Toward Gender-Inclusive Coreference Resolution: An Analysis of Gender and Bias Throughout the Machine Learning Lifecycle*

Yang Trista Cao
Department of Computer Science
University of Maryland
ycao95@umd.edu

Hal Daumé III
Microsoft Research
University of Maryland
me@hal3.name

Correctly resolving textual mentions of people fundamentally entails making inferences about those people. Such inferences raise the risk of systematic biases in coreference resolution systems, including biases that can harm binary and non-binary trans and cis stakeholders. To better understand such biases, we foreground nuanced conceptualizations of gender from sociology and sociolinguistics, and investigate where in the machine learning pipeline such biases can enter a coreference resolution system. We inspect many existing data sets for trans-exclusionary biases, and develop two new data sets for interrogating bias in both crowd annotations and in existing coreference resolution systems. Through these studies, conducted on English text, we confirm that without acknowledging and building systems that recognize the complexity of gender, we will build systems that fail for: quality of service, stereotyping, and over- or under-representation, especially for binary and non-binary trans users.

1. Introduction

Coreference resolution—the task of determining which textual references resolve to the same real-world entity—requires making inferences about those entities. Especially when those entities are people, coreference resolution systems run the risk of making unlicensed inferences, possibly resulting in harms either to individuals or groups of people. Embedded in coreference inferences are varied aspects of gender, both because gender can show up explicitly (e.g., pronouns in English, morphology in Arabic) and because societal expectations and stereotypes around gender roles may be explicitly or

* This study extends the work of Cao and Daumé III (2020); the specific additions are highlighted in the bulleted list toward the end of the Introduction.

Submission received: 23 November 2020; revised version received: 9 March 2021; accepted for publication: 22 June 2021.

https://doi.org/10.1162/COLI_a_00413

implicitly assumed by speakers or listeners. This can lead to significant biases in coreference resolution systems: cases where systems "systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others" (Friedman and Nissenbaum 1996, page 332).

Gender bias in coreference resolution can manifest in many ways; work by Rudinger et al. (2018), Zhao et al. (2018a), and Webster et al. (2018) focused largely on the case of *binary* gender discrimination in trained coreference systems, showing that current systems over-rely on social stereotypes when resolving HE and SHE pronouns[1] (see §2). Contemporaneously, critical work in human–computer interaction has complicated discussions around gender in other fields, such as computer vision (Keyes 2018; Hamidi, Scheuerman, and Branham 2018).

Building on both lines of work, and inspired by Keyes's (2018) study of vision-based automatic gender recognition systems, we consider gender bias from a broader conceptual frame than the binary "folk" model. We investigate ways in which folk notions of gender—namely, that there are two genders, assigned at birth, immutable, and in perfect correspondence to gendered linguistic forms—lead to the development of technology that is exclusionary and harmful to binary and non-binary trans and non-trans people.[2] We take the normative position that although the folk model of gender is widespread even today, building systems that adhere to it implicitly or explicitly can lead to significant harms to binary and non-binary trans individuals, and that we should aim to understand and minimize those harms. We take this as particularly important not just from the perspective of potentially improving the quality of our systems when run on documents by or about trans people (as well as documents by or about non-trans people), but more pointedly to minimize the harms caused by our systems by reinforcing existing unjust social hierarchies (Lambert and Packer 2019).

Because coreference resolution is a component technology embedded in larger systems, directly implicating coreference errors in user harms is less straightforward than for user-facing technology. Nonetheless, there are several stakeholder groups that may easily face harms when coreference is used in the context of machine translation or search engine systems (discussed in detail in §4.6). Following Bender's (2019) taxonomy of stakeholders and Barocas et al.'s (2017) taxonomy of harms, there are several obvious ways in which trans exclusionary coreference resolution systems can hypothetically cause harm:

- ◇ *Indirect: subject of query.* If a person is the subject of a Web query, relevant Web pages about xem may be downranked if "multiple references to the same person" is an important feature in ranking and the coreference system cannot recognize and resolve xyr pronouns. This can lead to quality of service and erasure harms.
- ◇ *Direct: by choice.* If a grammar checker uses coreference features, it may insist that an author writing hir third-person autobiography is repeatedly

1 Throughout, we avoid mapping pronouns to a "gender" label, preferring to use the nominative case of the pronoun directly, including (in English) SHE, HE, the non-binary use of singular THEY, and neopronouns (e.g., ZE/HIR, XEY/XEM), which have been in usage since at least the 1970s (Spivak 1997; Bustillos 2011; Merriam-Webster 2016; Bradley et al. 2019; Hord 2016).

2 Following GLAAD (2007), transgender individuals are those whose gender differs from the sex they were assigned at birth. This is in opposition to cisgender (or non-trans) individuals, whose assigned sex at birth happens to correspond to their gender. Transgender individuals can either be binary (those whose gender falls in the "male/female" dichotomy) or non-binary (those for whom the relationship is more complex).

making errors in refering to hirself. This can lead to quality of service and stereotyping (by reinforcing the stereotype that trans identities are not "real").

- ◇ *Direct: not by choice.* If an information extraction on job applications uses a coreference system as a preprocessor, but the coreference system relies on cisnormative assumptions, then errors may disproportionately affect those who do not fit in the gender binary. This can lead to allocative harms (for hiring) as well as erasure harms.
- ◇ *Many stakeholders*. If a machine translation system needs to use discourse context to generate appropriate pronouns or gendered morphological inflections in a target language, then errors can result in directly misgendering[3] subjects of the document being translated.

To address these (and other) potential harms in more detail, as well as where and how they arise, we need to (a) complicate what "gender" means and (b) uncover how harms can enter into natural language processing (NLP) systems. Toward (a), we begin with a unifying analysis (§3) of how gender is socially constructed, and how social conditions in the world impose expectations around people's gender. Of particular interest is how gender is reflected in language, and how that both matches and potentially mismatches the way people experience their gender in the world. This reflection is highlighted, for instance, in folk notions such as an implicitly assumed one-to-one mapping between a gender and pronouns. Then, in order to understand social biases around gender, we find it necessary to consider the different ways in which gender can be realized linguistically, breaking down what previously have been considered "gendered words" in NLP papers into finer-grained categories of lexical, referential, grammatical, and social gender. Through this deconstruction (well-established in sociolinguistics), we can begin to interrogate what forms of gender stereotyping are prevalent in coreference resolution.

Toward (b), we ground our analysis by adapting Vaughan and Wallach's (2019) framework of how a prototypical machine learning lifecycle operates.[4] We analyze forms of gender bias in coreference resolution in six of the eight stages of the lifecycle in Figure 1. We conduct much of our analysis around task definition (§4.1), bias in underlying text (§4.2), model definition (§4.5), and evaluation methodologies (§4.6) by evaluating prior coreference data sets, their corresponding annotation guidelines, and through a critical read of "gender" discussions in natural language processing papers. For our analysis of bias in annotations due to annotator positionality (§4.3), and our analysis of model definition (§4.5), we construct two new coreference data sets: MAP (a similar data set to GAP [Webster et al. 2018] but without binary gender constraints on which we can perform counterfactual manipulations; see §4.2) and GICoref (a fully annotated coreference resolution data set written by and/or about trans

3 According to Clements (2018), misgendering occurs when you intentionally or unintentionally refer to a person, relate to a person, or use language to describe a person that doesn't align with their affirmed gender.

4 Vaughan and Wallach (2019) do not distinguish data collection and data annotation (and call the combined step "Dataset Construction"); for us, the separation is natural and useful for analysis. We also replace "Test Model" and "Deploy Model" with "Test System" and "Deploy System," where the system is more inclusive with training data, trained model, etc.

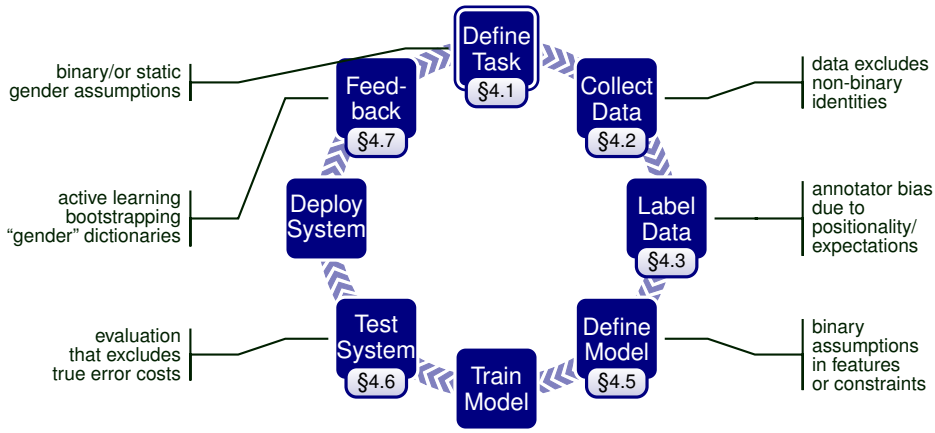

Figure 1
Machine learning lifecycle, with pointers to sections in this paper and some possible sources of bias listed for each state of the lifecycle; adapted from Vaughan and Wallach (2019).

people; see §4.5.3).[5] In all cases, we focus largely on harms due to over- and under-representation (Kay, Matuszek, and Munson 2015), replicating stereotypes (Sweeney 2013; Caliskan, Bryson, and Narayanan 2017) (particularly those that are cisnormative and/or heteronormative), and quality of service differentials (Buolamwini and Gebru 2018).
The primary contributions of this paper are:[6]

- ♢ Analyzing gender bias in the entire coreference resolution lifecycle, with a particular focus on how coreference resolution may fail to adequately process text involving binary and non-binary trans referents (§4).
- ⋄ Developing an ablation technique for measuring gender bias in coreference resolution annotations, focusing on the *human* biases that can enter into annotation tasks (§4.3).
- ⋄ Constructing a new data set, the Gender Inclusive Coreference data set (GICoref), for testing performance of coreference resolution systems on texts that discuss non-binary and binary transgender people (§4.5.3).
- ♢ Connecting existing work on gender bias in natural language processing to sociological and sociolinguistic conceptions of gender to provide a scaffolding for future work on analyzing "gender bias in NLP" (§3).

We conclude (§5) with a discussion of how the natural language processing community can move forward in this task in particular, and also how this case study can be generalized to other language settings. Our goal is to highlight issues in previous instantiations of coreference resolution in order to improve tomorrow's instantiations, continuing the

5 Both data sets are released under a BSD license at `github.com/TristaCao/into_inclusivecoref` with corresponding datasheets (Gebru et al. 2018).
6 As noted earlier, this work is an extension of our work published in ACL 2020 (Cao and Daumé III 2020). Contributions with ⋄ are published in Cao and Daumé III (2020) and contributions with ♢ are the new contributions of this paper. Note that the analysis of gender concepts in the last contribution is an extended version of the analysis in the ACL paper.

lifecycle of coreference resolutions' various task definition updates from MUC7 in 2001 through ACE in the mid 2000s and up to today.

Significant Limitations. The primary limitation of our study and analysis is that it is largely limited to English: Our consideration of task definition in §4.1 discusses other languages, but all the data and models we consider are for English. This is particularly limiting because English lacks a grammatical gender system (discussion in §3.2), and some extensions of our work to languages with grammatical gender are non-trivial. We also emphasize that while we endeavored to be inclusive, in particular in the construction of our data sets, our own positionality has undoubtedly led to other biases. One in particular is a largely Western bias, both in terms of what models of gender we use (e.g., the division of sex, gender, and sexuality along a Western frame; see §3) and also in terms of the data we annotated (§4.5.3). We have attempted to partially compensate for this latter bias by intentionally including documents with non-Western binary and non-binary trans expressions of gender in the GICoref data set, but the compensation is incomplete.

Additionally, our ability to collect naturally occurring data was limited because many sources simply do not yet permit (or have only recently permitted) the use of gender inclusive language in their articles (discussion in §4.2). This led us to counterfactual text manipulation in §4.3, which, while useful, is essentially impossible to do flawlessly (additional discussion in §4.4.1). Finally, because the social construct of gender is fundamentally contested (discussion in §3.1), some of our results may apply only under some frameworks. The use of "toward" in the title of this paper is intentional: We hope this work provides a useful stepping stone as the community continues to build technology and understanding of that technology, but this work is by no means complete.

2. Other Related Work

There are four recent papers that consider gender bias in coreference resolution systems. Rudinger et al. (2018) evaluate coreference systems for evidence of *occupational stereotyping*, by constructing Winograd-esque (Levesque, Davis, and Morgenstern 2012) test examples. They find that humans can reliably resolve these examples, but systems largely fail at them, typically in a gender-stereotypical way. In contemporaneous work, Zhao et al. (2018a) proposed a very similar, also Winograd-esque scheme, also for measuring gender-based occupational stereotypes. In addition to reaching similar conclusions as Rudinger et al. (2018), this work also used a similar "counterfactual" data process as we use in §4.3.1 in order to provide additional training data to a coreference resolution system. Webster et al. (2018) produced the GAP data set for evaluating coreference systems by specifically seeking examples where "gender" (left underspecified) could *not* be used to help coreference. They found that coreference systems struggle in these cases, also pointing to the fact that some success of current coreference systems is due to reliance on (binary) gender stereotypes. Finally, Ackerman (2019) presents an alternative breakdown of gender than we use (§3), and proposes matching criteria for modeling coreference resolution linguistically, taking a trans-inclusive perspective on gender.

Gender bias in NLP has been considered more broadly than just in coreference resolution, including, for instance, natural language inference (Rudinger, May, and Van Durme 2017), word embeddings (e.g., Bolukbasi et al. 2016; Romanov et al. 2019; Gonen and Goldberg 2019), sentiment analysis (Kiritchenko and Mohammad 2018), and machine translation (Font and Costa-jussà 2019; Prates, Avelar, and Lamb 2019), among many others (Blodgett et al. 2020, inter alia). Gender is also an object of study in gender

recognition systems (Hamidi, Scheuerman, and Branham 2018). Much of this work has focused on gender bias with a (usually implicit) binary lens, an issue that was also called out recently by Larson (2017).

Outside of NLP, there have been many studies looking at how gender information (particularly in languages with grammatical gender) are processed by people, using either psycholinguistic or neurolinguistic studies. For instance, Garnham et al. (1995) and Carreiras et al. (1996) use reading speed tests for gender-ambiguous contexts, and observe faster reading when the reference was "obvious" in Spanish. Relatedly, Esaulova, Reali, and von Stockhausen (2014) and Reali, Esaulova, and Von Stockhausen (2015) conduct eye movement studies around anaphor resolution in German, corresponding to stereotypical gender roles. In neurolinguistic studies, Osterhout and Mobley (1995) and Hagoort and Brown (1999) looked at event-related potential (ERP) violations for reflexive pronouns and antecedent in English, finding similar effects to violations of number agreement, but different effects from semantic violations. Osterhout, Bersick, and McLaughlin (1997) found ERP violations of the P600 type for violations of social gender stereotypes.

Issues of ambiguity in gender are also well documented in the translation studies literature, some of which have been discussed in the machine translation setting. For example, when translating from a language that can drop pronouns in subject position—the vast majority of the world's languages (Dryer 2013)—to a language like English that (mostly) requires pronominal subjects, a system is usually forced to infer some pronoun, significantly running the risk of misgendering. Frank et al. (2004) observe that human translators may be able to use more global context to resolve gender ambiguities than a machine translation system that does not take into account discourse context. However, in some cases using more context may be insufficient, either because the context simply does not contain the answer,[7] or because different languages mark for gender in different ways: For example, Hindi verbs agree with the gender of their objects, and Russian verbal forms sometimes inflect differently depending on the gender of the speaker, the addressee, or the person being discussed (Doleschal and Schmid 2001).

3. Background: Linguistic and Social Gender

The concept of gender is complex and contested, covering (at least) aspects of a person's internal experience, how they express this to the world, how social conditions in the world impose expectations on them (including expectations around their sexuality), and how they are perceived and accepted (or not). When this complex concept is realized in language, the situation becomes even more complex: Linguistic categories of gender do not even remotely map one-to-one to social categories. In order to properly discuss the role that "gender" plays in NLP systems in general (and coreference in particular), we first must work to disentangle these concepts. For without disentangling them (as few previous NLP papers have; see §4.1), we can end up conflating concepts that are fundamentally different and, in doing so, rendering ourselves unable to recognize certain forms of bias. As observed by Bucholtz (1999) page 80:

> Attempts to read linguistic structure directly for information about social gender are often misguided.

7 For instance, the gender of the *chef de cuisine* in Daphne du Maurier's *Rebecca* is never referenced, and different human translators have selected different genders when translating that book into languages with grammatical gender (Wandruszka 1969; Nissen 2002).

For instance, when working in a language like English, which formally marks gender on pronouns, it is all too easy to equate "recognizing the pronoun that corefers with this name" with "recognizing the real-world gender of referent of that name." Thus, possibly without even wishing to do so, we may effectively assume that "she" is equivalent to "female," "he" is equivalent to "male," and no other options are possible. This assumption can leak further—for instance by leading to an incorrect assumption that a single person cannot be referred to as both "she" and "he" (which can happen because a person's gender is contextual), nor by neither of those (which can happen when a person's gender does not align well with either of those English pronouns).

Furthermore, despite the impossibility of a perfect alignment with linguistic gender, it is generally clear that an incorrectly gendered reference to a person (whether through pronominalization or otherwise) can be highly problematic. This process of *misgendering* is problematic for both trans and cis individuals (the latter, for instance, in the all too common case of all computer science professors receiving "Dear Sir" emails), to the extent that transgender historian Stryker (2008, page 13) commented that:

> [o]ne's gender identity could perhaps best be described as how one feels about being referred to by a particular pronoun.

In what follows, we first discuss how gender is analyzed sociologically (§3.1), then how gender is reflected in language (§3.2), and finally how these two converge or diverge (§3.3). Only by carefully examining these two constructs, and their complicated relationship, will we be able to tease apart different forms of gender bias in NLP systems.

3.1 Sociological Gender

Many modern trans-inclusive models of gender recognize that *gender* encompasses many different aspects. These aspects include the experience that one has of gender (or lack thereof), the way that one expresses one's gender to the world, and the way that normative social conditions impose gender norms, typically as a dichotomy between masculine and feminine roles or traits (Kramarae and Treichler 1985). The latter two notions are captured by the "doing gender" model from social constructionism, which views gender as something that one *does* and to which one is socially *accountable* (West and Zimmerman 1987; Butler 1989; Risman 2009). However, viewing gender purely through the lens of expression and accountability does not capture the first aspect: one's experience of one's own gender (Serano 2007).

Such trans-inclusive views deconflate anatomical and biological traits and the sex that a person had assigned to them at birth from one's gendered position in society; this includes intersex people, whose anatomical/biological factors do not match the usual designational criteria for either sex. Trans-inclusive views further typically recognize that gender exists beyond the regressive "female"/"male" binary;[8] additionally, that one's gender may shift by time or context (often "genderfluid"), and some people do not experience gender at all (often "agender") (Kessler and McKenna 1978; Schilt and Westbrook 2009; Darwin 2017; Richards, Bouman, and Barker 2017). These models of gender contrast with "folk" views (that are prevalent both in linguistics, sociology, and science more broadly, as well as many societies at large), which assume that one's gen-

8 Some authors use female/male for sex and woman/man for gender; we do not need this distinction (which is itself contestable) and use female/male for gender.

der is defined by one's anatomy (and/or chromosomes), that gender is binary between "male" and "female," and that one's gender is immutable—all of which are inconsistent with reality as it has been known for at least two thousand years.[9] In §4.1 we will analyze the degree to which NLP papers make assumptions that are trans-inclusive or trans-exclusive.

Social gender[10] refers to the imposition of gender roles or traits based on normative social conditions (Kramarae and Treichler 1985), which often includes imposing a dichotomy between feminine and masculine (in behavior, dress, speech, occupation, societal roles, etc.). Taking gender role as an example, upon learning that a nurse is coming to their hospital room, a patient may form expectations that this person is likely to be "female," which in turn may generate expectations around how their face or body may look, how they are likely to be dressed, how and where hair may appear, how to refer to them, and so on. This process, often referred to as **gendering** (Serano 2007), occurs both in real-world interactions, as well as in purely linguistic settings (e.g., reading a newspaper), in which readers may use social gender clues to assign gender(s) to the real world people being discussed. For instance, it is social gender that may cause an inference that my cousin is female in "My cousin is a librarian" or "My cousin is beautiful."

3.2 Linguistic Gender

Our discussion of linguistic gender largely follows previous work (Corbett 1991; Ochs 1992; Craig 1994; Corbett 2013; Hellinger and Motschenbacher 2015), departing from earlier characterizations that postulate a direct mapping from language to gender (Lakoff 1975; Silverstein 1979). Here, it is useful to distinguish multiple ways in which gender is realized linguistically (see also Fuertes-Olivera [2007] for a similar overview). Our taxonomy is related but not identical to Ackerman (2019), which we discussed in §2.

Grammatical gender, similarly defined in Ackerman (2019), is nothing more than a classification of nouns based on a principle of *grammatical agreement*. It is useful to distinguish between "gender languages" and "noun class languages." The former have two or three grammatical genders that have, for animate or personal references, considerable correspondence between a FEM (resp., MASC) grammatical gender and referents with female- (resp., male-)[11] social gender. In comparison, "noun class languages" have no such correspondence, and typically have many more gender classes. Some languages have no grammatical gender at all; English is generally seen as one (viewing that referential agreement of personal pronouns does not count as a form of grammatical agreement, a view which we follow, but one that is contested [Nissen 2002; Baron 1971; Bjorkman 2017]).

9 As identified by Keyes (2018), references appear as early as CE 189 in the Mishnah (HaNasi 189). Similar references (with various interpretations) also appear in the Kama Sutra (Burton 1883, Chapter IX), which dates sometime between BCE 400 and CE 300. Archaeological and linguistic evidence also depicts the lives of trans individuals around 500 BCE in North America (Bruhns 2006) and around 2000 BCE in Assyria (Neill 2008).

10 Ackerman (2019) highlights a highly overlapping concept, "bio-social gender," which consists of gender role, gender expression, and gender identity.

11 One difficulty in this discussion is that linguistic gender and social gender use the terms "feminine" and "masculine" differently; to avoid confusion, when referring to the linguistic properties, we use FEM and MASC.

Referential gender (similar, but not identical to Ackerman [2019] "conceptual gender") relates linguistic expressions to extra-linguistic reality, typically identifying referents as "female," "male," or "gender-indefinite." Fundamentally, referential gender *only* exists when there is an entity being referred to, and their gender (or sex) is realized linguistically. The most obvious examples in English are gendered third person pronouns (SHE, HE), including neopronouns (ZE, EM) and singular THEY,[12] but also includes cases like "policeman" when the intended referent of this noun has social gender "male." (Note that this is *different* from the case when "policeman" is used exclusively non-referentially, as in "every policeman needs to hold others accountable," in which setting this is a case of *lexical gender*, as follows).

Lexical gender refers to extra-linguistic properties of female-ness or male-ness in a *non-referential* way, as in terms like "mother" or "uncle" as well as gendered terms of address like "Mrs." and "Sir." Importantly, lexical gender is a property of the linguistic unit, *not* a property of its referent in the real world, which may or may not exist. For instance, in "Every son loves his parents," there is no real world referent of "son" (and therefore no *referential* gender), yet it still (likely) takes HIS as a pronoun anaphor because "son" has lexical gender MASC.

We will make use of this taxonomy of linguistic gender in our ablation of annotation biases in §4.3, but first need to discuss ways in which notions of linguistic gender match (or mismatch) from notions of social gender.

3.3 Interplays Between Social and Linguistic Gender

The inter-relationship between all these types of gender is complex, and none is one-to-one. An individuals' gender identity may mismatch with their gender expression (at a given point in time). The referential gender of an individual (e.g., pronouns in the case of English) may or may not match either their gender identity or expression, and this may change by context. This can happen in the case of people whose everyday life experience of their gender fluctuates over time (at any interval), as well as in the case of drag performers (e.g., some men who perform drag are addressed as SHE while performing, and HE when not [Anonymous 2017; Butler 1989]).

The other linguistic forms of gender (grammatical, lexical) also need not match each other, nor match referential gender. For instance, a common example is the German term "Mädchen," meaning "girl" (e.g., Hellinger and Motschenbacher 2015). This term is grammatically neuter (due to the diminutive "-chen" suffix), has lexical gender as "female," and generally (but not exclusively) has female referential gender (by being used to refer to people whose gender is female). The idiom "Mädchen für alles" ("girl for everything," somewhat like "handyman") allows for male referents, sometimes with a derogatory connotation and sometimes with a connotation of appreciation.[13]

Social gender (societal expectations, in particular) captures the observation that upon hearing "My cousin is a librarian," many speakers will infer "female" for "cousin," because of either an entailment of "librarian" or some sort of probabilistic inference (Lyons 1977), but not based on either grammatical gender (which does not exist anyway in English) or lexical gender. Such inferences can also happen due to interplays between social gender and heteronormativity. This can happen in cases like "X's husband," in which some listeners may infer female social gender for "X," as well as in ambiguous

12 People's mental acceptability of singular THEY is still relatively low even with its increased usage (Prasad and Morris 2020), and depends on context (Conrod 2018a).

13 Dahl (2000) provides several complications of this analysis.

cases like "X's spouse," in which some listeners may infer "opposite" genders for "X" and their spouse (the inference of "opposite" additionally implies a gender binary assumption).

In this paper, we focus exclusively on English, which has no grammatical gender, but does have lexical gender (e.g., in kinship terms like "mother" and forms of address like "Mrs."). English also marks referential gender on singular third person pronouns.

English THEY, in particular, is tricky, because it can be used to refer to: plural non-humans (e.g., a set of boxes), plural humans (e.g., a group of scientists), a quantified human of unknown or irrelevant gender ("Every student loves their grade"), an indefinite human of unknown or irrelevant gender ("A student forgot their backpack"), a definite specific human of unknown gender, or one of non-binary gender ("Parker saw themself in the mirror").[14] This ontology is due to Conrod (2018b), who also investigates the degree to which these are judged grammatical by native English speakers, and which we will use to quantify data bias (§4.3).

Below, we use this more nuanced notion of different types of gender to inspect where in the machine learning lifecycle for English coreference resolution different types of bias play out. These biases may arise in the context of any of these notions of gender, and we encourage future work to extend care over what notions of gender are being utilized and when.

4. Sources of Bias

In this section, we analyze several ways in which harmful biases can and do enter into the machine learning lifecycle of coreference resolution systems (per Figure 1). Two stages discussed by Vaughan and Wallach (2019) that we exclude are Training Process and Deployment. It is rare (as they observed as well) for training processes (especially in batch learning settings) to lead to bias, and the same appears to be the case here. We do not consider the "Deployment" phase, because we are not aware of deployed coreference resolution systems to test—except, perhaps, those embedded in other systems, which we discuss in the context of testing (§4.6).

4.1 Bias in: Task Definition

Task definitions for linguistic annotations (like coreference) tend, in NLP, to be described in annotation guidelines (or, more recently, in datasheets or data statements [Gebru et al. 2018; Bender and Friedman 2018]). These guidelines naturally change over the years as the community understands more and more about both the task and the annotation process (this is part of what makes the lifecycle a *cycle*, rather than a pipeline). Getting annotation guidelines "right" is difficult, particularly in balancing informativeness with ability to achieve inter-annotator agreement, and important because poorly defined tasks lead to a substantial amount of wasted research effort.

For the purposes of this study, we consider here (and elsewhere in this paper) thirteen data sets on which coreference or anaphora are annotated in English (Table 1); eleven of these are corpora distributed by the Linguistic Data Consortium (LDC),[15] and two are not. According to the authors of the QB data set (personal communication),

14 The use of singular they to denote referents of unknown gender dates back to the late 1300s, while the non-binary use of they dates back at least to the 1950s (Merriam-Webster 2016).

15 See `https://catalog.ldc.upenn.edu/{LDC-ID}`.

Table 1
Corpora analyzed in this paper.

MUC7	Message Understanding Conference 7	2001T02
Zh-PB3	Chinese Propbank 3.0	2003T13
ACE04	Automatic Content Extraction 2004	2005T09
BBN	BBN Pronoun Coreference Corpus	2005T33
ACE05	Automatic Content Extraction 2005	2006T06
LUAC	Language Understanding Annotation Corpus	2009T10
NXT	NXT Switchboard Annotations	2009T26
MASC3	Manually Annotated Sub-Corpus	2013T12
Onto5	Ontonotes Version 5	2013T19
ACE07	Automatic Content Extraction 2007	2014T18
AMR2	Abstract Meaning Representation v 2	2017T10
GAP	Gendered Ambiguous Pronouns	Webster et al. (2018)
QB	QuizBowl Coreferences	Guha et al. (2015)

it was annotated under the OntoNotes guidelines, with the exception that singleton mentions were also annotated. The final data set, GAP, did not explicitly annotate full coreference, but rather annotated a binary choice of which of two names a pronoun refers to (as described in the associated paper). None of the annotation guidelines (or papers) give explicit guidance about what personal pronouns are to be considered (with the exception of GAP, which explicitly limits to HE/SHE pronouns), or otherwise what information a human annotator should use to resolve ambiguous situations. However, many of them do provide running examples, which we can analyze. Although the examples in annotation guidelines (or papers) are insufficient to fully tell what annotations were intended by the authors, they do provide a sense of what may have been top of mind in data set construction.

To assess task definition bias, we count, for each of the annotation guidelines, how many examples use different pronominal forms. For examples that use THEY, we separate four different subtypes, following Conrods (2018b) categorization (see §3.2):

NH: Plural non-human group – "The knives are put away in their carrier."

PL: Plural group of humans – "The children are friendly, and they are happy."

QI: Quantified/indefinite – "Most chefs harshly critique their own dishes."

SP: Specific singular referent – "Jun enjoys teaching their students."

The results are shown in Table 2 (data sets for which no relevant examples were provided are not listed). Overall, we see that in total across these seven data sets, examples with HE occur more than twice as frequently as all others combined. Furthermore, THEY is never used in a specific setting and, somewhat interestingly, is only used as an example for quantification in *older* data sets (2005 and before). Moreover, none of the annotation guidelines have examples using neopronouns. This lack does nothing to counterbalance a general societal bias that tends to erase non-binary identities. In the case of GAP, it is explicitly mentioned that only SHE and HE examples are considered (and only in cases where the gender of two possible referents "matches"—though it is unspecified what type of gender this is and how it is determined). Even on the binary spectrum, there is also an obvious gender bias between HE and SHE examples.

Tasks defined in these thirteen data sets only consider binary gender and are mostly male-dominated. Systems built along with such task definitions can hardly function for

Table 2
Frequency counts of different pronouns in example annotations given in annotation instructions for seven of the data sets that provide examples. Zeros are omitted from the table. No data sets contain examples using neopronouns, nor any examples using "they" to refer to a singular specific entity (and only older data sets included any examples of quantified usages of "they").

				THEY			
	SHE	HE	NEO	SP	QI	PL	NH
MUC7		7			2	1	
Zh-PB3		4			3		
ACE04	2	6			1		
LUAC	1	2					
Onto5	1						
AMR2	5	17					
QB		1					
Total	9	37	0	0	6	1	0

non-binary and female users. See §4.6 for detailed analysis of system performances on data with both binary and non-binary pronouns.

4.2 Bias in: Data Input

In coreference resolution, as in most NLP data collection settings, one typically first collects raw text and then has human annotators label that text. Here, we consider biases that arise due to the selection of what texts to have annotated. As an example, if a data curator chooses newswire text from certain sources as source material, xey are unlikely to observe singular uses of THEY which, for instance, was only added to the Washington Post style guide in late 2015 (Walsh 2015) and by the Associated Press Stylebook in early 2017 (Andrews 2017). If the raw data does not contain certain phenomena, this fundamentally limits all further stages in the machine learning lifecycle (a system that has never seen "hir" is unlikely to even know it is a pronoun, much less how to link it; indeed the off-the-shelf tokenizer we used often failed to separate "xey're" into two tokens, as it does for "they're").

To analyze the possible impact of input data, we consider our thirteen coreference data sets, and count how many instances of different types of pronouns are used in the raw data. We focus on SHE, HE, and THEY pronouns (in all their morphological forms); we additionally counted several neopronoun forms (HIR, XEY, and EY) and found no occurrences nor their morphological variances.[16] In the case of THEY, we again distinguish between its four usage cases: plural, singular, quantified, and non-human. To achieve this, we annotated 100 examples uniformly at random by hand from OntoNotes (Weischedel et al. 2011). Furthermore, we compare them to the raw counts in a 2015 dump from some of the Reddit Discussion Forum,[17] and also limited to the `genderqueer` subforum. We additionally include a new data set for this study, GICoref.

16 There was one instance of "hir" but that was a almost certainly a typo for "his" (given the context), and several instances of "ey" used as contractions for plural THEY in transcripts of spoken English.

17 It is a data set with publicly available comments from Reddit. The data set has about 1.7 billion comments with their related fields, such as score, author, subreddits, etc. Here is the link to the data set `www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/`

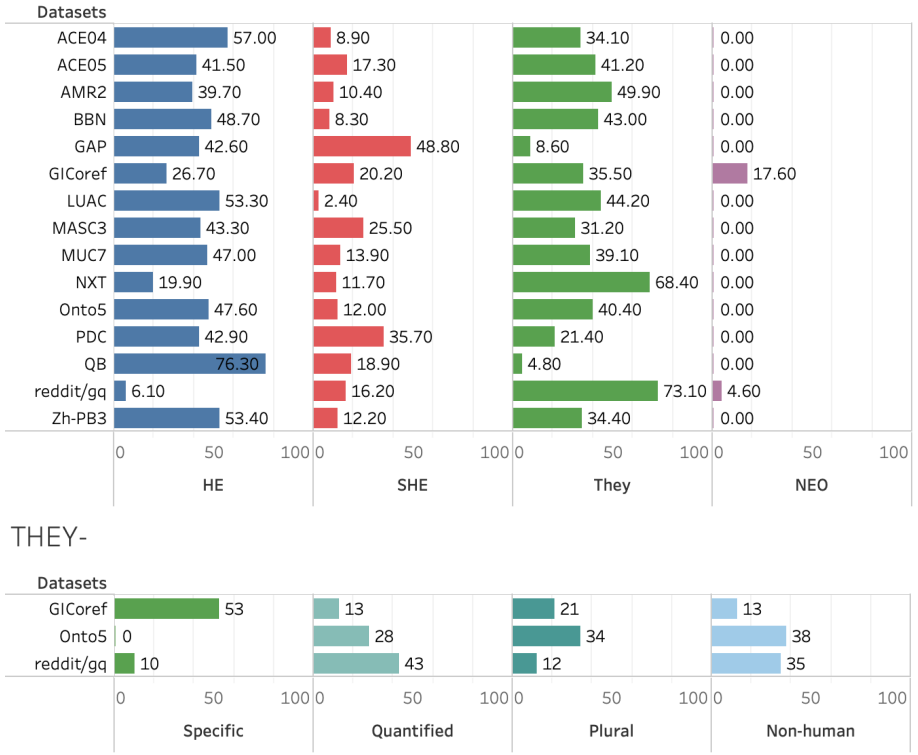

Figure 2
(Top) For each data set under consideration, the fraction of pronouns with different forms. Only our data set (GICoref) and the genderqueer subreddit include neopronouns. (In the case of GAP, there are some occurrences of THEY, but they are never considered targets for coreference and so we exclude them from these counts.) (Bottom) For three data sets, the fraction (out of 100 annotated) of each *they* into one of the four usage cases.

This new data set is collected to evaluate current coreference resolution systems on gender-inclusive and naturally occuring texts. Details of the data set are described in §4.5.3.

The results of this analysis are in Figure 2. Overall, the examples used in the documentation of each of these data sets focuses entirely on binary gendered pronouns, generally with many more HE examples than SHE examples. Only the older data sets (MUC7 and Zh-PB3) include any examples of THEY, some of which are in a quantified form.

Systems trained from these data sets never see non-binary pronouns during training. Thus, when generalizing, system performance for non-binary users on singular THEY or neopronouns surely will be worse.

4.3 Bias in: Data Annotation

A significant possible source of bias comes from annotations themselves, arising from a combination of (possibly) underspecified annotations guidelines and the positionality of annotators themselves. Ackerman (2019, page 14) analyzes how humans cognitively encode gender in resolving coreferences through a Broad Matching Criterion, which

posits "matching gender requires at least one level of the mental or the environment context] to be identical to the candidate antecedent in order to match." In this section, we delve into the linguistic notions of gender and study how different aspects of linguistic notions impact an annotator's judgments of anaphora.

Our study can be seen as evaluating which conceptual properties of gender are most salient in human annotation judgments. We start with natural text in which we can cast the coreference task as a binary classification problem ("which of these two names does this pronoun refer to?") inspired by Webster et al. (2018). We then generate "counterfactual augmentations" of this data set by ablating the various notions of linguistic gender described in §3.2, similar to Zmigrod et al. (2019) and Zhao et al. (2018a). We finally evaluate the impact of these ablations on human annotation behavior to answer the question: Which forms of linguistic knowledge are most essential for human annotators to make consistent judgments?

As motivation, consider (1) below, in which an annotator is likely to determine that "her" refers to "Mary" and not "John" due to assumptions on likely ways that names may map to pronouns (or possibly by not considering that SHE pronouns could refer to someone named "John"). Whereas in (2), an annotator is likely to have difficulty making a determination because both "Sue" and "Mary" suggest "her." In (3), an annotator lacking knowledge of name stereotypes on typical Chinese and Indian names (plus the fact that given names in Chinese—especially when romanized—generally do not signal gender strongly), respectively, will likewise have difficulty.

(1) John and Mary visited *her* mother.

(2) Sue and Mary visited *her* mother.

(3) Liang and Aditya visited her mother.

In all these cases, the plausible rough inference is that a reader takes a name, and uses it to infer the sociological gender of the extra-linguistic referent. Later the reader sees the SHE pronoun, infers the referential gender of that pronoun, and checks to see if they match.

An equivalent inference happens not just for names, but also for lexical gender references (both gendered nouns (4) and terms of address (5)), grammatical gender references (in gender languages like Arabic (6)), and social gender references (7). The last of these ((7)) is the case in which the correct referent is likely to be least clear to most annotators, and also the case studied by Rudinger et al. (2018) and Zhao et al. (2018a).

(4) My brother and niece visited *her* mother.

(5) Mr. Hashimoto and Mrs. Iwu visited *her* mother.

(6) المطرب و الممثلة شاهدا والدتها
walidatu *-ha* shahadaa al-momathela w almutarab
mother *-her* saw actor[FEM] and singer[MASC]
The singer[MASC] and actor[FEM] saw *her* mother.

(7) The nurse and the actor visited *her* mother.

4.3.1 Ablation Methodology. In order to determine which cues annotators are using and the degree to which they use them, we construct an ablation study in which we hide

various aspects of gender and evaluate how this impacts annotators' judgments of anaphoricity. To make the task easier for crowdsourcing, we follow the methodology of Webster et al.'s (2018) GAP data set for studying ambiguous binary gendered pronouns. In particular, we construct binary classification examples taken from Wikipedia pages, in which a single pronoun is selected, and two possible antecedent names are given, and the annotator must select which one. We cannot use the GAP data set directly, because their data is constrained so that the "gender" of the two possible antecedents is "the same";[18] for us, we are specifically interested in how annotators make decisions even when additional gender information is available. Thus, we construct a data set called *Maybe Ambiguous Pronoun* (MAP), which is similar to the GAP data set, but where we do not restrict the two names to match gender so that we can measure the influence of different gender cues.

In ablating gender information, one challenge is that removing social gender cues (e.g., "nurse" tending female) is not possible because they can exist anywhere. Likewise, it is not possible to remove syntactic cues in a non-circular manner. For example, in (8), syntactic structure strongly suggests the antecedent of "herself" is "Liang," making it less likely that "He" corefers with Liang later (though it is possible, and such cases exist in natural data due either to genderfluidity or misgendering).

(8) Liang saw herself in the mirror… *He* …

Fortunately, it is possible to enumerate a high coverage list of English terms that signal lexical gender: terms of address (Mrs., Mr.) and semantically gendered nouns (mother).[19] We assembled a list by taking many online lists (mostly targeted at English language learners), merging them, and manual filtering. The assembling process and the final list is published with the MAP data set and its datasheet.
To execute the "hiding" of various aspects of gender, we use the following substitutions:

(a) ¬PRO: Replace all third person singular pronouns with a gender neutral program (THEY, XEY, ZE).

(b) ¬NAME: Replace all names (e.g., "Aditya Modi") by a random name with only a first initial and last name (e.g., "B. Hernandez").

(c) ¬SEM: Replace all semantically gendered nouns (e.g., "mother") with a gender-indefinite variant (e.g., "parent").

(d) ¬ADDR: Remove all terms of address (e.g., "Mrs.," "Sir").[20]

See Figure 3 for an example of all substitutions.

We perform two sets of experiments, one following a "forward selection" type ablation (start with everything removed and add each back in one-at-a-time) and one following "backward selection" (remove each separately). Forward selection is necessary in order to de-conflate syntactic cues from stereotypes, whereas backward selection

18 It is unclear from the GAP data set what notion of "gender" is used, nor how it was determined to be "the same."

19 These are, however, sometimes complex. For instance, "actress" signals *lexical* gender of female, while "actor" may signal *social* gender of male and, in certain varieties of English, may also signal *lexical* gender of male.

20 An alternative suggested by Cassidy Henry that we did not explore would be to replace all with Mx. or Dr.

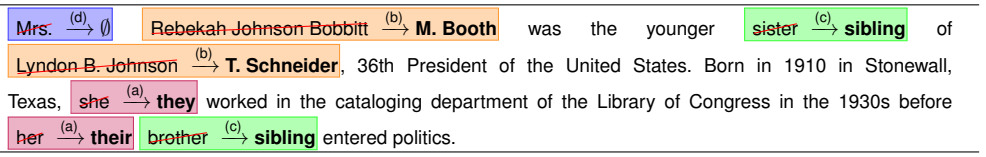

Figure 3
Example of applying all ablation substitutions for an example context in the MAP corpus. Each substitution type is marked over the arrow and separately color-coded.

gives a sense of how much impact each type of gender cue has in the context of all the others.

We begin with ZERO, in which we apply all four substitutions. Since this also removes gender cues from the pronouns themselves, an annotator cannot substantially rely on social gender to perform these resolutions. We next consider adding back in the original pronouns (always HE or SHE here), yielding ¬NAME ¬SEM ¬ADDR. Any difference in annotation behavior between ZERO and ¬NAME ¬SEM ¬ADDR can only be due to social gender stereotypes.

To see why, consider the example from Figure 3. In this case, the only difference between the Zero setting and the ¬NAME ¬SEM ¬ADDR is whether the pronouns SHE/HER are substituted with THEY/THEIR—all other substitutions are applied in both cases. In the ZERO case, there are no gender cues at all to help with the resolution, precisely because gender has been removed even from the pronoun. So even if there were gendered information in the rest of the text, that logically cannot help with the resolution of either of the pronouns. In the ¬NAME ¬SEM ¬ADDR case, all lexical and referential gender information *except* that on the pronoun have been removed (as English has no grammatical gender). If one accepts that the taxonomy of gender from §3.2 is complete, then this means that the only gender information that exists in the rest of this example is social gender. (And indeed, there is social gender—even in a fictitious world in which someone named T. Schneider was the 36th President of the U.S., social gender roles suggest that this person is relatively unlikely to be the referent of *she*). Thus, it is likely that in this latter case, readers and annotators can and will use the gender information on the pronoun to decide that SHE does not refer to Schneider and therefore likely refers to Booth. On the other hand, in the ZERO case, there is no such information available, and a reader must rely, perhaps, on parallel syntactic structure or centering (Joshi and Weinstein 1981; Grosz, Joshi, and Weinstein 1983; Poesio et al. 2004) if ey are to correctly identify that the referent is Booth.

The next setting, ¬SEM ¬ADDR, removes both forms of lexical gender (semantically gendered nouns and terms of address); differences between ¬SEM ¬ADDR and ¬NAME ¬SEM ¬ADDR show how much names are relied on for annotation. Similarly, ¬NAME ¬ADDR removes names and terms of address, showing the impact of semantically gendered nouns, and ¬NAME ¬SEM removes names and semantically gendered nouns, showing the impact of terms of address.

In the backward selection case, we begin with ORIG, which is the unmodified original text. To this, we can apply the pronoun filter to get ¬PRO; differences in annotation between ORIG and ¬PRO give a measure of how much *any* sort of gender-based inference is used. Similarly, we obtain ¬NAME by only removing names, which gives a measure of how much names are used (in the context of all other cues); we obtain ¬SEM by only removing semantically gendered words; and ¬ADDR by only removing terms of address.

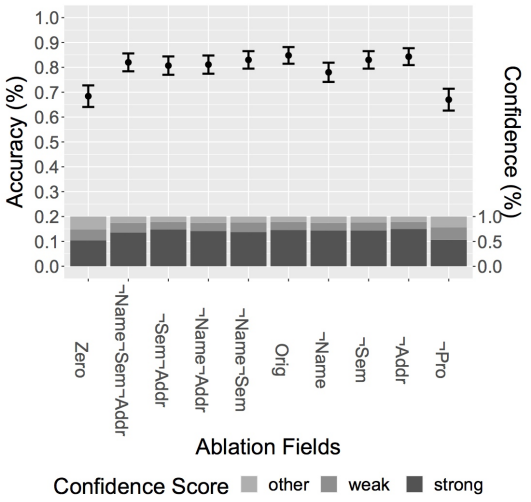

Figure 4
Human annotation results for the ablation study on MAP data set. Each column is a different ablation, and the *y*-axis is the degree of accuracy with 95% significance intervals. Bottom bar plots are annotator certainties as how sure they are in their choices.

4.4 Annotation Results

We construct examples using the methodology defined above. We then conduct annotation experiments using crowdworkers on Amazon Mechanical Turk following the methodology by which the original GAP corpus was created.[21] Because we wanted to also capture uncertainty, we ask the crowdworkers how sure they are in their choices, between "definitely" sure, "probably" sure, and "unsure."[22]

Figure 4 shows the human annotation results as binary classification accuracy for resolving the pronoun to the antecedent. We can see that removing pronouns leads to a significant drop in accuracy. This indicates that gender-based inferences, especially social gender stereotypes, play the most significant role when annotators resolve coreferences. This confirms the findings of Rudinger et al. (2018) and Zhao et al. (2018a) that human-annotated data incorporates bias from stereotypes.

Moreover, if we compare ORIG with columns to the left, we see that name is another significant cue for annotator judgments, while lexical gender cues do not have significant impacts on human annotation accuracies. This is likely in part due to the

21 Our study was approved by the Microsoft Research Ethics Board. We recruited workers from countries with large native English-speaking populations (Australia, Canada, New Zealand, United Kingdom, and United States), and who have greater than 80% HIT approval rate and more than 100 HITs approved. Workers were paid $1 to annotate ten contexts (the average annotation time was seven minutes). Crowdworkers were informed as part of the instructions and examples that they should expect to see both singular THEY and neopronouns, with examples of each.

22 In some of the examples, a crowdworker may apply knowledge of the situation or entities involved, for instance, "President of the United States" has, to date, always been referred to using HE pronouns. To capture this, we additionally asked the crowdworkers if they recognized any of the entities or the situation involved. Note, however, that even though this is true in the real world, it is not hard to imagine fictional contexts in which a human annotator would have no difficulty finding "President of the United States" to corefer with SHE or THEY pronouns.

low appearance frequency of lexical gender cues in our data set. Every example has pronouns and names, whereas 49% of the examples have semantically gendered nouns but only 3% of the examples include terms of address. We also note that if we compare ¬NAME ¬SEM ¬ADDR to ¬SEM ¬ADDR and ¬NAME ¬ADDR, accuracy drops when removing gender cues. Though the differences are not statistically significant, we did not expect the accuracy drop.

Finally, we find annotators' confidence follow the same trend as the accuracy: Annotators have a reasonable sense of when they are unsure. We also note that accuracy scores are essentially the same for ZERO and ¬PRO, which suggests that once explicit binary gender is gone from pronouns, the impact of any other form of linguistic gender in annotator decisions is also removed.

Overall, we can see that annotators may make unlicenced inferences of various gender cues by conflating gender concepts. Thus, systems trained by treating these annotator judgments as ground truth can be problematic for both binary and non-binary people.

4.4.1 Limitations of (Approximate) Counterfactual Text Manipulation. Any text manipulation—like we have done in this section—runs the risk of missing out on how a human author might truly have written that text under the presumed counterfactual. For example, a speaker uttering 1 may assume that aer interlocutor shares, or at least recognizes, social biases that lead one to assume that the person named "John" is likely referred to as HE and "Mary" as SHE. This speaker may use this assumption of the listener to determine that "her" is sufficiently unambiguous in this case as to be an acceptable reference (trading off brevity and specificity; see, for instance Arnold [2008], Frank and Goodman [2012], Orita et al. [2015a]). However, if we "counterfactually" replaced the names "John" and "Mary" to "H. Martinez" and "R. Modi" (respectively), it is unlikely that the supposed speaker would make the same decision. In this case, the speaker may well have said "Modi's mother" or some other reference that would have been sufficiently specific to resolve, even at the cost of being more wordy. That is to say, the counterfactual replacements here and their effect on human annotation agreement should be taken as a sort of *upper bound* on the effect one would expect in a truly counterfactual setting.

Moreover, although we studied crowdworkers on Mechanical Turk (because they are often employed as annotators for NLP resources), if other populations are used for annotation, it becomes important to consider their positionality and how that may impact annotations. This echoes a related finding in annotation of hate-speech that annotator positionality matters (Olteanu et al. 2019).

4.5 Bias in: Model Definition

Bias in machine learning systems can also come from how models are structured—for instance, what features they use, and what baked-in decisions are made. For instance, some models may simply fail to recognize anything other than a dictionary of fixed pronouns as possible entities. Others may use external resources, such as lists that map names to guesses of "gender," that bake in stereotypes around naming.

In this section, we analyze prior work in systems for coreference resolution in three ways. First, we do a literature study to quantify how NLP papers discuss gender, broadly. Second, similar to Zhao et al. (2018a) and Rudinger et al. (2018), we evaluate a handful of freely available systems on the ablated data from §4.3. Third, we evaluate these systems on the data set we created: Gender Inclusive Coreference (GICoref).

Table 3
Analysis of a corpus of 150 NLP papers that mention "gender" along the lines of what assumptions around gender are implicitly or explicitly made in the work.

	All Papers		Coref Papers	
Paper Discusses Linguistic Gender?	52.6%	(79/150)	95.4%	(21/22)
Paper Discusses Social Gender?	58.0%	(87/150)	86.3%	(19/22)
Paper Distinguishes Linguistic from Social Gender?	11.1%	(3/27)	5.5%	(1/18)
Paper assumes Social Gender is Binary?	92.8%	(78/84)	94.4%	(17/18)
Paper Assumes Social Gender is Immutable?	94.5%	(70/74)	100.0%	(14/14)
Paper Allows for Neopronouns/THEY-SP?	3.5%	(2/56)	7.1%	(1/14)

4.5.1 Cis-normativity in Published NLP Papers. In our first study, we adapt the approach Keyes (2018) took for analyzing the degree to which computer vision papers encoded trans-exclusive models of gender. In particular, we begin with a random sample of 150 papers from the ACL anthology that mention the word "gender" and coded them according to the following questions:

- Does the paper discuss coreference or anaphora resolution?
- Does the paper study English (and possibly other languages)?
- Does the paper deal with linguistic gender (i.e., grammatical gender or gendered pronouns)?
- Does the paper deal with social gender?
- (If yes to the previous two:) Does the paper explicitly distinguish linguistic from social gender?
- (If yes to social gender:) Does the paper explicitly recognize that social gender is not binary?
- (If yes to social gender:) Does the paper explicitly or implicitly assume social gender is immutable?[23]
- (If yes to social gender and to English:) Does the paper explicitly consider uses of definite singular "they" or neopronouns?

The results of this coding are in Table 3 and the list of the full set of annotations is in Appendix A. Here, we see out of the 22 coreference papers analyzed, the vast majority conform to a "folk" theory of language:

- ◇ Only 5.5% distinguish social from linguistic gender (despite it being relevant);
- ◇ Only 5.6% explicitly model gender as inclusive of non-binary identities;

23 The most common ways in which papers implicitly assume that social gender is immutable is either 1) by relying on external knowledge bases that map names to "gender"; or 2) by scraping a history of a user's social media posts or emails and assuming that their "gender" today matches the gender of that historical record.

- ◇ No papers treat gender as anything other than completely immutable;
- ◇ Only 7.1% (one paper) considers neopronouns and/or specific singular THEY.

The situation for papers not specifically about coreference is similar (the majority of these papers are either purely linguistic papers about grammatical gender in languages other than English, or papers that do "gender recognition" of authors based on their writing; May [2019] discusses the (re)production of gender in automated gender recognition in NLP in much more detail). Overall, the situation more broadly is equally troubling, and generally also fails to escape from the folk theory of gender. In particular, none of the differences between papers and papers about coreference are significant at a $p < 0.05$ level except for the first two questions, due to the small sample size (according to an $n - 1$ chi-squared test). The result of this analysis is that although we do not know exactly what decisions are baked in to all models, the vast majority in our study (including two papers by one of the authors [Daumé and Marcu 2005; Orita et al. 2015b]) come with strong gender binary assumptions, and exist within a broader sphere of literature which erases non-binary identities.

4.5.2 Coreference System Performance on MAP. Next, we analyze the effect that our different ablation mechanisms have on existing coreference resolutions systems. In particular, we run five coreference resolution systems on our ablated data: the AI2 system (AI2; Gardner et al. 2017), hugging face (HF; Wolf 2017), which is a neural system based on spacy (Honnibal and Montani 2017), and the Stanford deterministic (SfdD; Raghunathan et al. 2010), statistical (SfdS; Clark and Manning 2015), and neural (SfdN; Clark and Manning 2016) systems. Figure 5 shows the results. We can see that the system accuracies mostly follow the same pattern as human accuracy scores, though all are significantly lower than human results. Accuracy scores for systems drop dramatically when we ablate out referential gender in pronouns. This reveals that those coreference

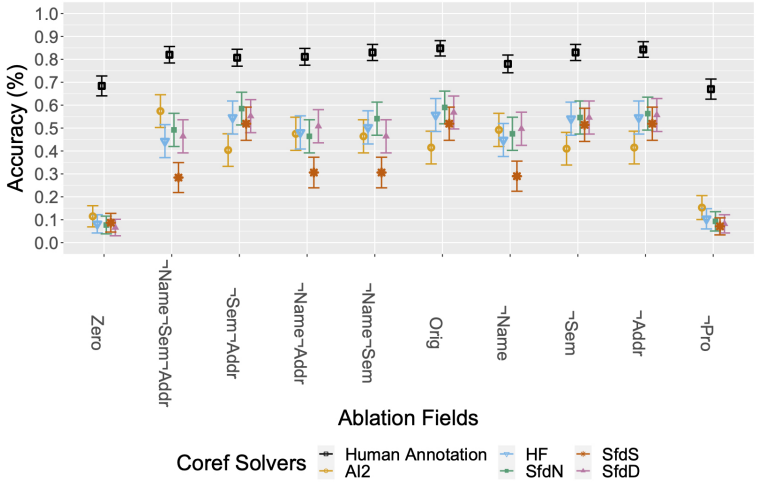

Figure 5
Coreference resolution systems results for the ablation study on MAP data set. The *y*-axis is the degree of accuracy with 95% significance intervals.

resolution systems rely heavily on gender-based inferences. In terms of each system, HF and SfdN systems have similar results and outperform other systems in most cases. SfdD accuracy drops significantly once names are ablated.

These results echo and extend previous observations made by Zhao et al. (2018a), who focus on detecting stereotypes within occupations. They detect gender bias by checking if the system accuracies are the same for cases that can be resolved by syntactic cues and cases that cannot, with original data and reversed-gender data. Similarly, Rudinger et al. (2018) focus on detecting stereotypes within occupations as well. They construct the data set without any gender cues other than stereotypes, and check how systems perform with different pronouns—THEY, SHE, HE. Ideally, they should all perform the same because there is not any gender cues in the sentence. However, they find that systems do not work on "they" and perform better on "he" than "she." Our analysis breaks this stereotyping down further to detect which aspects of gender signals are most leveraged by current systems.

4.5.3 Coreference System Performance on GICoref. We introduce a new data set, GICoref, for the purpose of evaluating current coreference resolution systems in the contexts where a broader range of gender identities are reflected, where linguistic examples of genderfluidity are encountered, where non-binary pronouns are used, and where misgendering happens. In comparison to Zhao et al. (2018a) and Rudinger et al. (2018) (as well as in contrast to our MAP data set), we focused on naturally occurring data, but sampled specifically to surface more gender-related phenomena than may be found in, say, the Wall Street Journal.

The GICoref data set consists of 95 documents from three types of sources: articles on English Wikipedia about people with non-binary gender identities, articles from LGBTQ periodicals, and fan-fiction stories from Archive Of Our Own[24] (with the respective authors' permission). Each author of this paper manually annotated each of these documents and then we jointly adjudicated the results.[25] To reduce annotation time, any article that was substantially longer than 1,000 words (pre-tokenization) was trimmed at the 1,000th word.[26] This data includes many examples of people who use pronouns other than SHE or HE, people who are genderfluid and whose name or pronouns changes through the article, people who are misgendered, and people in relationships that are not heteronormative. One annotation decision we made was around the specific case of people who perform drag. Following Butler (1989), in our annotation we considered drag performance as a form of genderfluidity; as such, we annotate the performer and the drag persona as coreferent with each other (as well as the relevant pronouns), akin to how we believed a reasonable model for handling stage names (e.g., Christopher Wallace / Notorious B.I.G.) would mark them as coreferent, while famous roles played by multiple actors (e.g., Carrie, played by Sissy Spacek, Angela Bettis, and Chloë Grace-Moretz) would be marked as non-coreferent. In addition,

24 See `https://archiveofourown.org`; thanks to Os Keyes for this suggestion.

25 We evaluate inter-annotator agreement by treating one annotation as the gold standard and the other as system output and computing the LEA metric; the resulting F1-score is 92%. During the adjudication process we found that most of the disagreements are due to one of the authors missing/overlooking mentions, and rarely due to true "disagreement."

26 There are four documents we accidentally trimmed at the 2,000th word and so we keep the longer version of them with 2,000 words in the data set.

incorrect references (misgendering and deadnaming[27]) are explicitly annotated.[28] Two example annotated documents, one from Wikipedia, and one from Archive of Our Own, are provided in Appendix B and Appendix C, respectively.

Although the majority of the examples in the data set are set in a Western context, we endeavored to have a broader range of experiences represented. We included articles about people who are gender non-conforming, but where sociological notions of gender mismatch the general sex/gender/sexuality taxonomy of the West. This includes people who identify as *hijra* (Indian subcontinent), *phuying* (Thailand, sometimes referred to as *kathoey*), *muxe* (Oaxaca), *two-spirit* (Americas), *fa'afafine* (Samoa), and *māhū* (Hawaii) individuals.

We run the same systems as before on this data set. Table 4 reports results according to the standard coreference resolution evaluation metric LEA (Moosavi and Strube 2016). It is not clear how systems or evaluation metrics should handle incorrect references (misgendering and deadnaming). Taking (9) as an example, should the misgendering entities and pronouns (cluster c) be included as a coreference to the person (cluster a) or not? If the person is a real human, including the misgendering reference as a ground truth may be potentially harmful to the person. Because no systems are implemented to explicitly mark incorrect references, and no current evaluation metrics address this case, we perform the same evaluation twice. One with incorrect references included as regular references in the ground truth (cluster a and cluster c are the same cluster); and the other with incorrect references excluded (cluster a and cluster c are separate clusters). Due to the limited number of incorrect references (0.6% of total references of people) in the data set, the difference in the results are not significant—the difference is less than 0.2% for each entry. Nonetheless, although these are rare, they constitute significant potential harms. Here we only report the results for the latter.

(9) Frisk$_A$ sat in the back of the classroom, silently praying that their$_A$ teacher wouldn't call on them$_A$. They$_A$ were having a bad day and didn't think they$_A$ could be misgendered today. But just their$_A$ luck, their$_A$ teacher$_B$ was staring straight at them$_A$. "Felix$_C$? Do you know the answer?" ... Chara$_D$'s hand shot up. "Ms. Richards$_B$, their$_A$ name is Frisk$_A$, remember?" "Christine$_E$," Ms. Richards$_B$ sighed, ignoring Chara$_D$'s flinch. "His$_C$ name is whatever is on the sheet, the same way yours is. We$_F$ have had this discussion, remember?"

The first observation is that there is still plenty of room for coreference systems to improve; the best performing system achieves an F1 score of 34%, but the Stanford neural system's F1 score on CoNLL-2012 test set reaches 60% (Moosavi 2020). Here are some examples where the HF and the Stanford deterministic system output erroneous resolutions: (10), (11), and (12). As demonstrated, even when there are clear syntactic cues and declaration of preferred pronouns, both systems fail to resolve the coreferences correctly due to various internal biases.

(10) **HF:** The artwork$_B$ consisted of Sulkowicz$_A$, who uses they$_B$/them$_B$ pronouns, carrying a mattress wherever they$_B$ went on campus.

27 According to Clements (2017), deadnaming occurs when someone, intentionally or not, refers to a person who is transgender by the name they used before they transitioned.

28 Thanks to an anonymous reader of a draft version of this paper for this suggestion.

Table 4
(Left) LEA scores on GICoref data set with various coreference resolution systems. Rows are different systems and columns are precision, recall, and F1 scores. When evaluating, we only count exact matches or pronouns and name entities. (Right) Recall scores of coreference resolution systems for detecting binary pronouns, THEY (of any type), and neopronouns.

LEA	Precision	Recall	F1
AI2	40.4%	**2**9.2%	33.9%
HF	**6**8.8%	22.3%	33.6%
SfdD	50.8%	23.9%	32.5%
SfdS	59.8%	24.1%	**3**4.3%
SfdN	59.4%	24.0%	34.2%

Recall	HE/SHE	THEY	NEO
AI2	96.4%	93.0%	25.4%
HF	95.4%	85.0%	21.1%
SfdD	97.6%	96.6%	0.0%
SfdS	96.2%	86.8%	20.4%
SfdN	96.6%	90.1%	0.0%

SfdD: The artwork consisted of Sulkowics$_A$, who uses they/them pronouns$_B$, carrying a mattress wherever they$_C$ went on campus.

(11) **HF & SfdD:** As the son of a military father$_{C,B}$, she$_A$ faced many challenges to be accepted.

(12) **HF:** Soon after Vasold$_A$'s win was announced, ze$_B$ spoke with the school newspaper about zir surprise at winning.
SfdD: Soon after Vasold$_A$'s win was announced, ze spoke with the school newspaper about zir surprise$_B$ at winning.

Additionally, we can see that system precision dominates recall. This is likely partially due to poor recall of pronouns other than HE and SHE. To analyze this, we compute the recall of each system for finding referential pronouns at all, regardless of whether they are correctly linked to their antecedents. Results are shown in Table 4. We find that all systems achieve a recall of at least 95% for binary pronouns, a recall of around 90% on average for THEY, and a recall of around a paltry 13% for neopronouns (two systems—Stanford deterministic and Stanford neural—never identify any neopronouns at all).

Overall, we have shown that current coreference resolution systems fail to escape from the folk theory of gender and rely heavily on gender-based inferences. Therefore, when deployed, these systems can easily make biased inferences that will lead to both direct and indirect harms to binary and non-binary users.

4.6 Bias in: System Testing

Bias can also show up at testing time, due either to data or metrics. For instance, if one evaluates on highly biased data, it will be difficult to capture disparities (akin to the over-representation of light skinned men in computer vision data sets [Buolamwini and Gebru 2018]). Alternatively, evaluation metrics may weight different errors in a way that is incongruous with their harm. For example, depending on the use case, corefering someone's name with an incorrectly gendered pronoun may produce a harm askin to misgendering, potentially leading to a high cost social error (Stryker 2008); evaluation metrics may or may not reflect the true cost of such mistakes.

In terms of data, most coreference resolution systems are evaluated intrinsically, by testing them against gold standard annotations using a variety of metrics; in this case,

all the observations on data bias (§4.2 and §4.3) apply. Sometimes coreference resolution is used as part of a larger system. For instance, information retrieval can use coreference resolution to help accurately rank documents by up-weighting the importance of entities that are referred to frequently (Du and Liddy 1990; Pirkola and Järvelin 1996; Edens et al. 2003). In machine translation, producing correct gendered forms in gender languages often requires coreference to have been solved (Mitkov 1999; Hardmeier and Federico 2010; Guillou 2012; Hardmeier and Guillou 2018). In the translation case, this then raises the question: Which data is being used and how biased is it? It turns out, "quite biased." Even limiting to just SHE and HE pronouns, the bias is significant: four times as many HE than SHE in Europarl (Koehn 2005) and the Common Crawl (Smith et al. 2013), six times as many in News Commentaries (Tiedemann 2012), and fifty times as many in Hong Kong Laws corpus.

In terms of metrics, most intrinsic evaluation is carried out using metrics like MUC (Vilain et al. 1995), ACE (Mitchell et al. 2005), B^3 (Bagga and Baldwin 1998; Stoyanov et al. 2009), or CEAF (Luo 2005) (see also Cai and Strube [2010] for additional discussion and variants). As observed by Agarwal et al. (2019), most of these metrics are rather insensitive to arguably large errors, like the inability to link pronouns to names; to address this, they introduce a new metric to focus specifically on this named entity coreference task. These metrics also generally treat all errors similarly, regardless of whether the error compounds societal injustices (e.g., ignoring an instance of XYR) or not (e.g., ignore an instance of HE), despite the fact that these have vastly different implications from the perspective of justice (e.g., Fraser 2008).

For extrinsic evaluation, the metrics used are those that are appropriate for the downstream task (e.g., machine translation). In the case of machine translation, one can ask whether standard evaluation metrics like BLEU (Papineni et al. 2002) are sensitive to misgendering. To quantify this, we use the sacreBLEU toolkit (Post 2018) and compute BLEU scores between the ground truth reference outputs and those same references where all SHE pronouns were replaced with morphologically equivalent HE forms (none of these data sets contain neopronouns and analysis of a small sample did not find any singular specific uses of THEY). From wmt08–wmt18 and iwslt17 test sets, the average percentage drop in BLEU score from this error is 0.67% ($\pm$0.22), which is barely statistically significant according to a bootstrap test at sensitivity 0.05. Evaluated only on the $\approx$ 17% of the sentences in these data sets containing either HE or SHE, the degradation is about 3.1%. While this degradation is noticeable, it perhaps does not reflect the real cost of such translation errors due to the high, and asymmetric, societal cost of misgendering.

4.7 Bias in: Feedback Loops

The final sources of bias we consider are feedback loops—essentially, when the bias from a coreference system feeds back on itself, or onto other coreference systems.

The most straightforward way in which this can happen is through coreference resolution systems that engage in statistically biased active learning (or bootstrapping) techniques.[29] Active learning for coreference has been popular since the early 2000s, perhaps largely because coreference annotation is quite costly. Considering the approaches

29 Here, the overloading of "bias" is unfortunate. By "statistically biased," we mean bias in the technical sense that of a learning system that even in the limit of infinite data does not infer the optimal model, a fundamentally different concept from the sort of "bias" we consider in the rest of this paper.

used in dominant papers, the active learning algorithms used are not statistically unbiased (Ng and Cardie 2003; Laws, Heimerl, and Schütze 2012; Miller, Dligach, and Savova 2012; Sachan, Hovy, and Xing 2015; Guha et al. 2015).

Another example is the use of external dictionaries that encode world knowledge that is potentially useful to coreference resolution systems. The earliest example we know of that uses such knowledge sources is the end-to-end machine learning approach of Daumé and Marcu (2005), which found substantial benefit by using mined mappings between names and professions to help resolve named entities like "Bill Clinton" to nominals like "president" (later examples include that of Rahman and Ng [2011] and Bansal and Klein [2012], who found less benefit from a similar approach).

More frequent is the almost ubiquitous use of "name lists" that map names (either full names or simply given names) to "gender." And the most frequently used of these is the resource developed by Bergsma and Lin (2006) (henceforth, B+L), in which a large quantity of text was processed with "high precision" anaphora resolution links to associate names with "genders." The process specifically mapped names to *pronouns*, from which gender (presumably an approximation of referential gender) was inferred. This leads to a resource that pairs a full name or name substring (like "Bill Clinton") counts for identified coreference with HE (8,150, 97.7%), SHE (70, 0.8%), IT (42, 0.5%), and THEY (82, 1%); these are referred to, respectively, as "male," "female," "neuter," and "plural," and seemingly largely used as such in work that leverages this resource. We focus on this resource only because it has become ubiquitous, both in coreference resolution and in gender analysis in NLP more broadly.

The first question we ask is: What happens when this "gender" inference data is used to infer the gender of prominent non-binary individuals? To this end, we took the names of 104 non-binary people referenced on Wikipedia[30] and queried the B+L data with them. In almost all cases, the full name was unknown in the B+L data (or had counts less than five), and in such cases we backed off to simply querying on the given name. We cross-tabulated the correct (according to Wikipedia) pronouns for these 90 people with the "gender" inferred by the B+L data.

The results are shown in Table 5, where we can see that of those who use a pronoun other than SHE or HE (exclusively) are, essentially always, misgendered. Even on binary pronouns SHE and HE, the accuracy is only 50%. For the case of people who use THEY pronouns, one might ask what the ideal behavior would be given the framing of this resource—given that "Plural" is interpreted to be "coreferent with 'they'," we might hope that (aside from the naming issue), people who use THEY are considered Plural: This only happens in 5% of cases, though. Expanding on this, one might hope that people who use THEY $\vee$SHE or THEY $\wedge$SHE are mapped to either Fem or Plural, but this only happens in 2 of 6 cases (and always to Fem in those cases). For the neopronoun cases, the manner in which the resource is constructed nearly excludes any reasonable behavior (aside from, perhaps, a "least bad" option of simply abstaining with an output—which only happens in 3 of 19 cases). This approach actively misgenders individuals, is harmful, and demonstrates that assigning gender to "names" does not work: anybody can have any combination of names and pronouns.

30 https://en.wikipedia.org/wiki/List_of_people_with_non-binary_gender_identities, accessed Jan 5, 2019.

Table 5
For non-binary individuals in our Wikipedia sample (for whom Wikipedia attests current pronouns), a confusion matrix between the pronoun(s) they use (rows) and the inferred gender of their name (based on Bergsma and Lin [2006]) in columns (where Masc="he," Fem="she," Neut="it," and Plur="they"; "Unk" means the name was not found). The final column is the total count. The semantics of "they∨she" is that the person accepts both "they" and "she" pronouns, while "they∧she" indicates that the person uses "they" or "she," depending on context (for instance, "she" while performing drag and "they" otherwise).

	"Gender" % Likelihood					
Pronoun	**Masc**	**Fem**	**Neut**	**Plur**	**Unk**	**count**
she	27	41	18	0	14	22
he	64	7	7	0	22	14
they	39	32	12	5	12	41
ze	22	33	33	11	0	9
they∨she	25	50	25	0	0	4
s/he	0	0	50	0	50	2
they∧she	50	0	0	0	50	2
all/any	50	50	0	0	0	2
xe	50	0	0	0	50	2
ey	50	0	0	0	50	2
v	100	0	0	0	0	1
ae	0	100	0	0	0	1
ne	100	0	0	0	0	1
ze∧she	0	0	100	0	0	1

5. Discussion and Moving Forward

Our goal in this paper was to take a singular task—coreference resolution—and identify how different sources of bias enter into machine learning-based systems for that task. We found varying amounts of bias entering in task definitions (including, in particular, strong assumptions around binary and immutable gender), data collection and annotation (in particular how sources of data impact the sorts of linguistic gender phenomena observed), testing, and feedback. In order to do so, we made substantial use of sociological and sociolinguistic notions of gender, in order to separate out different types of bias.

To run many of these studies, we additionally created—and released—two data sets for studying gender inclusion in coreference resolution. The MAP data set we created counterfactually (and therefore it is subject to general concerns about counterfactual data construction), which allowed us to very precisely control different types of gender information. The GICoref data set we created by targetting specific linguistic phenomena (searching for uses of neopronouns in LGBTQ periodicals) or social aspects (Wikipedia articles and fan fiction about people with non-binary gender). Both data sets show significant gaps in system performance, but perhaps more so, show that taking crowdworker judgments as "gold standard" can be problematic, especially when the annotators are judging referents of singular THEY or neopronouns. It may be the case that to truly build gender inclusive data sets and systems, we need to hire or consult experiential experts (Patton et al. 2019; Young, Magassa, and Friedman 2019).

Moreover, we realized that both human and coreference systems rely heavily on gender cues in resolving coreferences. Though it is natural for humans, we want to emphasize that both humans and systems should not overrely on the risky cues such as

names, semantically gendered nouns, and terms of address, compared to relatively safe cues like syntax. In annotating the data set, we only had about three ambiguous coreferences where both annotators agreed either reference was possible, thus demonstrating that people are able to resolve coreferences without relying extensively on the riskier cues. One cue that we explored in detail is that around names, and it is worth pointing out recent work by Agarwal et al. (2020) in the context of named entity recognition. In that paper, the authors found that state-of-the-art systems perform poorly on documents from non-U.S. contexts, due in large part to systems' unfamiliarity with non-Western names. We expect similar results would hold in the coreference case, where it would be particularly interesting to evaluate in the context of name-gender lists.

When building a coreference system, a developer must make decisions about what features to include or exclude, and therefore what grammatical or social notions of gender are incorporated. Our view is not that "risky" features must be excluded in order to build an inclusive system, but rather that developers should be aware of the risks when such features are included. After all, in a speaker-listener model of language understanding (Bard and Aylett 2004; Frank and Goodman 2012), it is rational for a human speaker to assume that *outside of additional context*, a listener will resolve "his" to "Tom" in "Tom and Mary went to his house." However, human speakers know how to adjust the context when default expectations cannot be used, as in Examples (10), (11), and (12) in §4.5.3. Recall that there and Example (13) here, we found that even given very explicit cues, systems are unable to override their internal biases. If the goal is to understand human communication, having a system that can understand speaker intent is highly important.

(13) **HF & SfdD:** Tom$_{A}$ and Mary$_{B}$$_{C}$ are at home. Tom$_{A}$ regards herself$_{B}$ in the mirror.

This analysis potentially changes if such a model is "flipped" and used, for example, as a method for performing referring expression generation (Krahmer and van Deemter 2012). Depending on a developer's normative stance, ze may need to make a decision about whether hir system will conform to, or challenge, hegemonic language usage, particularly around gender binaries, even though that may produce text that reads as unusual to some (or many) readers. For example, along masculine-as-default lines, does a system generate "engineer" or "man engineer" (when the referent is known to be male), and along non-trans-as-default lines, does a system generate "he/him" or "she/her" in the previous sentence, or "ze/hir"? What a system "should" do in such cases is highly contextual, and perhaps varies even depending on the population expected to use the system. What does not change is that these questions should be addressed head on, so that explicit decisions can be made and consequences understood, rather than being surprised later.

More broadly than in coreference resolution, we found that natural language processing papers also tend to make strong, binary assumptions around *gender* (typically implicitly), a practice that we hope to see change in the future. In more recent papers, we begin to see footnotes that acknowledge that the discussion omits questions around trans or non-binary, issues. We hope to see these be promoted from footnotes to objects of study in future work; mentioning the existence of non-binary people in a footnote does little to minimize the harms a system may cause them. Much inspiration here may come from third wave feminism and queer theory (De Lauretis 1990; Jagose 1996), and perhaps more closely the recent movement within human–computer interaction (HCI) toward Queering HCI (Light 2011) and Feminist HCI (Bardzell and Churchill 2011). The goal that queer theory has of deconstructing social norms and associated taxonomies is

particularly important as NLP technology addresses more and more socially relevant issues, including but not limited to issues around gender, sex, and sexuality.

We hope that this paper can also serve as a roadmap for future studies, both of gender in NLP and of bias in NLP systems. In particular, the gender taxonomy we presented, although not novel, is (to our knowledge) previously unattested in discussions around gender bias in NLP systems; we hope future work in this area can draw on these ideas. It can also be applied to other language settings though grammatical gender can be more complex in some languages. In addition, the specific ways we look into each stage of the machine learning lifecycle can be adapted to similar studies in other language settings too. Finally, we hope that developers of data sets, or systems, in the future, can use some of our analysis as inspiration for how one can attempt to measure—and then root out—different forms of bias throughout the development lifecycle.

A. Annotation of ACL Anthology Papers

Below we list the complete set of annotations we did of the papers described in §4.5.1. For each of the papers considered, we annotate the following items:

- Coref: Does the paper discuss coreference resolution?
- L.G: Does the paper deal with linguistic gender (grammatical gender or gendered pronouns)?
- S.G: Does the paper deal with social gender?
- Eng: Does the paper study English?
- L≠G: (If yes to L.G and S.G:) Does the paper distinguish linguistic from social gender?
- 0/1: (If yes to S.G:) Does the paper explicitly or implicitly assume that social gender is binary?
- Imm: (If yes to S.G:) Does the paper explicitly or implicitly assume social gender is immutable?
- Neo: (If yes to S.G and to English:) Does the paper explicitly consider uses of definite singular "they" or neopronouns?

For each of these, we mark with [Y] if the answer is yes, [N] if the answer is no, and [−] if this question is not applicable (i.e., it doesn't pass the conditional checks).

Citation	Coref	L.G	S.G	Eng	L≠S	0/1	Imm	Neo
Sidner (1981)	Y	Y	Y	Y	N	–	–	–
Bainbridge (1985)	Y	Y	N	Y	–	–	–	–
Kameyama (1986)	Y	Y	Y	Y	N	Y	Y	N
Mellish (1988)	N	Y	N	Y	–	–	–	–
Danlos and Namer (1988)	N	Y	N	N	–	–	–	–
Yoshimoto (1988)	N	Y	N	N	–	–	–	–
Zock, Francopoulo, and Laroui (1988)	N	Y	N	N	–	–	–	–
Popowich (1989)	N	Y	N	Y	–	–	–	–

Citation	Coref	L.G	S.G	Eng	L≠S	0/1	Imm	Neo
Mani et al. (1993)	Y	N	Y	Y	–	Y	–	–
Narayanan and Hashem (1993)	N	Y	N	N	–	–	–	–
Soloman and Wood (1994)	N	Y	N	Y	–	–	–	–
Quantz (1994)	N	Y	N	Y	–	–	–	–
Baker, Gillick, and Roth (1994)	–	–	–	–	–	–	–	–
Genthial, Courtin, and Menezo (1994)	N	Y	N	N	–	–	–	–
Levinger, Ornan, and Itai (1995)	N	Y	N	N	–	–	–	–
Holan, Kuboň, and Plátek (1997)	N	Y	N	N	–	–	–	–
Dorna et al. (1998)	N	N	N	Y	–	–	–	–
Harabagiu and Maiorano (1999)	Y	Y	Y	Y	N	Y	Y	N
Avgustinova and Uszkoreit (2000)	N	Y	N	N	–	–	–	–
Channarukul, McRoy, and Ali (2000)	N	Y	N	Y	–	–	–	–
Abuleil, Alsamara, and Evens (2002)	N	Y	N	N	–	–	–	–
Cucerzan and Yarowsky (2003)	N	Y	N	N	–	–	–	–
Pakhomov, Buntrock, and Chute (2003)	N	N	Y	Y	–	–	–	–
Tadić and Fulgosi (2003)	N	Y	N	N	–	–	–	–
Debowski (2003)	N	Y	N	N	–	–	–	–
Navarretta (2004)	Y	Y	Y	N	N	Y	Y	–
Carl et al. (2004)	Y	Y	Y	N	N	Y	Y	–
Mota, Carvalho, and Ranchhod (2004)	N	Y	N	Y	–	–	–	–
Eisner and Karakos (2005)	N	Y	N	Y	–	–	–	–
Boulis and Ostendorf (2005)	N	N	Y	Y	–	Y	Y	N
Smith, Smith, and Tromble (2005)	N	Y	N	N	–	–	–	–
Bergsma and Lin (2006)	Y	Y	Y	Y	N	Y	Y	N
Vogt and André (2006)	N	N	Y	N	–	Y	Y	–
Quirk and Corston-Oliver (2006)	N	Y	N	Y	–	–	–	–
Dada (2007)	N	Y	N	N	–	–	–	–
Streiter, Voltmer, and Goudin (2007)	N	N	Y	N	–	–	–	–
Jing, Kambhatla, and Roukos (2007)	Y	Y	Y	Y	N	Y	–	N
Badr, Zbib, and Glass (2008)	N	Y	N	N	–	–	–	–
Marchal et al. (2008)	N	Y	N	N	–	–	–	–
van Peursen (2009)	N	Y	N	N	–	–	–	–
Badr, Zbib, and Glass (2009)	N	Y	N	N	–	–	–	–
Garera and Yarowsky (2009)	N	Y	Y	Y	N	Y	Y	N
Bergsma, Lin, and Goebel (2009)	Y	Y	Y	Y	N	Y	Y	N
Nastase and Popescu (2009)	N	Y	N	N	–	–	–	–
Nanba et al. (2009)	N	N	N	Y	–	–	–	–
Robaldo and Di Carlo (2009)	N	N	N	Y	–	–	–	–
Mukherjee and Liu (2010)	N	N	Y	Y	–	Y	Y	–
Ng (2010)	Y	Y	Y	Y	N	Y	Y	N
Burkhardt et al. (2010)	N	N	Y	N	–	Y	Y	–
Marton, Habash, and Rambow (2010)	N	Y	N	N	–	–	–	–
Le Nagard and Koehn (2010)	Y	Y	Y	Y	N	Y	Y	N
Rojas-Barahona et al. (2011)	N	Y	N	N	–	–	–	–
Mukund, Ghosh, and Srihari (2011)	N	Y	N	N	–	–	–	–
Sarawgi, Gajulapalli, and Choi (2011)	N	N	Y	Y	–	Y	Y	N
Li, Miller, and Schuler (2011)	Y	Y	Y	Y	N	Y	Y	N

Citation	Coref	L.G	S.G	Eng	L≠S	0/1	Imm	Neo
Burger et al. (2011)	N	N	Y	Y	–	Y	Y	N
Mohammad and Yang (2011)	N	N	Y	Y	–	Y	Y	N
Sapena, Padró, and Turmo (2011)	Y	Y	Y	Y	N	Y	Y	N
Charton and Gagnon (2011)	Y	Y	Y	Y	N	Y	Y	N
Alkuhlani and Habash (2011)	N	Y	N	N	–	–	–	–
Mareček et al. (2011)	N	Y	N	N	–	–	–	–
López-Ludeña et al. (2011)	N	Y	N	N	–	–	–	–
Declerck, Koleva, and Krieger (2012)	Y	Y	N	Y	–	–	–	–
Bergsma, Post, and Yarowsky (2012)	N	N	Y	Y	–	Y	Y	N
Alkuhlani and Habash (2012)	N	Y	N	N	–	–	–	–
Filippova (2012)	N	N	Y	Y	–	Y	–	–
Dinu, Niculae, and, Şulea (2012)	N	Y	N	N	–	–	–	–
El Kholy and Habash (2012)	N	Y	N	N	–	–	–	–
Yu (2012)	N	N	N	N	–	–	–	–
Guillou (2012)	Y	Y	Y	Y	Y	Y	–	–
Vogel and Jurafsky (2012)	N	N	Y	Y	–	Y	Y	N
Goldberg and Elhadad (2013)	N	Y	N	N	–	–	–	–
Marton, Habash, and Rambow (2013)	N	Y	N	N	–	–	–	–
Weller, Fraser, and Schulte im Walde (2013)	N	Y	N	Y	–	–	–	–
Ciot, Sonderegger, and Ruths (2013)	N	N	Y	N	–	Y	Y	–
Volkova, Wilson, and Yarowsky (2013)	N	N	Y	Y	–	Y	Y	N
Levitan (2013)	N	N	Y	Y	–	N	N	N
Bojar, Rosa, and Tamchyna (2013)	N	Y	N	N	–	–	–	–
Glavaš, Korenčić, and Šnajder (2013)	N	Y	N	N	–	–	–	–
Liu et al. (2013)	N	N	N	N	–	–	–	–
Kestemont (2014)	N	N	N	Y	–	–	–	–
Novák and Žabokrtský (2014)	Y	Y	N	Y	–	–	–	–
Babych et al. (2014)	N	Y	N	N	–	–	–	–
Soler-Company and Wanner (2014)	N	N	Y	Y	–	Y	Y	N
Chen and Ng (2014)	Y	Y	Y	Y	N	Y	Y	N
Sap et al. (2014)	N	N	Y	Y	–	Y	Y	–
Nguyen et al. (2014)	N	N	Y	Y	–	Y	Y	N
Prabhakaran, Reid, and Rambow (2014)	N	N	Y	Y	–	Y	Y	N
Sidorov, Ultes, and Schmitt (2014)	N	N	Y	Y	–	Y	Y	N
Darwish, Abdelali, and Mubarak (2014)	N	Y	N	N	–	–	–	–
Ahmed Khan (2014)	N	Y	N	N	–	–	–	–
Nguyen, Trieschnigg, and Meder (2014)	N	N	Y	N	–	Y	Y	–
Stewart (2014)	N	N	Y	Y	–	Y	Y	–
Matthews et al. (2014)	N	Y	N	N	–	–	–	–
Vaidya, Rambow, and Palmer (2014)	N	Y	N	N	–	–	–	–
Kokkinakis, Ighe, and Malm (2015)	N	Y	Y	N	N	Y	–	–
Johannsen, Hovy, and Søgaard (2015)	N	N	Y	Y	–	Y	Y	–
Schwartz et al. (2015)	N	N	N	Y	–	–	–	–
Hovy (2015)	N	N	Y	Y	–	Y	Y	N
Agarwal et al. (2015)	N	Y	Y	Y	N	Y	Y	N
Preoţiuc-Pietro et al. (2015)	N	N	Y	Y	N	Y	Y	–
Ramakrishna et al. (2015)	N	Y	Y	Y	N	Y	Y	N

Citation	Coref	L.G	S.G	Eng	L≠S	0/1	Imm	Neo
Taniguchi et al. (2015)	N	N	Y	Y	–	N	Y	N
Schofield and Mehr (2016)	N	N	Y	Y	–	Y	Y	N
Levitan et al. (2016)	N	N	Y	Y	–	Y	Y	N
Flekova et al. (2016)	N	N	Y	Y	–	Y	Y	N
Tran and Ostendorf (2016)	N	N	N	Y	–	–	–	–
Qian, Qiu, and Huang (2016)	N	Y	N	Y	–	–	–	–
Li et al. (2016)	N	N	Y	Y	–	Y	Y	N
Zhang et al. (2016)	N	N	Y	Y	–	Y	Y	N
Garimella and Mihalcea (2016)	N	N	Y	Y	–	Y	Y	N
Reddy and Knight (2016)	N	N	Y	Y	–	Y	Y	N
Li and Dickinson (2017)	N	N	Y	N	–	Y	Y	–
Pérez Estruch, Paredes Palacios, and Rosso (2017)	N	N	Y	Y	–	Y	Y	N
Pérez-Rosas et al. (2017)	N	N	Y	Y	–	Y	Y	N
Rabinovich et al. (2017)	N	N	Y	N	–	Y	Y	–
Costa-jussà (2017)	N	Y	N	N	–	–	–	–
Sap et al. (2017)	N	N	Y	Y	–	Y	–	–
Zhao et al. (2017)	N	N	Y	Y	–	Y	Y	N
Mandravickaitė and Krilavičius (2017)	N	N	Y	Y	–	Y	Y	N
Verhoeven, Škrjanec, and Pollak (2017)	N	N	Y	Y	–	Y	Y	N
Larson (2017)	N	Y	Y	Y	Y	N	N	Y
Koolen and van Cranenburgh (2017)	N	N	Y	N	–	N	Y	–
Tatman (2017)	N	N	Y	Y	–	Y	Y	N
Soler-Company and Wanner (2017)	N	N	Y	Y	–	Y	Y	N
Ljubešić, Fišer, and Erjavec (2017)	N	N	Y	N	–	Y	Y	–
Litvinova et al. (2017)	N	N	Y	N	–	Y	Y	–
Mohammad et al. (2018)	N	N	Y	Y	–	Y	–	–
Wang and Jurgens (2018)	N	Y	Y	Y	Y	N	N	N
Kraus et al. (2018)	N	N	Y	Y	–	Y	–	–
Martinc and Pollak (2018)	N	N	Y	Y	–	Y	Y	N
Chan and Fyshe (2018)	N	N	Y	Y	–	Y	Y	N
Durmus and Cardie (2018)	N	N	N	Y	–	–	–	–
Zaghouani and Charfi (2018)	N	Y	Y	N	N	Y	Y	–
Plank (2018)	N	N	Y	Y	–	Y	Y	N
Wood-Doughty et al. (2018)	N	N	Y	Y	–	Y	Y	N
Moorthy et al. (2018)	N	N	Y	Y	–	Y	–	–
Levitan, Maredia, and Hirschberg (2018)	N	N	Y	Y	–	Y	Y	N
Webster et al. (2018)	Y	Y	Y	Y	N	Y	Y	N
Park, Shin, and Fung (2018)	N	Y	Y	Y	N	Y	Y	N
Vanmassenhove, Hardmeier, and Way (2018)	N	Y	Y	N	N	Y	Y	–
Kleinberg, Mozes, and van der Vegt (2018)	N	N	Y	Y	–	Y	Y	N
Zhao et al. (2018b)	N	N	Y	Y	–	Y	Y	N
Balusu, Merghani, and Eisenstein (2018)	N	N	N	Y	–	–	–	–
Rudinger et al. (2018)	Y	Y	Y	Y	N	N	–	Y
Zhao et al. (2018a)	Y	Y	Y	Y	N	Y	Y	N
Kiritchenko and Mohammad (2018)	–	–	–	–	–	–	–	–
Barbieri and Camacho-Collados (2018)	N	N	Y	Y	–	Y	N	–
van der Goot et al. (2018)	N	N	Y	N	–	Y	Y	–
Karlekar, Niu, and Bansal (2018)	N	N	Y	Y	–	Y	Y	N
de Gibert et al. (2018)	N	N	N	Y	–	–	–	–
Mickus, Bonami, and Paperno (2019)	N	Y	N	N	–	–	–	–

B. Example GICoref Document from Wikipedia: Dana Zzyym

[[Source: `https://en.wikipedia.org/wiki/Dana_Zzyym`**]]**

Dana Alix Zzyym$_A$ is an Intersex activist and former sailor who was the first military veteran in the United States to seek a non - binary gender U.S. passport, in a lawsuit Zzyym$_A$ v. Pompeo$_C$.

Early life

Zzyym$_A$ has expressed that their$_A$ childhood as a military brat made it out of the question for them$_A$ to be associated with the queer community as a youth due to the prevalence of homophobia in the armed forces . Their$_A$ parents$_B$ hid Zzyym$_A$'s status as intersex from them$_A$ and Zzyym$_A$ discovered their$_A$ identity and the surgeries their$_A$ parents$_B$ had approved for them$_A$ by themselves$_B$ after their$_A$ Navy service . In 1978, Zzyym$_A$ joined the Navy as a machinist 's mate .

Activism

Zzyym$_A$ has been an avid supporter of the Intersex Campaign for Equality .

Legal case

Zzyym$_A$ is the first veteran to seek a non - binary gender U.S. passport . In light of the State Department 's continuing refusal to recognize an appropriate gender marker, on June 27, 2017 a federal court granted Lambda Legal 's motion to reopen the case . On September 19, 2018, the United States District Court for the District of Colorado enjoined the U.S. Department of State from relying upon its binary - only gender marker policy to withhold the requested passport .

C. Example GICoref Document from AO3: Scar Tissue

[[Source: `https://archiveofourown.org/works/14476524`**]]**
[[Author: cornheck]]

Despite dreading their$_A$ first true series of final exams, Crona$_A$'s relieved to have a particularly absorbative memory, lucky to recall all the material they$_A$ 'd been required to catch up on . Half a semester of attendance, a whole year of course content .

The only true moment of discomfort came when they$_A$ 'd arrived at the essay portion . Thankful it was easy enough to answer, however, their$_A$ subtle eye - roll stemmed entirely from just how much writing it asked of them$_A$, hands already beginning to ache at the thought of scrawling out two pages on the origins, history, and importance of partnered and grouped soul resonance .

By the end of it all, their$_A$ neck, wrist, back, and ribs ached from the strain of their$_A$ typical, hunched posture – a habit they$_A$ defaulted to, and Miss Marie$_B$ silently wished they$_A$ 'd be more mindful of . It was a relief, at least to them$_A$, not to be the last one out of the lecture hall . Booklet turned in, they$_A$ left the room as quietly as possible and lingered just outside, an air of hesitance settling upon them$_A$ as they$_A$ considered what to do now that, it seemed, everything was over with . No more class, no more lessons, just . . . students on break from their studies for the season .

" Kind of a breeze, was n't it ? " Evans$_C$ ' voice echoes in the arched hall and Crona$_A$'s shoulders jump, their$_A$ frame still a tense and anxious mess .

" Oh, " they$_A$ sigh, " I$_A$ . . . I$_A$ suppose so . It was n't . . . necessarily hard . " Crona$_A$ answers, putting forth a vaguely forced smile .

Smiling with the assumed purpose of making Soul$_C$ comfortable with the interaction . A defense mechanism .

" I$_A$ - I$_A$ guess, for a final, it was easier than I$_A$ expected . . . everyone . . . made it sound like it 'd be difficult . "

" If by everyone, you$_A$ mean Black Star$_D$, then yeah, " Soul$_C$ chuckles, " he$_D$ does n't really do well on ' em . . . bad test - taker . "

" Ah, " their$_A$ facade falls just in time to be replaced by a much more genuine grin .

Of the little they$_A$ 'd spent talking to Black Star$_D$, he$_D$ certainly had confidence and skill enough to make up for the lost exam points given his$_D$ performance in every other grading category .

" That . . . makes sense . "

" Maka$_E$'s always the first one done when it comes to this stuff, she$_E$ practically studies in her$_E$ sleep . I$_C$ 'm convinced she$_E$ must be practicing clairvoyance the way she$_E$ burns through essay questions, " Soul$_C$ laughs, turning to the meek teen$_A$ who gives him$_C$ a simple nod in response .

Determined not to let an impending awkward silence fall between them$_F$, Soul$_C$ pipes up again, " So, are you$_A$ staying here for break ? "

" Ye - well, I$_A$ . . . I$_A$ think so, " they$_A$ begin, stuttering, but encouraged to continue by a cock of Soul$_C$'s head; a social cue even they$_A$ could read, " The professor$_H$ . . . and Miss Marie$_{B\,G}$ asked if I$_A$ 'd like to come and stay with them$_G$ for the time being . "

" Oh, huh, Stein$_H$ and Marie$_{B\,G}$? Nice, " his$_C$ brows lift, clearly some varying degree of happy for the other$_A$.

The optimism is short - lived, observing as Crona$_A$'s expression falls back to its characteristic expressionless gaze .

" It seems like you$_A$ 've got a good thing going with those two$_G$. "

" I$_A$ have n't decided, yet, if I$_A$ should accept the invitation, " they$_A$ shift a bit where they$_A$ stand .

Never having been the best at reassuring others, even his$_C$ own meister$_A$, Soul$_C$ kept his$_C$ mouth shut to avoid stuttering while he$_C$ searched for the right words a web of thoughts .

" Y '$_A$ know, I$_C$ think it 's less of an invitation and more of an extended welcome . "

The other$_A$ raises their$_A$ head, taken aback, " Oh, " Crona$_A$ mutters, in a poignant tone, " I$_A$ . . . never considered something like that . "

Soul$_C$ does n't leave much wiggle room for their$_A$ mood to fall any further (nothing past a flat - lipped frown) , " They$_G$ 'd probably love to have you$_A$, I$_C$ bet they$_G$ drive each other nuts sometimes all by themselves$_G$. "

Though Evans$_C$ wo n't admit it, he$_C$ knows it 's all too likely Stein$_H$ might actually put some more effort into taking care of himself$_H$ if he$_H$ had someone else besides Marie$_B$ to look after .

" I$_A$ - I$_A$ see, " they$_A$ exhale with a nod, giving Soul$_C$ a hint of affirmation that he$_C$ 'd done something to boost the kid$_A$'s confidence .

" I$_C$ mean, it 's got ta be lonely not to mention boring hanging here all summer . . . and the weather, " Soul$_C$ nearly gasps, dramatizing it for added effect, " Oh, man, I$_C$ do n't know how you$_A$ can stay cooped up in that room of yours$_A$ when it 's so nice out, " he$_C$ grins .

" But . . . meh . Different strokes . I$_C$ ca n't judge . "

His$_C$ comments comfort them$_A$, an for a moment they$_A$ forget how this came to be . The cathedral in Italy, Lady Medusa 's wrath, and the black blood that infected him$_C$. Every moment they$_A$ spent in the presence of Soul Evans$_C$ builds always up to this; fixation on the memories of their$_U$ first encounters and all the pain they$_A$ 've caused him$_C$, the pain they$_A$ 've caused he$_C$ and Maka$_{E,K}$ both . As quickly as Soul$_C$ had lifted the swordsman$_A$'s spirits, they$_A$ 'd weighed themselves$_A$ down once more . It seemed so normal, though . Soul$_C$ could n't bring himself$_C$ to feel any sense of accomplishment in the coaxing - out of Crona$_A$'s smile when the return of their$_A$ self doubt was as certain as the sun in the sky . His$_C$ own stubbornness could n't let his$_C$ diminished self worth lie . With another encouraging smile, rows of sharpened incisors appearing oddly charismatic, he$_C$ opens his$_C$ mouth to speak – but finds himself$_C$ cut off before he$_C$ can even squeeze a word in .

" Soul$_C$, I$_A$ 'm sorry, " the meister$_A$ blurts .

Having been pent - up for months, the apology comes forth without inhibition, rolling effortlessly off their$_A$ tongue . " Sorry . . . ? For what ? " Evans$_C$ quirks a brow, chuckling .

He$_C$ adjusts his$_C$ stance to face Crona$_A$ with the whole of his$_C$ body, maintaining his$_C$ positive demeanor .

" F - for what . . . ? "

They$_A$ stammer, shaking their$_A$ head . For all their$_A$ remorse, they$_A$ thought this would have been obvious .

" For everything, it 's . . . the first time we$_F$ dueled, I$_A$ was the enemy ! I$_A$ - I$_A$ almost killed you$_C$, I$_A$ - I$_A$ ... I$_A$ really, really hurt you$_C$, " they$_A$ answer, still so sick with guild that even their$_A$ confession of responsibility is tainted with frustration .

Soul$_C$ seems stunned for a moment before harnessing his$_C$ quick wit .

" Hey, now, you$_A$ ca n't take all the credit like that, Ragnarok$_L$ did most of the damage, " he$_C$ . . .

Acknowledgments

The authors are grateful to a number of people who have provided pointers, edits, suggestions, and annotation facilities to improve this work: Lauren Ackerman, Cassidy Henry, Os Keyes, Chandler May, Hanyu Wang, Marion Zepf, and reviewers all contributed to various aspects of this work, including suggestions for data sources for the GICoref data set. We also thank the CLIP lab at the University of Maryland for comments on previous drafts.

References

Abuleil, Saleem, Khalid Alsamara, and Martha Evens. 2002. Acquisition system for Arabic noun morphology. In *Proceedings of the ACL-02 Workshop on Computational Approaches to Semitic Languages*, Philadelphia, PA.

Ackerman, Lauren. 2019. Syntactic and cognitive issues in investigating gendered coreference. *Glossa*, 4.

Agarwal, Apoorv, Jiehan Zheng, Shruti Kamath, Sriramkumar Balasubramanian, and Shirin Ann Dey. 2015. Key female characters in film have more to talk about besides men: Automating the Bechdel test. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 830–840, Denver, CO.

Agarwal, Oshin, Sanjay Subramanian, Ani Nenkova, and Dan Roth. 2019. Evaluation of named entity coreference. In *Proceedings of the Second Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 1–7, Minneapolis, USA.

Agarwal, Oshin, Yinfei Yang, Byron C. Wallace, and A. Nenkova. 2020. Entity-switched datasets: An approach to auditing the in-domain robustness of named entity recognition models. *ArXiv*, abs/2004.04123.

Ahmed Khan, Tafseer. 2014. Automatic acquisition of Urdu nouns (along with gender and irregular plurals). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2846–2850, Reykjavik.

Alkuhlani, Sarah and Nizar Habash. 2011. A corpus for modeling morpho-syntactic

agreement in Arabic: Gender, number and rationality. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 357–362, Portland, OR.

Alkuhlani, Sarah and Nizar Habash. 2012. Identifying broken plurals, irregular gender, and rationality in Arabic text. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 675–685, Avignon.

Andrews, Travis M. 2017. The singular, gender-neutral 'they' added to the Associated Press Stylebook. Washington Post; https://web.archive.org/web/20170328082411/https://www.washingtonpost.com/news/morning-mix/wp/2017/03/28/the-singular-gender-neutral-they-added-to-the-associated-press-stylebook/?utm_term=.a90680e0725c.

Anonymous. 2017. Understanding drag. Blog post, https://transequality.org/issues/resources/understanding-drag.

Arnold, Jennifer E. 2008. Reference production: Production-internal and addressee-oriented processes. *Language and Cognitive Processes*, 23(4):495–527. https://doi.org/10.1080/01690960801920099

Avgustinova, Tania and Hans Uszkoreit. 2000. An ontology of systematic relations for a shared grammar of Slavic. In *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*, pages 28–34.

Babych, Bogdan, Jonathan Geiger, Mireia Ginestí Rosell, and Kurt Eberle. 2014. Deriving de/het gender classification for Dutch nouns for rule-based MT generation tasks. In *Proceedings of the 3rd Workshop on Hybrid Approaches to Machine Translation (HyTra)*, pages 75–81, Gothenburg.

Badr, Ibrahim, Rabih Zbib, and James Glass. 2008. Segmentation for English-to-Arabic statistical machine translation. In *Proceedings of ACL-08: HLT, Short Papers*, pages 153–156, Columbus, OH.

Badr, Ibrahim, Rabih Zbib, and James Glass. 2009. Syntactic phrase reordering for English-to-Arabic statistical machine translation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 86–93, Athens.

Bagga, Amit and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, volume 1, pages 563–566, Granada.

Bainbridge, R. I. 1985. Montagovian definite clause grammar. In *Second Conference of the European Chapter of the Association for Computational Linguistics*, pages 25–34, Geneva.

Baker, Janet, Larry Gillick, and Robert Roth. 1993. Research in large vocabulary continuous speech recognition. In *Human Language Technology: Proceedings of a Workshop*, page 394, Plainsboro, NJ.

Balusu, Murali Raghu Babu, Taha Merghani, and Jacob Eisenstein. 2018. Stylistic variation in social media part-of-speech tagging. In *Proceedings of the Second Workshop on Stylistic Variation*, pages 11–19, New Orleans, LA.

Bansal, Mohit and Dan Klein. 2012. Coreference semantics from web features. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 389–398.

Barbieri, Francesco and Jose Camacho-Collados. 2018. How gender and skin tone modifiers affect emoji semantics in twitter. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 101–106, New Orleans, LA.

Bard, Ellen and Matthew Aylett. 2004. Referential form word duration and modeling the listener in spoken dialogue.

Bardzell, Shaowen and Elizabeth F. Churchill. 2011. IwC special issue "feminism and HCI: New perspectives" special issue editors' introduction. *Interacting with Computers*, 23(5).

Barocas, Solon, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: Allocative versus representational harms in machine learning. In *Proceedings of SIGCIS*. Philadelphia, PA, USA.

Baron, Naomi S. 1971. A reanalysis of English grammatical gender. *Lingua*, 27:113–140. https://doi.org/10.1016/0024-3841(71)90082-9

Bender, Emily M. 2019. A typology of ethical risks in language technology with an eye towards where transparent documentation can help. *The Future of Artificial Intelligence: Language, Ethics, Technology*.

Bender, Emily M. and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science, *TACL*. 6:587–604.

Bergsma, Shane and Dekang Lin. 2006. Bootstrapping path-based pronoun resolution. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 33–40, Sydney.

Bergsma, Shane, Dekang Lin, and Randy Goebel. 2009. Glen, Glenda or Glendale: Unsupervised and semi-supervised learning of English noun gender. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 120–128, Boulder, CO.

Bergsma, Shane, Matt Post, and David Yarowsky. 2012. Stylometric analysis of scientific articles. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 327–337, Montréal.

Bjorkman, Bronwyn M. 2017. Singular they and the syntactic representation of gender in English. *Glossa*, 2(1):80. https://doi.org/10.5334/gjgl.374

Blodgett, Su Lin, Solon Barocas, Hal Daumé, III, and Hanna Wallach. 2020. Language (technology) is power: The need to be explicit about NLP harms. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Bojar, Ondřej, Rudolf Rosa, and Aleš Tamchyna. 2013. Chimera – three heads for English-to-Czech translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 92–98, Sofia.

Bolukbasi, Tolga, 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 NeurIPS*.

Boulis, Constantinos and Mari Ostendorf. 2005. A quantitative analysis of lexical differences between genders in telephone conversations. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 435–442, Ann Arbor, MI.

Bradley, Evan, Julia Salkind, Ally Moore, and Sofi Teitsort. 2019. Singular 'they' and novel pronouns: gender-neutral, nonbinary, or both? In *Proceedings of the Linguistic Society of America*, 4(1):36–1–7.

Bruhns, Karen Olsen. 2006. Gender archaeology in native North America. In Nelson, Sarah, editor, *Handbook of Gender in Archaeology*.

Bucholtz, Mary. 1999. Gender. *Journal of Linguistic Anthropology*. Special issue: Lexicon for the New Millennium, ed. Alessandro Duranti.

Buolamwini, Joy and Timnit Gebru. 2018. *Gender shades: Intersectional accuracy disparities in commercial gender classification*.

Burger, John D., John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on Twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1301–1309, Edinburgh.

Burkhardt, Felix, Martin Eckert, Wiebke Johannsen, and Joachim Stegmann. 2010. A database of age and gender annotated telephone speech. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, European Language Resources Association (ELRA), Valletta.

Burton, Richard. 1883. *Kama Sutra, Translation*.

Bustillos, Maria. 2011. Our desperate, 250-year-long search for a gender-neutral pronoun. https://web.archive.org/web/20110110061401/https://www.theawl.com/2011/01/our-desperate-250-year-long-search-for-a-gender-neutral-pronoun/.

Butler, Judith. 1989. *Gender Trouble: Feminism and the Subversion of Identity*, Routledge.

Cai, Jie and Michael Strube. 2010. Evaluation metrics for end-to-end coreference resolution systems. In *Proceedings of the SIGDIAL 2010 Conference*, pages 28–36.

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Cao, Yang Trista and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online.

Carl, Michael, Sandrine Garnier, Johann Haller, Anne Altmayer, and Bärbel Miemietz. 2004. Controlling gender equality with shallow NLP techniques. In *COLING 2004: Proceedings of the 20th International Conference on*

Computational Linguistics, pages 820–826, COLING, Geneva.

Carreiras, Manuel, Alan Garnham, Jane Oakhill, and Kate Cain. 1996. The use of stereotypical gender information in constructing a mental model: Evidence from English and Spanish. *The Quarterly Journal of Experimental Psychology Section A*, 49(3):639–663.

Chan, Sophia and Alona Fyshe. 2018. Social and emotional correlates of capitalization on Twitter. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 10–15, New Orleans, LA.

Channarukul, Songsak, Susan W. McRoy, and Syed S. Ali. 2000. Enriching partially-specified representations for text realization using an attribute grammar. In *INLG'2000 Proceedings of the First International Conference on Natural Language Generation*, pages 163–170, Mitzpe Ramon.

Charton, Eric and Michel Gagnon. 2011. Poly-co: A multilayer perceptron approach for coreference detection. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 97–101, Portland, OR.

Chen, Chen and Vincent Ng. 2014. Chinese zero pronoun resolution: An unsupervised probabilistic model rivaling supervised resolvers. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 763–774, Doha.

Ciot, Morgane, Morgan Sonderegger, and Derek Ruths. 2013. Gender inference of Twitter users in non-English contexts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1136–1145, Seattle, WA.

Clark, Kevin, and Christopher D. Manning. 2015. Entity-centric coreference resolution with model stacking. In *ACL*.

Clark, Kevin, and Christopher D. Manning. 2016. Deep reinforcement learning for mention-ranking coreference models. In *EMNLP*.

Clements, K. C. 2017. What is deadnaming? Blog post, `https://www.healthline.com/health/transgender/deadnaming`.

Clements, K. C. 2018. Misgendering: What is it and why is it harmful? Blog post, `https://www.healthline.com/health/transgender/misgendering`.

Conrod, Kirby. 2018a. Changes in singular they. In *Cascadia Workshop in Sociolinguistics*.

Conrod, Kirby. 2018b. What does it mean to agree? Coreference with singular they. In *Pronouns in Competition workshop*.

Corbett, Greville G. 1991. *Gender*. Cambridge University Press.

Corbett, Greville G. 2013. Number of genders. In Dryer, Matthew S. and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*, Max Planck Institute for Evolutionary Anthropology, Leipzig.

Costa-jussà, Marta R. 2017. Why Catalan-Spanish neural machine translation? Analysis, comparison and combination with standard rule and phrase-based technologies. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 55–62, Valencia.

Craig, Colette G. 1994. Classifier languages. *The Encyclopedia of Language and Linguistics* 2:565–569.

Cucerzan, Silviu and David Yarowsky. 2003. Minimally supervised induction of grammatical gender. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 40–47.

Dada, Ali. 2007. Implementation of the Arabic numerals and their syntax in GF. In *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*, pages 9–16, Prague.

Dahl, Osten. 2000. Animacy and the notion of semantic gender. *Trends in Linguistics Studies and Monographs*, 124:99–116. `https://doi.org/10.1515/9783110802603.99`

Danlos, Laurence and Fiametta Namer. 1988. Morphology and cross dependencies in the synthesis of personal pronouns in romance languages. In *Coling 1988 Volume 1: International Conference on Computational Linguistics*, Budapest.

Darwin, Helana. 2017. Doing gender beyond the binary: A virtual ethnography. *Symbolic Interaction*, 40(3):317–334. `https://doi.org/10.1002/symb.316`

Darwish, Kareem, Ahmed Abdelali, and Hamdy Mubarak. 2014. Using stem-templates to improve Arabic POS and gender/number tagging. In *Proceedings of the Ninth International Conference on Language Resources and*

Evaluation (LREC'14), pages 2926–2931, European Language Resources Association (ELRA), Reykjavik.

Daumé, Hal, III and Daniel Marcu. 2005. *A large-scale exploration of effective global features for a joint entity detection and tracking model*, pages 97–104.

de Gibert, Ona, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels.

de Lauretis, Teresa. 1990. Feminism and its differences. *Pacific Coast Philology*. pages 24–30.

Debowski, Łukasz. 2003. A reconfigurable stochastic tagger for languages with complex tag structure. In *Proceedings of the 2003 EACL Workshop on Morphological Processing of Slavic Languages*, pages 63–70, Budapest.

Declerck, Thierry, Nikolina Koleva, and Hans-Ulrich Krieger. 2012. Ontology-based incremental annotation of characters in folktales. In *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 30–34, Avignon.

Dinu, Liviu P., Vlad Niculae, and Octavia-Maria Şulea. 2012. The Romanian neuter examined through a two-gender n-gram classification system. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 907–910, European Language Resources Association (ELRA), Istanbul.

Doleschal, Ursula and Sonja Schmid. 2001. Doing gender in Russian. *Gender Across Languages. The Linguistic Representation of Women and Men*, 1:253–282. https://doi.org/10.1075/impact.9.16dol

Dorna, Michael, Anette Frank, Josef van Genabith, and Martin C. Emele. 1998. Syntactic and semantic transfer with f-structures. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 341–347, Montreal.

Dryer, Matthew S. 2013, Expression of pronominal subjects. In Dryer, Matthew S. and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*, Max Planck Institute for Evolutionary Anthropology, Leipzig.

Du, Elizabeth and Ross Liddy. 1990. Anaphora in natural language processing and information retrieval. *Information Processing & Management*, 26.

Durmus, Esin and Claire Cardie. 2018. Understanding the effect of gender and stance in opinion expression in debates on "abortion." In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 69–75, New Orleans, LA.

Edens, Richard J., Helen L. Gaylard, Gareth J. F. Jones, and Adenike M. Lam-Adesina. 2003. An investigation of broad coverage automatic pronoun resolution for information retrieval. In *SIGIR*.

Eisner, Jason and Damianos Karakos. 2005. Bootstrapping without the boot. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 395–402, Vancouver.

El Kholy, Ahmed and Nizar Habash. 2012. Rich morphology generation using statistical machine translation. In *INLG 2012 Proceedings of the Seventh International Natural Language Generation Conference*, pages 90–94, Utica, IL.

Esaulova, Yulia, Chiara Reali, and Lisa von Stockhausen. 2014. Influences of grammatical and stereotypical gender during reading: eye movements in pronominal and noun phrase anaphor resolution. *Language, Cognition and Neuroscience*, 29(7):781–803.

Filippova, Katja. 2012. User demographics and language in an implicit social network. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1478–1488, Jeju Island.

Flekova, Lucie, Jordan Carpenter, Salvatore Giorgi, Lyle Ungar, and Daniel Preoţiuc-Pietro. 2016. Analyzing biases in human perception of user age and gender from text. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 843–854, Berlin.

Font, Joel Escudé and Marta R. Costa-jussà. 2019. Equalizing gender biases in neural machine translation with word embeddings techniques. In *Proceedings of the 1st ACL Workshop on Gender Bias for Natural Language Processing*.

Frank, Anke, Christiane Hoffmann, Maria Strobel, et al. 2004. Gender issues in machine translation. *Amsterdam (J. Benjamins)*.

Frank, Michael C. and Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084): 998–998. https://doi.org/10.1086/589478

Fraser, Nancy. 2008. Abnormal justice. *Critical Inquiry*, 34(3):393–422.

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

Fuertes-Olivera, Pedro A. 2007. A corpus-based view of lexical gender in written business English. *English for Specific Purposes*, 26(2):219–234. https://doi.org/10.1016/j.esp.2006.07.001

Gardner, Matt, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. AllenNLP: A deep semantic natural language processing platform. *arXiv:1803.07640*.

Garera, Nikesh and David Yarowsky. 2009. Modeling latent biographic attributes in conversational genres. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 710–718, Suntec.

Garimella, Aparna and Rada Mihalcea. 2016. Zooming in on gender differences in social media. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 1–10, The COLING 2016 Organizing Committee, Osaka.

Garnham, Alan, Jane Oakhill, Marie-France Ehrlich, and Manuel Carreiras. 1995. Representations and processes in the interpretation of pronouns: New evidence from Spanish and French. *Journal of Memory and Language*, 34(1):41–62.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé, III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv:1803.09010*.

Genthial, Damien, Jacques Courtin, and Jacques Menezo. 1994. Towards a more user-friendly correction. In *COLING 1994 Volume 2: The 15th International Conference on Computational Linguistics*.

GLAAD. 2007. Media reference guide–transgender. https://www.glaad.org/reference/transgender.

Glavaš, Goran, Damir Korenčić, and Jan Šnajder. 2013. Aspect-oriented opinion mining from user reviews in Croatian. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 18–23, Sofia.

Goldberg, Yoav and Michael Elhadad. 2013. Word segmentation, unknown-word resolution, and morphological agreement in a Hebrew parsing system. *Computational Linguistics*, 39(1):121–160. https://doi.org/10.1162/COLI_a_00137

Gonen, Hila and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of NAACL-HLT*.

Grosz, Barbara J., Aravind K. Joshi, and Scott Weinstein. 1983. Providing a unified account of definite noun phrases in discourse. In *21st Annual Meeting of the Association for Computational Linguistics*, pages 44–50, Cambridge, MA.

Guha, Anupam, Mohit Iyyer, Danny Bouman, and Jordan Boyd-Graber. 2015. Removing the training wheels: A coreference dataset that entertains humans and challenges computers. In *NAACL*.

Guillou, Liane. 2012. Improving pronoun translation for statistical machine translation. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–10, Avignon.

Hagoort, Peter and Colin M. Brown. 1999. Gender electrified: ERP evidence on the syntactic nature of gender processing. *Journal of psycholinguistic research*, 28(6):715–728.

Hamidi, Foad, Morgan Klaus Scheuerman, and Stacy M. Branham. 2018. Gender recognition or gender reductionism? The social implications of embedded gender recognition systems. In *CHI*, page 8, ACM.

HaNasi, Judah. 189. Mishnah bikkurim. In *Mishnah*, Chapter 4.

Harabagiu, Sanda M. and Steven J. Maiorano. 1999. Knowledge-lean coreference resolution and its relation to textual cohesion and coherence. In *The Relation of Discourse/Dialogue Structure and Reference*.

Hardmeier, Christian and Marcello Federico. 2010. Modelling pronominal anaphora in statistical machine translation. *IWSLT*.

Hardmeier, Christian and Liane Guillou. 2018. Pronoun translation in English-French machine translation: An analysis of error types. *arXiv preprint arXiv:1808.10196*.

Hellinger, Marlis and Heiko Motschenbacher. 2015. *Gender Across Languages*, volume 4, John Benjamins Publishing Company.

Holan, Tomáš, Vladislav Kuboň, and Martin Plátek. 1997. A prototype of a grammar checker for Czech. In *Fifth Conference on Applied Natural Language Processing*, pages 147–154, Washington, DC.

Honnibal, Matthew and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Hord, Levi C. R. 2016. Bucking the Linguistic Binary: Gender Neutral Language in English, Swedish, French, and German. *Western Papers in Linguistics / Cahiers linguistiques de Western:* Vol. 3, Article 4.

Hovy, Dirk. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 752–762, Beijing.

Jagose, Annamarie. 1996. *Queer Theory: An Introduction*, NYU Press.

Jing, Hongyan, Nanda Kambhatla, and Salim Roukos. 2007. Extracting social networks and biographical facts from conversational speech transcripts. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 1040–1047, Prague.

Johannsen, Anders, Dirk Hovy, and Anders Søgaard. 2015. Cross-lingual syntactic variation over age and gender. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 103–112, Beijing.

Joshi, Aravind K. and Scott Weinstein. 1981. Control of inference: Role of some aspects of discourse structure-centering. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'81, pages 385–387, San Francisco, CA.

Kameyama, Megumi. 1986. A property-sharing constraint in centering. In *24th Annual Meeting of the Association for Computational Linguistics*, pages 200–206, New York, NY.

Karlekar, Sweta, Tong Niu, and Mohit Bansal. 2018. Detecting linguistic characteristics of Alzheimer's dementia by interpreting neural models. 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 701–707, New Orleans, LA.

Kay, Matthew, Cynthia Matuszek, and Sean A. Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. In *CHI*.

Kessler, Suzanne J. and Wendy McKenna. 1978. *Gender: An Ethnomethodological Approach*, University of Chicago Press.

Kestemont, Mike. 2014. Function words in authorship attribution. From black magic to theory? In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, pages 59–66, Gothenburg.

Keyes, Os. 2018. The misgendering machines: Trans/HCI implications of automatic gender recognition. *CHI*.

Kiritchenko, Svetlana 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, New Orleans, LA.

Kleinberg, Bennett, Maximilian Mozes, and Isabelle van der Vegt. 2018. Identifying the sentiment styles of YouTube's vloggers. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3581–3590, Brussels.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*, volume 5, pages 79–86.

Kokkinakis, Dimitrios, Ann Ighe, and Mats Malm. 2015. Gender-based vocation identification in Swedish 19th century prose fiction using linguistic patterns, NER and CRF learning. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 89–97, Denver, CO.

Koolen, Corina and Andreas van Cranenburgh. 2017. These are not the stereotypes you are looking for: Bias and fairness in authorial gender attribution. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 12–22, Valencia.

Krahmer, Emiel and Kees van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Kramarae, Cheris and Paula A. Treichler. 1985. *A Feminist Dictionary*, Pandora Press.

Kraus, Matthias, Johannes Kraus, Martin Baumann, and Wolfgang Minker. 2018.

Effects of gender stereotypes on trust and likability in spoken human-robot interaction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, European Language Resources Association (ELRA), Miyazaki.

Lakoff, Robin. 1975. *Language and Woman's Place*. New York: Harper and Row.

Lambert, Max and Melina Packer. 2019. How gendered language leads scientists astray. Washington Post. https://www.washingtonpost.com/outlook/2019/06/10/how-gendered-language-leads-scientists-astray/

Larson, Brian. 2017. Gender as a variable in natural-language processing: Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia.

Laws, Florian, Florian Heimerl, and Hinrich Schütze. 2012. Active learning for coreference resolution. In *NAACL*.

Le Nagard, Ronan and Philipp Koehn. 2010. Aiding pronoun translation with co-reference resolution. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 252–261, Uppsala.

Levesque, Hector, Ernest Davis, and Leora Morgenstern. 2012. The Winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Levinger, Moshe, Uzzi Ornan, and Alon Itai. 1995. Learning morpho-lexical probabilities from an untagged corpus with an application to Hebrew. *Computational Linguistics*, 21(3):383–404.

Levitan, Rivka. 2013. Entrainment in spoken dialogue systems: Adopting, predicting and influencing user behavior. In *Proceedings of the 2013 NAACL HLT Student Research Workshop*, pages 84–90, Atlanta, GA.

Levitan, Sarah Ita, Yocheved Levitan, Guozhen An, Michelle Levine, Rivka Levitan, Andrew Rosenberg, and Julia Hirschberg. 2016. Identifying individual differences in gender, ethnicity, and personality from dialogue for deception detection. In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, pages 40–44, San Diego, CA.

Levitan, Sarah Ita, Angel Maredia, and Julia Hirschberg. 2018. Linguistic cues to deception and perceived deception in interview dialogues. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1941–1950, New Orleans, LA.

Li, Dingcheng, Tim Miller, and William Schuler. 2011. A pronoun anaphora resolution system based on factorial hidden Markov models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1169–1178, Portland, Oregon, OR.

Li, Shoushan, Bin Dai, Zhengxian Gong, and Guodong Zhou. 2016. Semi-supervised gender classification with joint textual and social modeling. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2092–2100, Osaka.

Li, Wen and Markus Dickinson. 2017. Gender prediction for Chinese social media data. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 438–445, Varna.

Light, Ann. 2011. HCI as heterodoxy: Technologies of identity and the queering of interaction with computers. *Interacting with Computers*, 23(5):430–438.

Litvinova, Olga, Pavel Seredin, Tatiana Litvinova, and John Lyell. 2017. Deception detection in Russian texts. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 43–52, Valencia.

Liu, Yuanchao, Ming Liu, Xiaolong Wang, Limin Wang, and Jingjing Li. 2013. PAL: A chatterbot system for answering domain-specific questions. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 67–72, Sofia.

Ljubešić, Nikola, Darja Fišer, and Tomaž Erjavec. 2017. Language-independent gender prediction on Twitter. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 1–6, Vancouver.

López-Ludeña, Verónica, Rubén San-Segundo, Syaheerah Lufti, Juan Manuel Lucas-Cuesta, Julián David Echevarry, and Beatriz Martínez-González. 2011. Source language categorization for improving a speech into sign language translation system. In *Proceedings of the*

Second Workshop on Speech and Language Processing for Assistive Technologies, pages 84–93, Edinburgh.

Luo, Xiaoqiang. 2005. On coreference resolution performance metrics. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 25–32.

Lyons, John. 1977. *Semantics*. Cambridge University Press.

Mandravickaitė, Justina and Tomas Krilavičius. 2017. Stylometric analysis of parliamentary speeches: Gender dimension. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 102–107, Valencia.

Mani, Inderjeet, T. Richard Macmillan, Susann Luperfoy, Elaine Lusher, and Sharon Laskowski. 1993. Identifying unknown proper names in newswire text. In *Acquisition of Lexical Knowledge from Text*.

Marchal, Harmony, Benoît Lemaire, Maryse Bianco, and Philippe Dessus. 2008. A MDL-based model of gender knowledge acquisition. In *CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 73–80, Manchester.

Mareček, David, Rudolf Rosa, Petra Galuščáková, and Ondřej Bojar. 2011. Two-step translation with grammatical post-processing. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 426–432, Edinburgh.

Martinc, Matej and Senja Pollak. 2018. Reusable workflows for gender prediction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, European Language Resources Association (ELRA), Miyazaki.

Marton, Yuval, Nizar Habash, and Owen Rambow. 2010. Improving Arabic dependency parsing with lexical and inflectional morphological features. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 13–21, Los Angeles, CA.

Marton, Yuval, Nizar Habash, and Owen Rambow. 2013. Dependency parsing of modern standard Arabic with lexical and inflectional features. *Computational Linguistics*, 39(1):161–194. https://doi.org/10.1162/COLI_a_00138

Matthews, Austin, Waleed Ammar, Archna Bhatia, Weston Feely, Greg Hanneman, Eva Schlinger, Swabha Swayamdipta, Yulia Tsvetkov, Alon Lavie, and Chris Dyer. 2014. The CMU machine translation systems at WMT 2014. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 142–149, Baltimore, MD.

May, Chandler. 2019. Neurips2019.

Mellish, C. S. 1988. Implementing systemic classification by unification. *Computational Linguistics*, 14(1).

Merriam-Webster. 2016. Words we're watching: Singular 'they'. https://web.archive.org/web/20160625170705/https://www.merriam-webster.com/words-at-play/singular-nonbinary-they.

Mickus, Timothee, Olivier Bonami, and Denis Paperno. 2019. Distributional effects of gender contrasts across categories. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pages 174–184.

Miller, Timothy A., Dmitriy Dligach, and Guergana K. Savova. 2012. Active learning for coreference resolution. In *Workshop on Biomedical NLP*.

Mitchell, Alexis, Stephanie Strassel, Shudong Huang, and Ramez Zakhary. 2005. Ace 2004 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 1:1–1.

Mitkov, Ruslan. 1999. Introduction: Special issue on anaphora resolution in machine translation and multilingual NLP. *Machine Translation*, 14(3).

Mohammad, Saif, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in Tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, LA.

Mohammad, Saif and Tony Yang. 2011. Tracking sentiment in mail: How genders differ on emotional axes. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, pages 70–79, Portland, OR.

Moorthy, Sridhar, Ruth Pogacar, Samin Khan, and Yang Xu. 2018. Is Nike female? Exploring the role of sound symbolism in predicting brand name gender. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1128–1132, Brussels.

Moosavi, Nafise and Michael Strube. 2016. Which coreference evaluation metric do you trust? A proposal for a link-based entity aware metric, pages 632–642. Berlin, Germany.

Moosavi, Nafise Sadat. 2020. Robustness in Coreference Resolution. PhD dissertation, University of Heidelberg.

Mota, Cristina, Paula Carvalho, and Elisabete Ranchhod. 2004. Multiword lexical acquisition and dictionary formalization. In *Proceedings of the Workshop on Enhancing and Using Electronic Dictionaries*, pages 73–76, Geneva.

Mukherjee, Arjun and Bing Liu. 2010. Improving gender classification of blog authors. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 207–217, Cambridge, MA.

Mukund, Smruthi, Debanjan Ghosh, and Rohini Srihari. 2011. Using sequence kernels to identify opinion entities in Urdu. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 58–67, Portland, OR.

Nanba, Hidetsugu, Haruka Taguma, Takahiro Ozaki, Daisuke Kobayashi, Aya Ishino, and Toshiyuki Takezawa. 2009. Automatic compilation of travel information from automatically identified travel blogs. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 205–208, Suntec.

Narayanan, Ajit and Lama Hashem. 1993. On abstract finite-state morphology. In *Sixth Conference of the European Chapter of the Association for Computational Linguistics*, Utrecht.

Nastase, Vivi and Marius Popescu. 2009. What's in a name? In some languages, grammatical gender. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1368–1377, Singapore.

Navarretta, Costanza. 2004. An algorithm for resolving individual and abstract anaphora in Danish texts and dialogues. In *Proceedings of the Conference on Reference Resolution and Its Applications*, pages 95–102, Barcelona.

Neill, James. 2008. *The Origins and Role of Same-Sex Relations in Human Societies*.

Ng, Vincent. 2010. Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1396–1411, Uppsala.

Ng, Vincent and Claire Cardie. 2003. Bootstrapping coreference classifiers with multiple machine learning algorithms. In *EMNLP*.

Nguyen, Dong, Dolf Trieschnigg, A. Seza Doğruöz, Rilana Gravel, Mariët Theune, Theo Meder, and Franciska de Jong. 2014. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1950–1961, Dublin.

Nguyen, Dong, Dolf Trieschnigg, and Theo Meder. 2014. TweetGenie: Development, evaluation, and lessons learned. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 62–66, Dublin.

Nissen, Uwe Kjær. 2002. Aspects of translating gender. *Linguistik online*, 11(2):02.

Novák, Michal and Zdeněk Žabokrtský. 2014. Cross-lingual coreference resolution of pronouns. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 14–24, Dublin.

Ochs, Elinor. 1992. Indexing gender. *Rethinking Context: Language as an Interactive Phenomenon*, 11:335.

Olteanu, Alexandra, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2:13.

Orita, Naho, Eliana Vornov, Naomi Feldman, and Hal Daumé, III. 2015a. Why discourse affects speakers' choices of referring expressions. In *ACL*.

Orita, Naho, Eliana Vornov, Naomi H. Feldman, and Hal Daumé, III. 2015b. Why discourse affects speakers' choice of referring expressions. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Osterhout, Lee, Michael Bersick, and Judith McLaughlin. 1997. Brain potentials reflect violations of gender stereotypes. *Memory & Cognition*, 25(3):273–285.

Osterhout, Lee and Linda A. Mobley. 1995. Event-related brain potentials elicited by failure to agree. *Journal of Memory and Language*, 34(6):739–773.

Pakhomov, Serguei V., James Buntrock, and Christopher G. Chute. 2003. Identification of patients with congestive heart failure using a binary classifier: A case study. In *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*, pages 89–96, Sapporo.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA.

Park, Ji Ho, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels.

Patton, Desmond, Philipp Blandfort, William Frey, Michael Gaskell, and Svebor Karaman. 2019. Annotating Twitter data from vulnerable populations: Evaluating disagreement between domain experts and graduate student annotators. Grand Wailea, Hawaii.

Pérez Estruch, Carlos, Roberto Paredes Palacios, and Paolo Rosso. 2017. Learning multimodal gender profile using neural networks. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 577–582, Varna.

Pérez-Rosas, Verónica, Quincy Davenport, Anna Mengdan Dai, Mohamed Abouelenien, and Rada Mihalcea. 2017. Identity deception detection. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 885–894, Taipei.

Pirkola, Ari and Kalervo Järvelin. 1996. The effect of anaphor and ellipsis resolution on proximity searching in a text database. *Information Processing & Management*, 32.

Plank, Barbara. 2018. Predicting authorship and author traits from keystroke dynamics. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 98–104, New Orleans, LA.

Poesio, Massimo, Rosemary Stevenson, Barbara Di Eugenio, and Janet Hitzeman. 2004. Centering: A parametric theory and its instantiations. *Computational Linguistics*, 30(3):309–363. https://doi.org/10.1162/0891201041850911.

Popowich, Fred. 1989. Tree unification grammar. In *27th Annual Meeting of the Association for Computational Linguistics*, pages 228–236, Vancouver.

Post, Matt. 2018. A call for clarity in reporting BLEU scores. In *WMT*.

Prabhakaran, Vinodkumar, Emily E. Reid, and Owen Rambow. 2014. Gender and power: How gender and gender environment affect manifestations of power. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1965–1976, Doha.

Prasad, Grusha and Joanna Morris. 2020. The p600 for singular "they": How the brain reacts when John decides to treat themselves to sushi, PsyArXiv. psyarxiv.com/hwzke, https://doi.org/10.31234/osf.io/hwzke

Prates, Marcelo, Pedro Avelar, and Luis C. Lamb. 2019. Assessing gender bias in machine translation—a case study with Google Translate. *Neural Computing and Applications*.

Preoţiuc-Pietro, Daniel, Johannes Eichstaedt, Gregory Park, Maarten Sap, Laura Smith, Victoria Tobolsky, H. Andrew Schwartz, and Lyle Ungar. 2015. The role of personality, age, and gender in tweeting about mental illness. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 21–30, Denver, CO.

Qian, Peng, Xipeng Qiu, and Xuanjing Huang. 2016. Investigating language universal and specific properties in word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1478–1488, Berlin.

Quantz, J. Joachim. 1994. An HPSG parser based on description logics. In *COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics*.

Quirk, Chris and Simon Corston-Oliver. 2006. The impact of parse quality on syntactically-informed statistical machine translation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 62–69, Sydney.

Rabinovich, Ella, Raj Nath Patel, Shachar Mirkin, Lucia Specia, and Shuly Wintner. 2017. Personalized machine translation: Preserving original author traits. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1074–1084, Valencia.

Raghunathan, Karthik, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky,

and Christopher Manning. 2010. A multi-pass sieve for coreference resolution. In *EMNLP*.

Rahman, Altaf and Vincent Ng. 2011. Coreference resolution with world knowledge. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 814–824.

Ramakrishna, Anil, Nikolaos Malandrakis, Elizabeth Staruk, and Shrikanth Narayanan. 2015. A quantitative analysis of gender differences in movies using psycholinguistic normatives. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1996–2001, Lisbon.

Reali, Chiara, Yulia Esaulova, and Lisa Von Stockhausen. 2015. Isolating stereotypical gender in a grammatical gender language: Evidence from eye movements. *Applied Psycholinguistics*, 36(4):977–1006.

Reddy, Sravana and Kevin Knight. 2016. Obfuscating gender in social media writing. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 17–26, Austin, TX.

Richards, Christina, Walter Pierre Bouman, and Meg-John Barker. 2017. *Genderqueer and Non-Binary Genders*, Springer.

Risman, Barbara J. 2009. From doing to undoing: Gender as we know it. *Gender & Society*, 23(1):81–84.

Robaldo, Livio and Jurij Di Carlo. 2009. Disambiguating quantifier scope in DTS. In *Proceedings of the Eight International Conference on Computational Semantics*, pages 195–209, Tilburg.

Rojas-Barahona, Lina Maria, Thierry Bazillon, Matthieu Quignard, and Fabrice Lefevre. 2011. Using MMIL for the high level semantic annotation of the French MEDIA dialogue corpus. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)*.

Romanov, Alexey, Maria De-Arteaga, Hanna Hanna, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, Anna Rumshisky, and Adam Tauman Kalai. 2019. What's in a name? Reducing bias in bios without access to protected attributes. In *NAACL*.

Rudinger, Rachel, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *ACL Workshop on Ethics in NLP*, pages 74–79.

Rudinger, Rachel, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. 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 8–14, New Orleans, LA.

Sachan, Mrinmaya, Eduard Hovy, and Eric P. Xing. 2015. An active learning approach to coreference resolution. In *IJCAI*.

Sap, Maarten, Gregory Park, Johannes Eichstaedt, Margaret Kern, David Stillwell, Michal Kosinski, Lyle Ungar, and Hansen Andrew Schwartz. 2014. Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1146–1151, Doha.

Sap, Maarten, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation frames of power and agency in modern films. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2329–2334, Copenhagen.

Sapena, Emili, Lluís Padró, and Jordi Turmo. 2011. RelaxCor participation in CoNLL shared task on coreference resolution. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 35–39, Portland, OR.

Sarawgi, Ruchita, Kailash Gajulapalli, and Yejin Choi. 2011. Gender attribution: Tracing stylometric evidence beyond topic and genre. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 78–86, Portland, OR.

Schilt, Kristen and Laurel Westbrook. 2009. Doing Gender, Doing Heteronormativity: Gender Normals, Transgender People, and the Social Maintenance of Heterosexuality. *Gender & Society*, 23(4):440–464. https://doi.org/10.1177/0891243209340034

Schofield, Alexandra and Leo Mehr. 2016. Gender-distinguishing features in film dialogue. In *Proceedings of the Fifth Workshop on Computational Linguistics for Literature*, pages 32–39, San Diego, CA.

Schwartz, H. Andrew, Gregory Park, Maarten Sap, Evan Weingarten, Johannes Eichstaedt, Margaret Kern, David Stillwell, Michal Kosinski, Jonah Berger, Martin Seligman, and Lyle Ungar. 2015. Extracting human temporal orientation from Facebook language. In *Proceedings of the 2015 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 409–419, Denver, CO.

Serano, Julia. 2007. *Whipping Girl: A Transsexual Woman on Sexism and the Scapegoating of Femininity*, Seal Press.

Sidner, Candace L. 1981. Focusing for interpretation of pronouns. *American Journal of Computational Linguistics*, 7(4):217–231.

Sidorov, Maxim, Stefan Ultes, and Alexander Schmitt. 2014. Comparison of gender- and speaker-adaptive emotion recognition. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3476–3480, European Language Resources Association (ELRA), Reykjavik.

Silverstein, Michael. 1979. Language structure and linguistic ideology. *The Elements: A Parasession on Linguistic Units and Levels*, pages 193–247.

Smith, Jason R., Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. 2013. Dirt cheap Web-scale parallel text from the common crawl. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1374–1383.

Smith, Noah A., David A. Smith, and Roy W. Tromble. 2005. Context-based morphological disambiguation with random fields. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 475–482, Vancouver.

Soler-Company, Juan and Leo Wanner. 2014. How to use less features and reach better performance in author gender identification. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1315–1319, European Language Resources Association (ELRA), Reykjavik.

Soler-Company, Juan and Leo Wanner. 2017. On the relevance of syntactic and discourse features for author profiling and identification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 681–687, Valencia.

Soloman, Danny and Mary McGee Wood. 1994. Learning a radically lexical grammar. *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*.

Spivak, Michael. 1997. *The Joy of TEX: A Gourmet Guide to Typesetting with the AMS-TEX Macro Package*, 1st edition. American Mathematical Society.

Stewart, Ian. 2014. Now we stronger than ever: African-American English syntax in Twitter. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 31–37, Gothenburg.

Stoyanov, Veselin, Nathan Gilbert, Claire Cardie, and Ellen Riloff. 2009. Conundrums in noun phrase coreference resolution: Making sense of the state-of-the-art. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 656–664.

Streiter, Oliver, Leonhard Voltmer, and Yoann Goudin. 2007. From tombstones to corpora: TSML for research on language, culture, identity and gender differences. In *Proceedings of the 21st Pacific Asia Conference on Language, Information and Computation*, pages 450–458, Seoul.

Stryker, Susan. 2008. *Transgender History*, Seal Press.

Sweeney, Latanya. 2013. Discrimination in online ad delivery. *ACM Queue*.

Tadić, Marko and Sanja Fulgosi. 2003. Building the Croatian morphological lexicon. In *Proceedings of the 2003 EACL Workshop on Morphological Processing of Slavic Languages*, pages 41–45, Budapest.

Taniguchi, Tomoki, Shigeyuki Sakaki, Ryosuke Shigenaka, Yukihiro Tsuboshita, and Tomoko Ohkuma. 2015. A weighted combination of text and image classifiers for user gender inference. In *Proceedings of the Fourth Workshop on Vision and Language*, pages 87–93, Lisbon.

Tatman, Rachael. 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, Valencia.

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in OPUS. In *LREC*.

Tran, Trang and Mari Ostendorf. 2016. Characterizing the language of online communities and its relation to community reception. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1030–1035, Austin, TX.

Vaidya, Ashwini, Owen Rambow, and Martha Palmer. 2014. Light verb constructions with 'do' and 'be' in Hindi: A TAG analysis. In *Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing*, pages 127–136, Dublin.

van der Goot, Rob, Nikola Ljubešić, Ian Matroos, Malvina Nissim, and Barbara Plank. 2018. Bleaching text: Abstract features for cross-lingual gender prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 383–389, Melbourne.

Vanmassenhove, Eva, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels.

van Peursen, Wido. 2009. How to establish a verbal paradigm on the basis of ancient Syriac manuscripts. In *Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages*, pages 1–9, Athens.

Vaughan, Jennifer Wortman and Hanna Wallach. 2019. Microsoft research webinar: Machine learning and fairness. https://note.microsoft.com/MSR-Webinar-Machine-Learning-and-Fairness-Registration-LIVE/.

Verhoeven, Ben, Iza Škrjanec, and Senja Pollak. 2017. Gender profiling for Slovene Twitter communication: The influence of gender marking, content and style. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 119–125, Valencia.

Vilain, Marc, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Conference on Message Understanding*, pages 45–52.

Vogel, Adam and Dan Jurafsky. 2012. He said, she said: Gender in the ACL anthology. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 33–41, Jeju Island.

Vogt, Thurid and Elisabeth André. 2006. Improving automatic emotion recognition from speech via gender differentiaion. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, European Language Resources Association (ELRA), Genoa.

Volkova, Svitlana, Theresa Wilson, and David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1815–1827, Seattle, WA.

Walsh, Bill. 2015. The post drops the 'mike'–and the hyphen in 'e-mail'. Washington Post; https://web.archive.org/web/20170316091712/https://www.washingtonpost.com/opinions/the-post-drops-the-mike--and-the-hyphen-in-e-mail/2015/12/04/ccd6e33a-98fa-11e5-8917-653b65c809eb_story.html?tid=a_inl.

Wandruszka, Mario. 1969. *Sprachen: vergleichbar und unvergleichlich*, R. Piper & Company.

Wang, Zijian and David Jurgens. 2018. It's going to be okay: Measuring access to support in online communities. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 33–45, Brussels.

Webster, Kellie, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617. https://doi.org/10.1162/tacl_a_00240

Weischedel, Ralph, Eduard Hovy, Mitchell Marcus, Martha Palmer, Robert Belvin, Sameer Pradhan, Lance Ramshaw, and Nianwen Xue. 2011. OntoNotes: A Large Training Corpus for Enhanced Processing.

Weller, Marion, Alexander Fraser, and Sabine Schulte im Walde. 2013. Using subcategorization knowledge to improve case prediction for translation to German. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 593–603, Sofia.

West, Candace and Don H. Zimmerman. 1987. Doing gender. *Gender & society*, 1(2):125–151.

Wolf, Thomas. 2017. State-of-the-art neural coreference resolution for chatbots. Blog post, https://medium.com/huggingface/state-of-the-art-neural-coreference-resolution-for-chatbots-3302365dcf30.

Wood-Doughty, Zach, Nicholas Andrews, Rebecca Marvin, and Mark Dredze. 2018. Predicting Twitter user demographics from names alone. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 105–111, New Orleans, LA.

Yoshimoto, Kei. 1988. Identifying zero pronouns in Japanese dialogue. In *Coling Budapest 1988 Volume 2: International Conference on Computational Linguistics*.

Young, Meg, Lassana Magassa, and Batya Friedman. 2019. Toward inclusive tech policy design: A method for underrepresented voices to strengthen tech policy documents. *Ethics Inf Technol*. 21, 89–103. https://doi.org/10.1007/s10676-019-09497-z

Yu, Bei. 2012. Function words for Chinese authorship attribution. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 45–53, Montréal.

Zaghouani, Wajdi and Anis Charfi. 2018. Arap-Tweet: A large multi-dialect Twitter corpus for gender, age and language variety identification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, European Language Resources Association (ELRA), Miyazaki.

Zhang, Dong, Shoushan Li, Hongling Wang, and Guodong Zhou. 2016. User classification with multiple textual perspectives. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2112–2121, Osaka.

Zhao, Jieyu, 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 2979–2989, Copenhagen.

Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. 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, LA.

Zhao, Jieyu, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels.

Zmigrod, Ran, Sebastian J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. *arXiv preprint arXiv:1906.04571*.

Zock, Michael, Gil Francopoulo, and Abdellatif Laroui. 1988. Language learning as problem solving. In *Coling Budapest 1988 Volume 2: International Conference on Computational Linguistics*.