Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science

Emily M. Bender
Department of Linguistics
University of Washington
ebender@uw.edu

Batya Friedman
The Information School
University of Washington
batya@uw.edu

Abstract

In this paper, we propose data statements as a design solution and professional practice for natural language processing technologists, in both research and development. Through the adoption and widespread use of data statements, the field can begin to address critical scientific and ethical issues that result from the use of data from certain populations in the development of technology for other populations. We present a form that data statements can take and explore the implications of adopting them as part of regular practice. We argue that data statements will help alleviate issues related to exclusion and bias in language technology, lead to better precision in claims about how NLP research can generalize and thus better engineering results, protect companies from public embarrassment, and ultimately lead to language technology that meets its users in their own preferred linguistic style and furthermore does not misrepresent them to others.

1 Introduction

As technology enters widespread societal use it is important that we, as technologists, think critically about how the design decisions we make and systems we build impact people—including not only users of the systems but also other people who will be affected by the systems without directly interacting with them. For this paper, we focus on natural language processing (NLP) technology. Potential adverse impacts include NLP systems that fail to work for specific subpopulations (e.g., children or speakers of language varieties that are not supported by training or test data) or systems that reify and reinforce biases present in training data (e.g., a resume-review system that ranks female candidates as less qualified for computer programming jobs because of biases present in training text).

There are both scientific and ethical reasons to be concerned. Scientifically, there is the issue of generalizability of results; ethically, the potential for significant real-world harms. Although there is increasing interest in ethics in NLP,[1] there remains the open and urgent question of how we integrate ethical considerations into the everyday practice of our field. This question has no simple answer, but rather will require a constellation of multi-faceted solutions.

Toward that end, and drawing on value sensitive design (Friedman et al., 2006), this paper contributes one new professional practice—called **data statements**—which we argue will bring about improvements in engineering and scientific outcomes while also enabling more ethically responsive NLP technology. A data statement is a characterization of a data set 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. In developing this practice, we draw on analogous practices from the fields of psychology and medicine that require some standardized information about the populations studied (e.g., APA, 2009; Moher et al., 2010; Furler et al., 2012; Mbuagbaw et al., 2017). Though the construct of data statements applies more broadly, in this paper we focus specifically on data statements for NLP systems. Data statements should be included in most writing on NLP including: papers presenting new datasets, papers reporting experimental work with datasets, and documentation for NLP systems. Data statements should

[1] This interest has manifested in workshops (Fort et al., 2016; Devillers et al., 2016; Hovy et al., 2017) and papers (Hovy and Spruit, 2016) in NLP, as well as workshops in related fields, notably the FATML series (http://www.fatml.org/) held annually since 2014.

help us as a field engage with the ethical issues of *exclusion*, *overgeneralization*, and *underexposure* (Hovy and Spruit, 2016). Furthermore, as data statements bring our datasets and their represented populations into better focus, they should also help us as a field deal with scientific issues of *generalizability* and *reproducibility*. Adopting this practice will position us to better understand and describe our results and, ultimately, do better and more ethical science and engineering.[2]

We begin by defining terms (§2), discuss why NLP needs data statements (§3), and relate our proposal to current practice (§4). Next is the substance of our contribution: a detailed proposal for data statements for NLP (§5), illustrated with two case studies (§6). In §7 we discuss how data statements can mitigate bias and use the technique of "value scenarios" to envision potential effects of their adoption. Finally, we relate data statements to similar emerging proposals (§8), make recommendations for how to implement and promote the uptake of data statements (§9), and lay out considerations for tech policy (§10).

2 Definitions

As this paper is intended for at least two distinct audiences (NLP technologists and tech policymakers), we use this section to briefly define key terms.

Dataset, Annotations An **(NLP) dataset** is a collection of speech or writing possibly combined with **annotations**.[3] Annotations include indications of linguistic structure like part of speech tags or syntactic parse trees, as well as labels classifying aspects of what the speakers were attempting to accomplish with their utterances. The latter includes annotations for sentiment (Liu, 2012) and for figurative language or sarcasm (e.g., Riloff et al., 2013; Ptáček et al., 2014). Labels can be naturally occurring, such as star ratings in reviews taken as indications of the overall sentiment of the review (e.g., Pang et al., 2002) or the hashtag *#sarcasm* used to identify sarcastic language (e.g., Kreuz and Caucci, 2007).

Speaker We use the term **speaker** to refer to the individual who produced some segment of linguistic behavior included in the dataset, even if the linguistic behavior is originally written.

Annotator The term **annotator** refers to people who assign annotations to the raw data, including transcribers of spoken data. Annotators may be crowdworkers or highly trained researchers, sometimes involved in the creation of the annotation guidelines. Annotation is often done semi-automatically, with NLP tools being used to create a first pass that is corrected or augmented by human annotators.

Curator A third role in dataset creation, less commonly discussed, is the **curator**. Curators are involved in the selection of which data to include, by selecting individual documents, by creating search terms that generate sets of documents, by selecting speakers to interview and designing interview questions, and so forth.

Stakeholders **Stakeholders** are people impacted directly or indirectly by a system (Friedman et al., 2006; Czeskis et al., 2010). **Direct stakeholders** include those who interact with the system, either by participating in system creation (developers, speakers, annotators and curators) or by using it. **Indirect stakeholders** do not use the system but are nonetheless impacted by it. For example, people whose Web content is displayed or rendered invisible by search engine algorithms are indirect stakeholders with respect to those systems.

Algorithm We use the term **algorithm** to encompass both rule-based and machine learning approaches to NLP. Some algorithms (typically rule-based ones) are tightly connected to the datasets they are developed against. Other algorithms can be easily ported to different datasets.[4]

System We use the term **(NLP) system** to refer to a piece of software that does some kind

[2] By arguing here that data statements promote both ethical practice and sound science, we do not mean to suggest that these two can be conflated. A system can give accurate responses as measured by some test set (scientific soundness) and yet lead to real-world harms (ethical issues). Accordingly, it is up to researchers and research communities to engage with both scientific and ethical ideals.

[3] Multi-modal data sets combine language and video or other additional signals. Here, our focus is on linguistic data.

[4] Datasets used during algorithm development can influence design choices in machine learning approaches too: Munro and Manning (2010) found that subword information, not helpful in English SMS classification, is extremely valuable in Chichewa, a morphologically complex language with high orthographic variability.

of natural language processing, typically involving algorithms trained on particular datasets. We use this term to refer to both components focused on specific tasks (e.g., the Stanford parser (Klein and Manning, 2003) trained on the Penn Treebank (Marcus et al., 1993) to do English parsing) and user-facing products such as Amazon's Alexa or Google Home.

Bias We use the term **bias** to refer to cases where computer systems "*systematically* and *unfairly discriminate* against certain individuals or groups of individuals in favor of others" (Friedman and Nissenbaum, 1996, page 332).[5] To be clear: (i) unfair discrimination does not give rise to bias unless it occurs systematically and (ii) systematic discrimination does not give rise to bias unless it results in an unfair outcome. Friedman and Nissenbaum (1996) show that in some cases, system bias reflects biases in society; these are **pre-existing biases** with roots in social institutions, practices and attitudes. In other cases, reasonable, seemingly neutral, technical elements (e.g., the order in which an algorithm processes data) can result in bias when used in real world contexts; these **technical biases** stem from technical constraints and decisions. A third source of bias, **emergent bias**, occurs when a system designed for one context is applied in another (e.g., with a different population).

3 Why Does NLP Need Data Statements?

Recent studies have documented the fact that limitations in training data lead to ethically problematic limitations in the resulting NLP systems. Systems trained on naturally occurring language data learn the pre-existing biases held by the speakers of that data: Typical vector-space representations of lexical semantics pick up cultural biases about gender (Bolukbasi et al., 2016) and race, ethnicity, and religion (Speer, 2017). Zhao et al. (2017) show that beyond picking up such biases, machine learning algorithms can amplify them. Furthermore, these biases, far from being inert or simply a reflection of the data, can have real-world consequences for both direct and indirect stakeholders. For example, Speer (2017) found that a sentiment analysis system rated reviews of Mexican restaurants as more negative than other types of food with similar star ratings, because of associations between the word *Mexican* and words with negative sentiment in the larger corpus on which the word embeddings were trained. (See also Kiritchenko and Mohammad, 2018.) In these and other ways, pre-existing biases can be trained into NLP systems. There are other studies showing that systems from part of speech taggers (Hovy and Søgaard, 2015; Jørgensen et al., 2015 to speech recognition engines (Tatman, 2017) perform better for speakers whose demographic characteristics better match those represented in the training data. These are examples of emergent bias.

Because the linguistic data we use will always include pre-existing biases and because it is not possible to build an NLP system in such a way that it is immune to emergent bias, we must seek additional strategies for mitigating the scientific and ethical shortcomings that follow from imperfect datasets. We propose here that foregrounding the characteristics of our datasets can help, by allowing reasoning about what the likely effects may be and by making it clearer which populations are and are not represented, for both training and test data. For training data, the characteristics of the dataset will affect how the system will work when it is deployed. For test data, the characteristics of the dataset will affect what can be measured about system performance and thus provides important context for scientific claims.

4 Current Practice and Challenges

Typical current practice in academic NLP is to present new datasets with a careful discussion of the annotation process as well as a brief characterization of the genre (usually by naming the underlying data source) and the language. NLP papers using datasets for training or test data tend to more briefly characterize the annotations and will sometimes leave out mention of genre and even language.[6] Initiatives such as the Open Language Archives Community (OLAC; Bird and Simon

[5] The machine learning community uses the term *bias* to refer to constraints on what an algorithm can learn, which may prevent it from picking up patterns in a dataset or lead it to relevant patterns more quickly (see Coppin 2004, Ch. 10). This use of the term does not carry connotations of unfairness.

[6] Surveys of EACL 2009 (Bender, 2011) and ACL 2015 (Munro, 2015) found 33–81% of papers failed to name the language studied. (It always appeared to be English.)

2000), the Fostering Language Resources Network (FLaReNet; Calzolari et al., 2012) and the Text Encoding Initiative (TEI; Consortium, 2008) prescribe metadata to publish with language resources, primarily to aid in the discoverability of such resources. FLaReNet also encourages documentation of language resources. And yet, it is very rare to find detailed characterization of the speakers whose data is captured or the annotators who provided the annotations, though the latter are usually characterized as being experts or crowdworkers.[7]

To fill this information gap, we argue that data statements should be included in every NLP publication that presents new datasets and in the documentation of every NLP system, as part of a chronology of system development including descriptions of the various datasets for training, tuning, and testing. Data statements should also be included in all NLP publications reporting experimental results. Accordingly, data statements will need to be both detailed and concise. To meet these competing goals, we propose two variants. For each dataset there should be a long-form version in an academic paper presenting the dataset or in system documentation. Research papers presenting experiments making use of datasets with existing long-form data statements should include shorter data statements and cite the longer one.[8]

We note another set of goals in competition: Although readers need as much information as possible in order to understand how the results can and cannot be expected to generalize, considerations of the privacy of the people involved (speakers, annotators) might preclude including certain kinds of information, especially with small groups. Each project will need to find the right balance, but this can be addressed in part by asking annotators and speakers for permission to collect and publish such information.

5 Proposed Data Statement Schema

We propose the following schema of information to include in long and short form data statements.

[7] A notable exception is Derczynski et al. (2016), who present a corpus of tweets collected to sample diverse speaker communities (location, type of engagement with Twitter), at diverse points in time (time of year, month, and day), and annotated with named entity labels by crowdworker annotators from the same locations as the tweet authors.

[8] Older datasets can be retrofitted with citeable long-form data statements published on project Web pages or archives.

5.1 Long Form

Long form data statements should be included in system documentation and in academic papers presenting new datasets, and should strive to provide the following information:

A. CURATION RATIONALE Which texts were included and what were the goals in selecting texts, both in the original collection and in any further sub-selection? This can be especially important in datasets too large to thoroughly inspect by hand. An explicit statement of the curation rationale can help dataset users make inferences about what other kinds of texts systems trained with them could conceivably generalize to.

B. LANGUAGE VARIETY Languages differ from each other in structural ways that can interact with NLP algorithms. Within a language, regional or social dialects can also show great variation (Chambers and Trudgill, 1998). The language and language variety should be described with:

- A language tag from BCP-47[9] identifying the language variety (e.g., en-US or yue-Hant-HK)
- A prose description of the language variety, glossing the BCP-47 tag and also providing further information (e.g. English as spoken in Palo Alto CA (USA) or Cantonese written with traditional characters by speakers in Hong Kong who are bilingual in Mandarin)

C. SPEAKER DEMOGRAPHIC Sociolinguistics has found that variation (in pronunciation, prosody, word choice, and grammar) correlates with speaker demographic characteristics (Labov, 1966), as speakers use linguistic variation to construct and project identities (Eckert and Rickford, 2001). Transfer from native languages (L1) can affect the language produced by non-native (L2) speakers (Ellis, 1994, Ch. 8). A further important type of variation is disordered speech (e.g., dysarthria). Specifications include:

- Age
- Gender
- Race/ethnicity
- Native language
- Socioeconomic status
- Number of different speakers represented
- Presence of disordered speech

[9] https://tools.ietf.org/rfc/bcp/bcp47.txt.

D. ANNOTATOR DEMOGRAPHIC What are the demographic characteristics of the annotators and annotation guideline developers? Their own "social address" influences their experience with language and thus their perception of what they are annotating. Specifications include:

- Age
- Gender
- Race/ethnicity
- Native language
- Socioeconomic status
- Training in linguistics/other relevant discipline

E. SPEECH SITUATION Characteristics of the speech situation can affect linguistic structure and patterns at many levels. The intended audience of a linguistic performance can also affect linguistic choices on the part of speakers.[10] The time and place provide broader context for understanding how the texts collected relate to their historical moment and should also be made evident in the data statement.[11] Specifications include:

- Time and place
- Modality (spoken/signed, written)
- Scripted/edited vs. spontaneous
- Synchronous vs. asynchronous interaction
- Intended audience

F. TEXT CHARACTERISTICS Both genre and topic influence the vocabulary and structural characteristics of texts (Biber, 1995), and should be specified.

G. RECORDING QUALITY For data that include audiovisual recordings, indicate the quality of the recording equipment and any aspects of the recording situation that could impact recording quality.

H. OTHER There may be other information of relevance as well (e.g., the demographic characteristics of the curators). As stated earlier, this is intended as a starting point and we anticipate best practices around writing data statements to develop over time.

I. PROVENANCE APPENDIX For datasets built out of existing datasets, the data statements for the source datasets should be included as an appendix.

[10]For example, people speak differently to close friends vs. strangers, to small groups vs. large ones, to children vs. adults and to people vs. machines (e.g., Ervin-Tripp, 1964).

[11]Mutable speaker demographic information, such as age, is interpreted as relative to the time of the linguistic behavior.

5.2 Short Form

Short form data statements should be included in any publication using a dataset for training, tuning, or testing a system and may also be appropriate for certain kinds of system documentation. The short form data statement does not replace the long form one, but rather should include a pointer to it. For short form data statements, we envision 60–100 word summaries of the description included in the long form, covering most of the main points.

5.3 Summary

We have outlined the kind of information data statements should include, addressing the needs laid out in §3, describing both long and short versions. As the field gains experience with data statements, we expect to see a better understanding of what to include as well as best practices for writing data statements to emerge.

Note that full specification of all of this information may not be feasible in all cases. For example, in datasets created from Web text, precise demographic information may be unavailable. In other cases (e.g., to protect the privacy of annotators) it may be preferable to provide ranges rather than precise values. For the description of demographic characteristics, our field can look to others for best practices, such as those described in the American Psychological Association's *Manual of Style*.

It may seem redundant to reiterate this information in every paper that makes use of well-trodden datasets. Nonetheless, it is critical to consider the data anew each time to ensure that it is appropriate for the NLP work being undertaken and that the results reported are properly contextualized. Note that the requirement is not that datasets be used only when there is an ideal fit between the dataset and the NLP goals but rather that the characteristics of the dataset be examined in relation to the NLP goals and limitations be reported as appropriate.

6 Case Studies

We illustrate the idea of data statements with two cases studies. Ideally, data statements are written at or close to the time of dataset creation. These data statements were constructed post hoc in conversation with the dataset curators. The first entails labels for a particular subset of all Twitter

data. In contrast, the second entails all available data for an intentionally generated interview collection, including audiofiles and transcripts. Both illustrate how even when specific information is not available, the explicit statement of its lack of availability provides a more informative picture of the dataset.

6.1 Hate Speech Twitter Annotations

The Hate Speech Twitter Annotations collection is a set of labels for ∼19,000 tweets collected by Waseem and Hovy (2016) and Waseem (2016). The dataset can be accessed via `https://github.com/zeerakw/hatespeech`.[12]

A. CURATION RATIONALE In order to study the automatic detection of hate speech in tweets and the effect of annotator knowledge (crowdworkers vs. experts) on the effectiveness of models trained on the annotations, Waseem and Hovy (2016) performed a scrape of Twitter data using contentious terms and topics. The terms were chosen by first crowdsourcing an initial set of search terms on feminist Facebook groups and then reviewing the resulting tweets for terms to use and adding others based on the researcher's intuition.[13] Additionally, some prolific users of the terms were chosen and their timelines collected. For the annotation work reported in Waseem (2016), expert annotators were chosen for their attitudes with respect to intersectional feminism in order to explore whether annotator understanding of hate speech would influence the labels and classifiers built on the dataset.

B. LANGUAGE VARIETY The data was collected via the Twitter search API in late 2015. Information about which varieties of English are represented is not available, but at least Australian (en-AU) and US (en-US) mainstream Englishes are both included.

C. SPEAKER DEMOGRAPHIC Speakers were not directly approached for inclusion in this dataset and thus could not be asked for demographic information. More than 1,500 different Twitter accounts are included. Based on independent information about Twitter usage and impressionistic observation of the tweets by the dataset curators, the data is likely to include tweets from both younger (18–30 years) and older (30+ years) adult speakers, the majority of whom likely identify as white. No direct information is available about gender distribution or socioeconomic status of the speakers. It is expected that most, but not all, of the speakers speak English as a native language.

D. ANNOTATOR DEMOGRAPHIC This dataset includes annotations from both crowdworkers and experts. A total of 1,065 crowdworkers were recruited through Crowd Flower, primarily from Europe, South America, and North America. Beyond country of residence, no further information is available about the crowdworkers. The expert annotators were recruited specifically for their understanding of intersectional feminism. All were informally trained in critical race theory and gender studies through years of activism and personal research. They ranged in age from 20–40 years, included 3 men and 13 women, and gave their ethnicity as white European (11), East Asian (2), Middle East/Turkey (2), and South Asian (1). Their native languages were Danish (12), Danish/English (1), Turkish/Danish (1), Arabic/Danish (1), and Swedish (1). Based on income levels, the expert annotators represented upper lower class (5), middle class (7), and upper middle class (2).

E. SPEECH SITUATION All tweets were initially published between April 2013 and December 2015. Tweets represent informal, largely asynchronous, spontaneous, written language, of up to 140 characters per tweet. About 23% of the tweets were in reaction to a specific Australian TV show (*My Kitchen Rules*) and so were likely meant for roughly synchronous interaction with other viewers. The intended audience of the tweets was either other viewers of the same show, or simply the general Twitter audience. For the tweets containing racist hate speech, the authors appear to intend them both for those who would agree but also for people whom they hope to provoke into having an agitational and confrontational exchange.

F. TEXT CHARACTERISTICS For racist tweets the topic was dominated by Islam and Islamophobia.

[12] This data statement was prepared based on information provided by Zeerak Waseem, personal conversation, Feb–Apr 2018 and reviewed and approved by him.

[13] In a standalone data statement, the search terms should be given in the main text. To avoid accosting readers with slurs in this article, we instead list them in this footnote. Waseem and Hovy (2016) provide the following complete list of terms used in their initial scrape: MKR, asian drive, feminazi, immigrant, nigger, sjw, WomenAgainstFeminism, blameonenotall, islam terrorism, notallmen, victimcard, victim card, arab terror, gamergate, jsil, racecard, race card.

For sexist tweets predominant topics were the TV show and people making sexist statements while claiming not to be sexist. The majority of tweets only used one modality (text) though some included links to pictures and Web sites.

G. RECORDING QUALITY N/A.

H. OTHER N/A.

I. PROVENANCE APPENDIX N/A.

Twitter Hate Speech Short Form This dataset includes labels for ∼19,000 English tweets from different locales (Australia and North America being well represented) selected to contain a high prevalence of hate speech. The labels indicate the presence and type of hate speech and were provided both by experts (mostly with extensive if informal training in critical race theory and gender studies and English as a second language) and by crowdworkers primarily from Europe and the Americas. [Include a link to the long form.]

6.2 Voices from the Rwanda Tribunal (VRT)

Voices from the Rwanda Tribunal is a collection of 49 video interviews in English and French with personnel from the International Criminal Tribunal for Rwanda (ICTR) comprising 50–60 hours of material with high quality transcription throughout (Nathan et al., 2011; Nilsen et al., 2012; Friedman et al., 2016). The dataset can be downloaded from `http://www.tribunalvoices.org`.[14]

A. CURATION RATIONALE The VRT project, funded by the United States National Science Foundation, is part of a research program on developing multi-lifespan design knowledge (Friedman and Nathan, 2010). It is independent from the ICTR, the United Nations, and the government of Rwanda. To help ensure accuracy and guard against breeches of confidentiality, interviewees had an opportunity to review and redact any material that was either misspoken or revealed confidential information. A total of two words have been redacted. No other review or redaction of content has occurred. The dataset includes all publicly released material from the collection; as of the writing of this data statement (28 September 2017) one interview and a portion of a second are currently sealed.

B. LANGUAGE VARIETY Of the interviews, 44 are conducted in English (en-US and international English on the part of the interviewees, en-US on the part of the interviewers) and 5 in French and English, with the interviewee speaking international French, the interviewer speaking English (en-US), and an interpreter speaking both.[15]

C. SPEAKER DEMOGRAPHIC The interviewees (13 women and 36 men, all adults) are professionals working in the area of international justice, such as judges or prosecutors, and support roles of the same, such as communications, prison warden, and librarian. They represent a variety of nationalities: Argentina, Benin, Cameroon, Canada, England, The Gambia, Ghana, Great Britain, India, Italy, Kenya, Madagascar, Mali, Morocco, Nigeria, Norway, Peru, Rwanda, Senegal, South Africa, Sri Lanka, St. Kitts and Nevis, Sweden, Tanzania, Togo, Uganda, and the US. Their native languages are not known, but are presumably diverse. The 7 interviewers (2 women and 5 men) are informataion and legal professionals from different regions in the US. All are native speakers of US English, all are white, and at the time of the interviews they ranged in age from early 40s to late 70s. The interpreters are language professionals employed by the ICTR with experience interpreting between French and English. Their age, gender, and native languages are unknown.

D. ANNOTATOR DEMOGRAPHIC The initial transcription was outsourced to a professional transcription company, so information about these transcribers is unavailable. The English transcripts were reviewed by English-speaking (en-US) members of the research team for accuracy and then reviewed a third time by an additional English speaking (en-US) member of the team. The French/English transcripts received a second and third review for accuracy by bilingual French/English doctoral students at the University of Washington. Because of the sensitivity of the topic, the high political status of some interviewees (e.g., prosecutor for the tribunal), and the international stature of the institution, it is very important that interviewees' comments be accurately transcribed. Accordingly, the bar for quality of transcription was set extremely high.

[14] This data statement was prepared based on information provided by co-author Batya Friedman.

[15] At the end of one interview, there are 38 seconds of untranscribed speech in Kinyarwanda (rw).

E. SPEECH SITUATION The interviews were conducted in Autumn 2008 at the ICTR in Arusha, Tanzania, and in Rwanda, face-to-face, as spoken language. The interviewers begin with a prepared set of questions, but most of the interaction is semi-structured. Most generally, the speech situation can be characterized as a dialogue, but some of the interviewees give long replies, so stretches may be better characterized as monologues. For the interviewees, the immediate interlocutor is the interviewer, but the intended audience is much larger (see Part F).

F. TEXT CHARACTERISTICS The interviews were intended to provide an opportunity for tribunal personnel to reflect on their experiences working at the ICTR and what they would like to share with the people of Rwanda, the international justice community, and the global public now, and 50 and 100 years from now. Professionals from all organs of the tribunal (judiciary, prosecution, registry) were invited to be interviewed, with effort made to include a broad spectrum of roles (e.g., judges, prosecutor, defense counsel, but also the warden, librarian, language services). Interviewees expected their interviews to be made broadly accessible.

G. RECORDING QUALITY The video interviews were recorded with high definition equipment in closed but not soundproof offices. There is some background noise.

H. OTHER N/A

I. PROVENANCE APPENDIX N/A.

VRT Short Form The data represent well-vetted transcripts of 49 spoken interviews with personnel from the International Criminal Tribunal for Rwanda (ICTR) about their experience at the tribunal and reflections on international justice, in international English (44 interviews) and French (5 interviews with interpreters). Interviewees are adults working in international justice and support fields at the ICTR; interviewers are adult information or legal professionals, highly fluent in en-US; and transcribers are highly educated, highly fluent English and French speakers. [Include a link to the long form.]

6.3 Summary

These sample data statements are meant to illustrate how the schema can be used to communicate the specific characteristics of datasets. They were both created post hoc, in communication with the dataset curators. Once data statements are created as a matter of best practice, however, they should be developed in tandem with the datasets themselves and may even inform the curation of datasets. At the same time, data statements will need to be written for widely used, pre-existing datasets, where documentation may be lacking, memories imperfect, and dataset curators no longer accessible. While retrospective data statements may be incomplete, by and large we believe they can still be valuable.

Our case studies also underscore how curation rationales shape the specific kinds of texts included. This is particularly striking in the case of the Hate Speech Twitter Annotations, where the specific search terms very clearly shaped the specific kinds of hate speech included and the ways in which any technology or studies built on this dataset will generalize.

7 A Tool for Mitigating Bias

We have explicitly designed data statements as a tool for mitigating bias in systems that use data for training and testing. Data statements are particularly well suited to mitigate forms of emergent and pre-existing bias. For the former, we see benefits at the level of specific systems and of the field: When a system is paired with data statement(s) for the data it is trained on, those deploying it are empowered to assess potential gaps between the speaker populations represented in the training and test data and the populations whose language the system will be working with. At the field level, data statements enable an examination of the entire catalog of testing and training datasets to help identify populations who are not yet included. All of these groups are vulnerable to **emergent bias**, in that any system would by definition have been trained and tested on data from datasets that do not represent them well.

Data statements can also be instrumental in the diagnosis (and thus mitigation) of **pre-existing bias**. Consider again Speer's (2017) example of Mexican restaurants and sentiment analysis. The information that the word vectors were trained on general Web text (together with knowledge of what kind of societal biases such text might contain) was key in figuring out why the system consistently underestimated the ratings associated with reviews of Mexican restaurants. In order to enable both more informed system development

and deployment and audits by users and others of systems in action, it is critical that characterizations of the training and test data underlying systems be available.

To be clear, data statements do not in and of themselves solve the entire problem of bias. Rather, they are a critical enabling infrastructure. Consider by analogy this example from Friedman (1997) about access to technology and employment for people with disabilities.

> In terms of computer system design, we are not so privileged as to determine rigidly the values that will emerge from the systems we design. But neither can we abdicate responsibility. For example, let us for the moment agree [...] that disabled people in the work place should be able to access technology, just as they should be able to access a public building. As system designers we can make the choice to try to construct a technological infrastructure which disabled people can access. If we do not make this choice, then we single-handedly undermine the principle of universal access. But if we do make this choice, and are successful, disabled people would still rely, for example, on employers to hire them. (page 3)

Similarly, with respect to bias in NLP technology, if we do not make a commitment to data statements or a similar practice for making explicit the characteristics of datasets, then we will single-handedly undermine the field's ability to address bias.

In NLP, we expect proposals to come with some kind of evaluation. In this paper, we have demonstrated the substance and "writability" of a data statement through two exemplars (§6). The positive effects of data statements that we anticipate (and negative effects we haven't anticipated) cannot be demonstrated and tested a priori, however, as their impact emerges through practice. Thus, we look to value sensitive design, which encourages us to consider what would happen if a proposed technology were to come into widespread use, over longer periods of time, with attention to a wide range of stakeholders, potential benefits, and harms (Friedman et al., 2006, 2017). We do this with value scenarios (Nathan et al., 2007; Czeskis et al., 2010).

Specifically, we look at two kinds of value scenarios: Those concerning NLP technology that fails to take into account an appropriate match between training data and deployment context and those that envision possible positive as well as negative consequences stemming from the widespread use of the specific "technology" we are proposing in this paper (data statements). Envisioning possible negative outcomes allows us to consider how to mitigate such possibilities before they occur.

7.1 Public Health and NLP for Social Media

This value scenario is inspired by Jurgens et al. (2017), who provide a similar one to motivate training language ID systems on more representative datasets.

Scenario. Big U Hospital in a town in the Upper Midwest collaborates with the Computer Science Department at Big U to create a Twitter-based early warning system for infectious disease called DiseaseAlert. Big U Hospital finds that the system improves patient outcomes by alerting hospital staff to emerging community health needs and alerting physicians to test for infectious diseases that currently are active locally.

Big U decides to make the DiseaseAlert project open source to provide similar benefits to hospitals across the Anglophone world and is delighted to learn that City Hospital in Abuja, Nigeria, is excited to implement DiseaseAlert locally. Big U supports City Hospital with installing the code, including localizing the system to draw on tweets posted from Abuja. Over time, however, City Hospital finds that the system is leading its physicians to order unnecessary tests and that it is not at all accurate in detecting local health trends. City Hospital complains to Big U about the poor system performance and reports that their reputation is being damaged.

Big U is puzzled, as the DiseaseAlert performs well in the Upper Midwest, and they had spent time localizing the system to use tweets from Abuja. After a good deal of frustration and investigation into Big U's system, the developers discover that the third-party language ID component they had included was trained on only highly edited US and UK English text. As a result, it tends to misclassify tweets in regional or non-standard varieties of English as "not English" and therefore not relevant. Most of the tweets posted

by people living in Abuja that City Hospital's system should have been looking at were thrown out by the system at the first step of processing.

Analysis. City Hospital adopted Big U's open source DiseaseAlert system in exactly the way Big U intended. The documentation for the language ID component lacked critical information needed to help ensure the localization process would be successful, however; namely, information about the training and test sets for the system. Had Big U included data statements for all system components (including third-party components) in their documentation, then City Hospital IT staff would have been positioned to recognize the potential limitation of DiseaseAlert and to work proactively with Big U to ensure the system performed well in City Hospital's context. Specifically, in reviewing data statements for all system components, the IT staff could note that the language ID component was trained on data unlike what they were seeing in their local tweets and ask for a different language ID component or ask for the existing one to be retrained. In this manner, an emergent bias and its concomitant harms could have been identified and addressed during the system adaptation process prior to deployment.

7.2 Toward an Inclusive Data Catalog

In §7.1 we consider data statements in relation to a particular system. Here, we explore their potential to enable better science in NLP overall.

Scenario. It's 2022 and "Data Statement" has become a standard section heading for NLP research papers and system documentation. Happily, reports of mismatch between dataset and community of application leading to biased systems have decreased. Yet, research community members articulate an unease regarding which language communities are and which are not part of the field's data catalog—the abstract total collection of data and associated meta-data to which the field has access—and the possibility for resulting bias in NLP at a systemic level.

In response, several national funding bodies jointly fund a project to discover gaps in knowledge. The project compares existing data statements to surveys of spoken languages and systematically maps which language varieties have resources (annotated corpora and standard processing tools) and which ones lack such resources. The study turns up a large number of language varieties lacking such resources; it also produces a precise list of underserved populations, some of which are quite sizable, suggesting opportunity for impactful intervention at the academic, industry, and government levels.

Study results in hand, the NLP community embarks on an intentional program to broaden the language varieties in the data catalog. Public discussions lead to criteria for prioritizing language varieties and funding agencies come together to fund collaborative projects to produce state of the art resources for understudied languages. Over time, the data catalog becomes more inclusive; bias in the catalog, although not wholly absent, is significantly reduced and NLP researchers and developers are able to run more comprehensive experiments and build technology that serves a larger portion of society.

Analysis. The NLP community has recognized critical limitations in the field's existing data catalog, leaving many language communities underserved (Bender, 2011; Munro, 2015; Jurgens et al., 2017).[16] The widespread uptake of data statements positions the NLP community to document the degree to which it leaves out certain language groups and empower itself to systematically broaden the data catalog. In turn, individual NLP systems could be trained on datasets that more closely align with the language of anticipated system users, thereby averting emergent bias. Furthermore, NLP researchers can more thoroughly test key research ideas and systems, leading to more reliable scientific results.

7.3 Anticipating and Mitigating Barriers

Finally, we explore one potential negative outcome and how with care it might be mitigated: that of data statements as a barrier to research.

Scenario. In response to widespread uptake, in 2026 the Association for Computational Linguistics (ACL) proposes that data statements be standardized and required components of research papers. A standards committee is formed, open public professional discussion is engaged, and in 2028 a standard is adopted. It mandates data

[16]The EU-funded project META-NET worked on identifying gaps at the level of whole languages for Europe, producing a series of 32 white papers, each concerning one European language, available from http://www.meta-net.eu/whitepapers/overview, accessed 6 August 2018.

statements as a requirement for publication, with standardized information fields and strict specifications for how these should be completed to facilitate automated meta-analysis. There is great hope that the field will experience increasing benefits from ability to compare, contrast, and build complementary data sets.

Many of those hopes are realized. However, in a relatively short period of time papers from underrepresented regions abruptly decline. In addition, the number of papers from everywhere producing and reporting on new datasets decline as well. Distressed by this outcome, the ACL constitutes an ad hoc committee to investigate. A survey of researchers reveals two distinct causes: First, researchers from institutions not yet well represented at ACL were having their papers desk-rejected because of missing or insufficient data statements. Second, researchers who might otherwise have developed a new dataset instead chose to use existing datasets whose data statements could simply be copied. In response, the ACL executive develops a mentoring service to assist authors in submitting standards-compliant data statements and considers relaxing the standard somewhat in order to encourage more dataset creation.

Analysis. With any new technology, there can be unanticipated ripple effects—data statements are no exception. Here, we envision two potential negative impacts, which could both be mitigated through other practices. Importantly, although we recommend the practice of creating data statements, we believe that they should be widely used before any standardization takes place. Furthermore, once a degree of expertise in this area is built up, we recommend that mentoring be put in place proactively. Community engagement and mentoring will also contribute to furthering ethical discourse and practice in the field.

7.4 Summary

The value scenarios described here point to key upsides to the widespread adoption of data statements and also help to provide words of caution. They are meant to be thought-provoking and plausible, but are not predictive. Importantly, the scenarios illustrate how, if used well, data statements could be an effective tool for mitigating bias in NLP systems.

8 Related Work

We see three strands of related work that lend support to our proposal and to the proposition that data statements will have the intended effect: similar practices in medicine (§8.1); emerging, independent proposals around similar ideas for transparency about datasets in AI (§8.2); and proposals for "algorithmic impact statements" (§8.3).

8.1 Guidelines for Reporting Medical Trials

In medicine, the CONSORT (CONsolidated Standards of Reporting Trials) guidelines were developed by a consortium of journal editors, specialists in clinical trial methodology, and others to improve reporting of randomized, controlled trials.[17] They include a checklist for authors to use to indicate where in their research reports each item is handled and a statement explaining the rationale behind each item (Moher et al., 2010). CONSORT development began in 1993, with the most recent release in 2010. It has been endorsed by 70 medical journals.[18]

Item 4a, "Eligibility criteria for participants", is most closely related to the concerns of this paper. Characterizing the population that participated in the study is critical for gauging the extent to which the results of the study are applicable to particular patients a physician is treating (Moher et al., 2010).

The inclusion of this information has also enabled further kinds of research. For example, Mbuagbaw et al. (2017) argue that careful attention to and publication of demographic data that may correlate with health inequities can facilitate further work through meta-analyses. In particular, individual studies usually lack the statistical power to do the kind of sub-analyses required to check for health inequities, and failing to publish demographic information precludes its use in the kind of aggregated, meta-analyses that could have sufficient statistical power. This echoes the field-level benefits we anticipate for data statements in building out the data catalog in the value scenario in §7.2.

[17] http://www.consort-statement.org/consort-2010, accessed 12 July 2017.

[18] http://www.consort-statement.org/about-consort/endorsement-of-consort-statement, accessed 12 July 2017.

8.2 Converging Proposals

At least three other groups are working in parallel on similar proposals regarding bias and Artificial intelligence (AI). Gebru et al. (in preparation) propose "datasheets for datasets," looking at AI more broadly (but including NLP); Chmielinski and colleagues at the MIT Media Lab propose "dataset nutrition labels";[19] and Yang et al. (2018) describe "Ranking Facts," a series of widgets that allow a user to explore how attributes influence a ranking. Of these, the datasheets proposal is most similar to ours in including a comparable schema.

The datasheets are inspired by those used in computer hardware to give specifications, limits, and appropriate use information for components. There is important overlap in the kinds of information called for in the datasheets schema and our data statement schema: For example, the datasheets schema includes a section on "Motivation for Dataset Creation," akin to our "Curation Rationale." The primary differences stem from the fact that the datasheets proposal is trying to accommodate all types of datasets used to train machine learning systems and, hence, tends toward more general, cross-cutting categories, whereas we elaborate requirements for linguistic datasets and, hence, provide more specific, NLP-focused categories. Gebru et al. note, like us, that their proposal is meant as an initial starting point to be elaborated through adoption and application. Having multiple starting points for this discussion will certainly make it more fruitful.

8.3 Algorithmic Impact Statements

Several groups have called for algorithmic impact statements (Diakopoulos, 2016; Shneiderman, 2016; AI Now Institute, 2018), modeled after environmental impact statements. Of these, AI Now's proposal is perhaps the most developed. All three groups point to the need to clarify information about the data: "Algorithm impact statements would document [...] data quality control for input sources" (Shneiderman, 2016, page 13539); "One avenue for transparency here is to communicate the quality of the data, including its accuracy, completeness, and uncertainty, [...] representativeness of a sample for a specific population, and assumptions or other limitations" (Diakopoulos, 2016, page 60); "AIAs should cover [...] input and training data" (AI Now Institute, 2018). However, none of these proposals specify how to do so. Data statements fill this critical gap.

9 Recommendations for Implementation

Data statements are meant to be something practical and concrete that NLP technologists can adopt as one tool for mitigating potential harms of the technology we develop. For this benefit to come about, data statements must be easily adopted. In addition, practical uptake will require coordinated effort at the level of the field. In this section we briefly consider possible costs to writers and readers of data statements, and then propose strategies for promoting uptake.

The primary cost we see for writers is time: With the required information to hand, writing a data statement should take no more than 2–3 hours (based on our experience with the case studies). The time to collect the information will depend on the dataset, however. The more speakers and annotators are involved, the more time it may take to collect demographic information. This can be facilitated by planning ahead, before the corpus is collected. Another possible cost is that collecting demographic information may mean that projects previously not submitted to institutional review boards for approval must now be, at least for exempt status. This process itself can take time, but is valuable in its own right. A further cost to writers is space. We propose that data statements, even the short form (60–100 words), be exempt from page limits in conference and journal publications.

As for readers, reviewers have more material to read and dataset (and ultimately system) users need to scrutinize data statements in order to determine which datasets are appropriate for their use case. But this is precisely the point: Data statements make critical information accessible that previously could only be found by users with great effort, if at all. The time invested in scrutinizing data statements prior to dataset adoption is expected to be far less than the time required to diagnose and retrofit an already deployed system should biases be identified.

Turning to uptake in the field, NLP technologists (both researchers and system developers) are key stakeholders of the technology of data statements. Practices that engage these stakeholders in the development and promotion of data statements

[19] http://datanutrition.media.mit.edu/, accessed 2 April 2018.

will both promote uptake and ensure that the ultimate form data statements take are responsive to NLP technologists' needs. Accordingly, we recommend that one or more professional organizations such as the Association for Computational Linguistics convene a working group on data statements.

Such a working group would engage in several related sets of activities, which would collectively serve to publicize and cultivate the use of data statements:

(i) Best Practices A clear first step entails developing best practices for how data statements are produced. This includes: steps to take before collecting a dataset to facilitate writing an informative data statement; heuristics for writing concise and effective data statements; how to incorporate material from institutional review board/ethics committee applications into the data statement schema; how to find an appropriate level of detail given privacy concerns, especially for small or vulnerable populations; and how to produce data statements for older datasets that predate this practice. In doing this work, it may be helpful to distill best practices from other fields, such as medicine and psychology, especially around collecting demographic information.

(ii) Training and Support Materials With best practices in place, the next step is providing training and support materials for the field at large. We see several complementary strategies to undertake: Create a digital template for data statements; run tutorials at conferences; establish a mentoring network (see §7.3); and develop an online "how-to" guide.

(iii) Recommendations for Field-level Policies There are a number of field-level practices that the working group could explore to support the uptake and successful use of data statements. Funding agencies could require data statements to be included in data management plans; conferences and journals could not count data statements against page limits (similar to references) and eventually require short form data statements in submissions; conferences and journals could allocate additional space for data statements in publications; and finally, once data statements have been in use for a few years, a standardized form could be established.

10 Tech Policy Implications

Transparency of datasets and systems is essential for preserving accountability and building more just systems (Kroll et al., 2017). Due process provides a critical case in point. In the United States, for example, due process requires that citizens who have been deprived of liberty or property by the government be afforded the opportunity to understand and challenge the government's decision (Citron, 2008). Without data statements or something similar, governmental decisions that are made or supported by automated systems deprive citizens of the ability to mount such a challenge, undermining the potential for due process.

In addition to challenging any specific decision by any specific system, there is a further concern about building systems that are broadly representative and fair. Here, too, data statements have much to contribute. 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. Once systems are deployed, data statements enable diagnosis of systemic unfairness when it is detected in system performance. At a societal level, such transparency is necessary for government and advocacy groups seeking to ensure protections and an inclusive society.

If data statements turn out to be useful as anticipated, then the following implications for standardization and tech policy likely ensue.

Long-Form Data Statements Required in System Documentation. For academia, industry, and government, inclusion of long-form data statements as part of system documentation should be a requirement. As appropriate, inclusion of long-form data statements should be a requirement for ISO and other certification. Even groups that are creating datasets that they don't share (e.g., US National Security Agency) would be well advised to make internal data statements. Moreover, under certain legal circumstances, such groups may be required to share this information.

Short-Form Data Statements Required for Academic and Other Publication. For academic publication in journals and conferences, inclusion of short-form data statements should be a requirement. As highlighted in §7.3, caution must be exercised to ensure that this requirement does not become a barrier to access for some researchers.

These two recommendations will need to be implemented with care. We have already noted the potential barrier to access. Secrecy concerns may also arise in some situations (e.g., some groups may be willing to share datasets but not demographic information, for fear of public relations backlash or to protect the safety of contributors to the dataset). That said, as consumers of datasets or products trained with them, NLP researchers, developers, and the general public would be well advised to use systems only if there is access to the information we propose should be included in data statements.

11 Conclusion and Future Work

As researchers and developers working on technology in widespread use, capable of impacting people beyond its direct users, we have an obligation to consider the ethical implications of our work. This will only happen reliably if we find ways to integrate such thought into our regular practice. In this paper, we have put forward one specific, concrete proposal that we believe will help with issues related to exclusion and bias in language technology: the practice of including "data statements" in all publications and documentation for all NLP systems.

We believe this practice will have beneficial effects immediately and into the future: In the short term, it will foreground how our data do and do not represent the world (and the people our systems will impact). In the long term, it should enable research that specifically addresses issues of bias and exclusion, promote the development of more representative datasets, and make it easier and more normative for researchers to take stakeholder values into consideration as they work. In foregrounding the information about the data we work with, we can work toward making sure that the systems we build work for diverse populations and also toward making sure we are not teaching computers about the world based on the world views of a limited subset of people.

Granted, it will take time and experience to develop the skill of writing carefully crafted data statements. However, we see great potential benefits: For the scientific community, researchers will be better able to make precise claims about how results should generalize and perform more targeted experiments around reproducing results for datasets that differ in specific characteristics. For industry, we believe that incorporating data statements will encourage the kind of conscientious software development that protects companies' reputations (by avoiding public embarrassment) and makes them more competitive (by creating systems used more fluidly by more people). For the public at large, data statements are one piece of a larger collection of practices that will enable the development of NLP systems that equitably serves the interests of users and indirect stakeholders.

Acknowledgments

We are grateful to the following people for helpful discussion and critical commentary as we developed this paper: the anonymous TACL reviewers, Hannah Almeter, Stephanie Ballard, Chris Curtis, Leon Derczynski, Michael Wayne Goodman, Anna Hoffmann, Bill Howe, Kristen Howell, Dirk Hovy, Jessica Hullman, David Inman, Tadayoshi Kohno, Nick Logler, Mitch Marcus, Angelina McMillan-Major, Rob Munro, Glenn Slayden, Michelle Stamnes, Jevin West Daisy Yoo, Olga Zamaraeva, and especially Zeerak Waseeem and Ryan Calo. We have presented talks based on earlier versions of this paper at New York University (Nov 2017), Columbia University (Nov 2017), University of Washington (Nov 2017), UC San Diego (Feb 2018), Microsoft (Mar 2018) and Macquarie University (July 2018) and thank the audiences at those talks for useful feedback. Finally, Batya Friedman's contributions to this paper were supported by the UW Tech Policy Lab and National Science Foundation grant IIS-1302709. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

AI Now Institute. 2018. Algorithmic impact assessments: Toward accountable automation in public agencies. Medium.com, https://medium.com/@AINowInstitute/algorithmic-impact-assessments-impact-assessments-toward-accountable-automation-in-public-agencies-bd9856e6fdde, accessed 6 April 2018.

American Psychological Association. 2009. *Publication Manual of the American Psychological Association*, 6th edition. Author, Washington DC.

Emily M. Bender. 2011. On achieving and evaluating language independence in NLP. *Linguistic Issues in Language Technology*, 6:1–26.

Douglas Biber. 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge University Press, Cambridge.

Steven Bird and Gary Simons. 2000. White paper on establishing an infrastructure for open language archiving. In *Workshop on Web-Based Language Documentation and Description, Philadelphia, PA*, pages 12–15.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4349–4357. Curran Associates, Inc.

Nicoletta Calzolari, Valeria Quochi, and Claudia Soria. 2012. The strategic language resource agenda. http://www.flarenet.eu/sites/default/files/FLaReNet_Strategic_Language_Resource_Agenda.pdf, accessed 6 August 2018.

Jack K. Chambers and Peter Trudgill. 1998. *Dialectology*, second edition. Cambridge University Press.

Danielle Keats Citron. 2008. Technological due process. *Washington University Law Review*, 85:1249–1313.

TEI Consortium. 2008. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. http://www.tei-c.org/guidelines/p5/, accessed 6 August 2018.

Ben Coppin. 2004. *Artificial Intelligence Illuminated*. Jones & Bartlevtt Publishers, Sudbury MA.

Alexei Czeskis, Ivayla Dermendjieva, Hussein Yapit, Alan Borning, Batya Friedman, Brian Gill, and Tadayoshi Kohno. 2010. Parenting from the pocket: Value tensions and technical directions for secure and private parent-teen mobile safety. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*. ACM.

Leon Derczynski, Kalina Bontcheva, and Ian Roberts. 2016. Broad twitter corpus: A diverse named entity recognition resource. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1169–1179. The COLING 2016 Organizing Committee.

Laurence Devillers, Björn Schuller, Emily Mower Provost, Peter Robinson, Joseph Mariani, and Agnes Delaborde, editors. 2016. *Proceedings of ETHI-CA2 2016: ETHics in Corpus Collection, Annotation & Application*. LREC.

Nicholas Diakopoulos. 2016. Accountability in algorithmic decision making. *Communications of the ACM*, 59(2):56–62.

Penelope Eckert and John R. Rickford, editors. 2001. *Style and Sociolinguistic Variation*. Cambridge University Press, Cambridge.

Rod Ellis. 1994. *The Study of Second Language Acquisition*. Oxford University Press, Oxford.

Susan Ervin-Tripp. 1964. An analysis of the interaction of language, topic, and listener. *American Anthropologist*, 66(6_PART2):86–102.

Karën Fort, Gilles Adda, and K. Bretonnel Cohen, editors. 2016. *TAL et Ethique, special issue of Traitement automatique des languages*, volume 57:2.

Batya Friedman. 1997. Introduction. In Batya Friedman, editor, *Human Values and the Design of Computer Technology*, pages 1–18. Stanford CA, Stanford.

Batya Friedman, David G Hendry, and Alan Borning. 2017. A survey of value sensitive design methods. *Foundations and Trends® in Human–Computer Interaction*, 11(2):63–125.

Batya Friedman, Peter H. Kahn, Jr., and Alan Borning. 2006. Value sensitive design and information systems. In Ping Zhang and Dennis F. Galletta, editors, *Human–Computer Interaction in Management Information Systems: Foundations*, pages 348–372. M. E. Sharpe, Armonk NY.

Batya Friedman and Lisa P. Nathan. 2010. Multi-lifespan information system design: A research initiative for the HCI community. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2243–2246. ACM.

Batya Friedman, Lisa P. Nathan, and Daisy Yoo. 2016. Multi-lifespan information system design in support of transitional justice: Evolving situated design principles for the long(er) term. *Interacting with Computers*, 29:80–96.

Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, 14(3):330–347.

John Furler, Parker Magin, Marie Pirotta, and Mieke van Driel. 2012. Participant demographics reported in "table 1" of randomised controlled trials: A case of "inverse evidence"? *International Journal for Equity in Health*, 11.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. in prep. Datasheets for datasets. ArXiv:1803.09010v1.

Dirk Hovy and Anders Søgaard. 2015. Tagging performance correlates with author age. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 483–488. Association for Computational Linguistics.

Dirk Hovy, Shannon Spruit, Margaret Mitchell, Emily M. Bender, Michael Strube, and Hanna Wallach, editors. 2017. *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Association for Computational Linguistics.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598. Association for Computational Linguistics.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2015. Challenges of studying and processing dialects in social media. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 9–18. Association for Computational Linguistics.

David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. 2017. Incorporating dialectal variability for socially equitable language identification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 51–57. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53. Association for Computational Linguistics.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430. Association for Computational Linguistics.

Roger Kreuz and Gina Caucci. 2007. Lexical influences on the perception of sarcasm. In *Proceedings of the Workshop on Computational Approaches to Figurative Language*, pages 1–4. Association for Computational Linguistics.

Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, and Harlan Yu. 2017. Accountable algorithms. *University of Pennsylvania Law Review*, 165. Fordham Law Legal Studies Research Paper No. 2765268. Available at SSRN: https://ssrn.com/abstract=2765268, accessed 6 August 2018.

William Labov. 1966. *The Social Stratification of English in New York City*. Center for Applied Linguistics, Washington, DC.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*, volume 5:1 of *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19:313–330.

Lawrence Mbuagbaw, Theresa Aves, Beverley Shea, Janet Jull, Vivian Welch, Monica Taljaard, Manosila Yoganathan, Regina Greer-Smith, George Wells, and Peter Tugwell. 2017. Considerations and guidance in designing equity-relevant clinical trials. *International Journal for Equity in Health*, 16(1):93.

Davida Moher, Sally Hopewell, Kenneth F. Schulz, Victor Montori, Peter C. Gøtzsche, P. J. Devereaux, Diana Elbourne, Matthias Egger, and Douglas G. Altman. 2010. CONSORT 2010 explanation and elaboration: Updated guidelines for reporting parallel group randomised trials. *The BMJ*, 340.

Robert Munro. 2015. Languages at ACL this year. Blog post, http://www.junglelightspeed.com/languages-at-acl-this-year/, accessed 22 September 2017.

Robert Munro and Christopher D. Manning. 2010. Subword variation in text message classification. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 510–518. Association for Computational Linguistics.

Lisa P. Nathan, Predrag V. Klasnja, and Batya Friedman. 2007. Value scenarios: A technique for envisioning systemic effects of new technologies. In *CHI'07 Extended Abstracts on Human Factors in Computing Systems*, pages 2585–2590. ACM.

Lisa P. Nathan, Milli Lake, Nell Carden Grey, Trond Nilsen, Robert F. Utter, Elizabeth J. Utter, Mark Ring, Zoe Kahn, and Batya Friedman. 2011. Multi-lifespan information system design: Investigating a new design approach in Rwanda. In *Proceedings of the 2011 iConference*, pages 591–597. ACM.

Trond T. Nilsen, Nell Carden Grey, and Batya Friedman. 2012. Public curation of a historic collection: A means for speaking safely in public. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work Companion*, pages 277–278. ACM.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 79–86. Association for Computational Linguistics.

Tomáš Ptáček, Ivan Habernal, and Jun Hong. 2014. Sarcasm detection on Czech and English Twitter. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 213–223. Dublin City University and Association for Computational Linguistics.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 704–714. Association for Computational Linguistics.

Ben Shneiderman. 2016. Opinion: The dangers of faulty, biased, or malicious algorithms requires independent oversight. *Proceedings of the National Academy of Sciences*, 113(48): 13538–13540.

Robyn Speer. 2017. Conceptnet numberbatch 17.04: better, less-stereotyped word vectors. Blog post, https://blog.conceptnet.io/2017/04/24/conceptnet-numberbatch-17-04-better-less-stereotyped-word-vectors/, accessed 6 July 2017.

Rachael Tatman. 2017. Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59. Association for Computational Linguistics.

Zeerak Waseem. 2016. Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research*

Workshop, pages 88–93. Association for Computational Linguistics.

Ke Yang, Julia Stoyanovich, Abolfazl Asudeh, Bill Howe, H. V. Jagadish, and Gerome Miklau. 2018. A nutritional label for rankings. In *Proceedings of the 2018 International Conference on Management of Data*, SIGMOD '18, pages 1773–1776, New York, NY, USA. ACM.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2941–2951. Association for Computational Linguistics.