Online*. Oxford University Press, December. [Online; accessed 17-October-2020].
- Oxford English Dictionary. 2013c. Racism. In *OED Online*. Oxford University Press, June. [Online; accessed 21-August-2020].
- Oxford English Dictionary. 2013d. Sexism. In *OED Online*. Oxford University Press, June. [Online; accessed 21-August-2020].
- Orestis Papakyriakopoulos, Simon Hegelich, Juan Carlos Medina Serrano, and Fabienne Marco. 2020. Bias in Word Embeddings. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT*’20*, pages 446–457, New York, NY. Association for Computing Machinery.
- Caroline Criado Perez. 2019. *Invisible Women: Exposing Data Bias in a World Designed for Men*. Vintage, London, GB.
- Colleen Reid and Wendy Frisby. 2008. 6 Continuing the Journey: Articulating Dimensions of Feminist Participatory Action Research (FPAR). In *The SAGE Handbook of Action Research*, pages 93–105. SAGE Publications Ltd, February.
- Roopika Risam. 2015. Beyond the Margins: Intersectionality and the Digital Humanities. *Digital Humanities Quarterly*, 9(2):14.

- Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating Gender Bias in Machine Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, IT. Association for Computational Linguistics.
- Marja Liisa Swantz. 2008. 2 Participatory Action Research as Practice. In *The SAGE Handbook of Action Research*, pages 31–48. SAGE Publications Ltd.
- Latanya Sweeney. 2013. Discrimination in online ad delivery. *Communications of the ACM*, 56(5):44–54, May.
- Nathaniel Swinger, Maria De-Arteaga, Neil Thomas Heffernan IV, Mark DM Leiserson, and Adam Tauman Kalai. 2019. What are the Biases in My Word Embedding? In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 305–311, Honolulu, US, January. Association for Computing Machinery.
- Mary Talbot. 2003. Gender Stereotypes: Reproduction and Challenge. In *The Handbook of Language and Gender*, pages 468–486, Oxford, GB, January. Blackwell Publishing Ltd.
- Sigal Vainapel, Opher Y. Shamir, Yulie Tenenbaum, and Gadi Gilam. 2015. The dark side of gendered language: The masculine-generic form as a cause for self-report bias. *Psychological Assessment*, 27(4):1513–1519.
- Theo van Leeuwen. 2009. Discourse as the Recontextualization of Social Practice: A Guide. In *Methods for Critical Discourse Analysis*. SAGE Publications.
- Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns. *Computing Research Repository*, arXiv:1810.05201, October.
- Anne Welsh. 2016. The Rare Books Catalog and the Scholarly Database. *Cataloging & Classification Quarterly*, 54(5–6):317–337, aug.
- Haoran Zhang, Amy X. Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. 2020. Hurtful words: quantifying biases in clinical contextual word embeddings. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 110–120, Toronto, CA, April. Association for Computing Machinery.
- Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, US. Association for Computational Linguistics.