Toward Gender-Inclusive Coreference Resolution

YANG TRISTA CAO, University of Maryland
HAL DAUMÉ III, Microsoft Research & University of Maryland

ABSTRACT

Correctly resolving textual mentions of people fundamentally entails making inferences about those people. Such inferences raise the risk of systemic biases in coreference resolution systems, including biases that reinforce cis-normativity and 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 system. We inspect many existing datasets for trans-exclusionary biases, and develop two new datasets for interrogating bias in 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.

1 INTRODUCTION

Coreference resolution—the task of determining which references in text resolve to the same real-world entity—fundamentally requires making inferences about those entities.[1] Especially when those entities are people, coreference resolution systems run the risk of making unlicensed inferences, which can result 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 (for instance, in pronouns in English, and morphology in Arabic) and because societal expectations and stereotypes around gender roles may be explicitly or implicitly assumed by speakers. 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, p. 332].

Gender bias in coreference resolution can manifest in many ways; previous work by Rudinger et al. [2018], Zhao et al. [2018], and Webster et al. [2018] have focused largely on the particular case of *binary* gender discrimination in trained coreference systems, showing that current systems tend to over-rely on social stereotypes (particularly occupational stereotypes) when resolving "he" and "she" pronouns (see §4). Building on this work, in this paper we consider gender bias from a broader conceptual frame: that of gender inclusion beyond the binary "folk" model of gender. In particular, 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 of binary and non-binary trans people[2]. Addressing such issues is critical not just to minimizing oppression and harms and improving the quality of our systems, but also to minimizing the effect our results can have in furthering bigotry in the world [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 who may easily face harms when coreference is used in the context of machine translation or search engine systems (discussed in detail in §3.5).

[1] We use "coreference resolution" to refer also to anaphora resolution and entity detection and tracking.

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

Working draft, in progress, feedback appreciated.
2019. ACM ISBN Prepublication. Author contact: ycao95@cs.umd.edu and me@hal3.name.

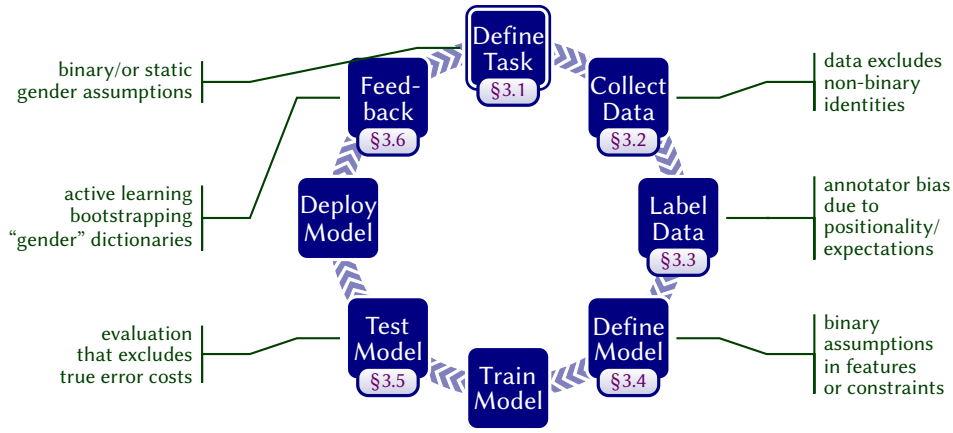


Fig. 1. Machine learning lifecycle, with pointers to sections in this paper and some possible sources of bias listed for each state of the lifecycle; adapted from Vaughan and Wallach [2019].

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 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 making errors in refering to hirself. This can lead to quality of service, stereotyping and denigration harms.
- ◇ *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 results in directly misgendering 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 complicate (a) what "gender" means and (b) how harms can enter into machine learning-based systems. Toward (a), we begin with a review (§2) 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. 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 socio-linguistics), 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[3]. We analyze forms of gender bias in coreference

[3]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.

resolution in six of the eight stages of the lifecycle in Figure 1. We conduct much of our analysis around task definition (§3.1), bias in underlying text (§3.2), model definition (§3.4), and evaluation methodologies (§3.5) by evaluating prior coreference datasets, 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 (§3.3), and our analysis of feedback loops (§3.6), we construct two new coreference datasets: MAP (a similar dataset to GAP [Webster et al. 2018] but without binary gender constraints on which we can perform counterfactual manipulations; see §3.2) and GICoref (a fully annotated coreference resolution dataset annotated on written by and about trans people; see §3.2.1).[4] In all cases, we focus largely on harms due to over- and under-representation [Kay et al. 2015], replicating stereotypes [Caliskan et al. 2017; Sweeney 2013] (particular those that are cisnormative and/or heteronormative), and quality of service differentials [Buolamwini and Gebru 2018].

The primary contributions of this paper are:

- ◇ 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 (§3).
- ◇ Developing a novel ablation technique for measuring gender bias in coreference resolution annotations, focusing on the *human* bias that can enter into annotation tasks (§3.3).
- ◇ Constructing a new dataset, the Gender Inclusive Coreference dataset (GICoref), for testing performance of coreference resolution systems on texts that discuss non-binary and binary transgender people (§3.2.1).
- ◇ 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" (§2).

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 lifecycle of coreference resolutions' various task definition updates from MUC7 in 2001 through ACE in the mid 2008 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 §3.1 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 §2.2), 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 datasets, 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 (§2) and also in terms of the data we annotated (§3.2.1). We have attempted to partially compensate for this bias by intentionally including documents with non-Western non-binary expressions of gender in the GICoref dataset, but the dataset is Western-dominant.

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 §3.2). This led us to counterfactual text manipulation in §3.3, which, while useful, is essentially impossible to do flawlessly (additional discussion in §3.3.3). Finally, because the social construct of gender is fundamentally contested (discussion in §2.1), some of our results may apply only under some frameworks. The use of "toward" in the title of this paper is

[4] All data and code will be released under MIT, BSD or more liberal license after review, and datasets will be accomonpanied with a corresponding datasheet/data statement [Bender and Friedman 2018; Gebru et al. 2018].

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 BACKGROUND: LINGUISTIC & 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 §3.1), 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]:

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

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] 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 (§2.1), then how gender is reflected in language (§2.2), and finally how these two converge or diverge (§2.3). 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.

2.1 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* [Butler 1990; Risman 2009; West and Zimmerman 1987]. 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 anatomy 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[5]; for example, that one's gender may shift by time or context (often "genderfluid"), or that some people do not experience gender at all (often "agender") [Darwin 2017; Kessler and McKenna 1978; Richards et al. 2017; Schilt and Westbrook 2009]. These models of gender contrast with historic "folk" views (both in linguistics and many societies at large) which assume that ones gender 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.[6] In §3.1 we will analyze the degree to which NLP papers make assumptions that are trans-inclusive or trans-exclusive.

Social gender 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.). For 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."

2.2 Linguistic Gender

Our discussion of linguistic gender largely follows [Corbett 1991, 2013; Craig 1994; Hellinger and Motschenbacher 2015; Ochs 1992], 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). 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.

Grammatical gender 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-) social gender. For example, On the other hand, "noun class languages" have no obvious such correspondence, and typically have many more gender classes. Some languages have no grammatical gender at all; English is 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]).

Referential 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

[5]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.

[6]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].

most obvious examples in English are gendered third person pronouns ("she", "he"), including neopronouns ("ze", "em", etc.), but also includes cases like "policeman" when the intended referent of this noun has social gender "male" (though not when "policeman" is used non-referentially, as in "every policeman needs to hold others accountable").

Lexical gender refers to an 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 real world 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 §3.3, but first need to discuss ways in which notions of linguistic gender match (or mismatch) from notions of social gender.

2.3 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).

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.[7]

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 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. Throughout this paper, instead of attempting to map pronouns to some specific labeled gender term we simply use the nominative case of the pronoun. English third person personal pronouns include SHE and HE, as well as the non-binary use of singular THEY and neopronouns[8] like ZE/HIR and XEY/XEM. 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

[7] Dahl [2000] provides several several complications of this analysis.

[8] Neopronouns have been in usage since at least the 1970s with varied forms [Bustillos 2011; Merriam-Webster 2016].

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]

Table 1. Corpora analyzed in this paper.

saw themself in the mirror").[9] This ontology is due to Conrod [2018], who also investigates the degree to which these are judged grammatical by native English speakers, and which we will use to quantify data bias (§3.3).

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.

3 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 (§3.5).

3.1 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 [Bender and Friedman 2018; Gebru et al. 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 datasets on which coreference or anaphora are annotated in English (Table 1); eleven of these are corpora distributed by the Linguistic Data Consortium (LDC)[10], and two are not. The final non-LDC corpora do not (to our knowledge) have corresponding annotation guidelines, and are instead described in

[9]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].
[10]See https://catalog.ldc.upenn.edu/{LDC-ID}.

				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

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

associated papers. None of the annotation guidelines (or papers) give explicit guidance about what personal pronouns are to be considered, 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.

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 Conrod's [2018] categorization (see §2.2):

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 (datasets for which no relevant examples were provided are not listed). Overall, we see that in total across these seven datasets, 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* datasets (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 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.

3.2 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 which has never seen "hir" is unlikely to even know it's 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").

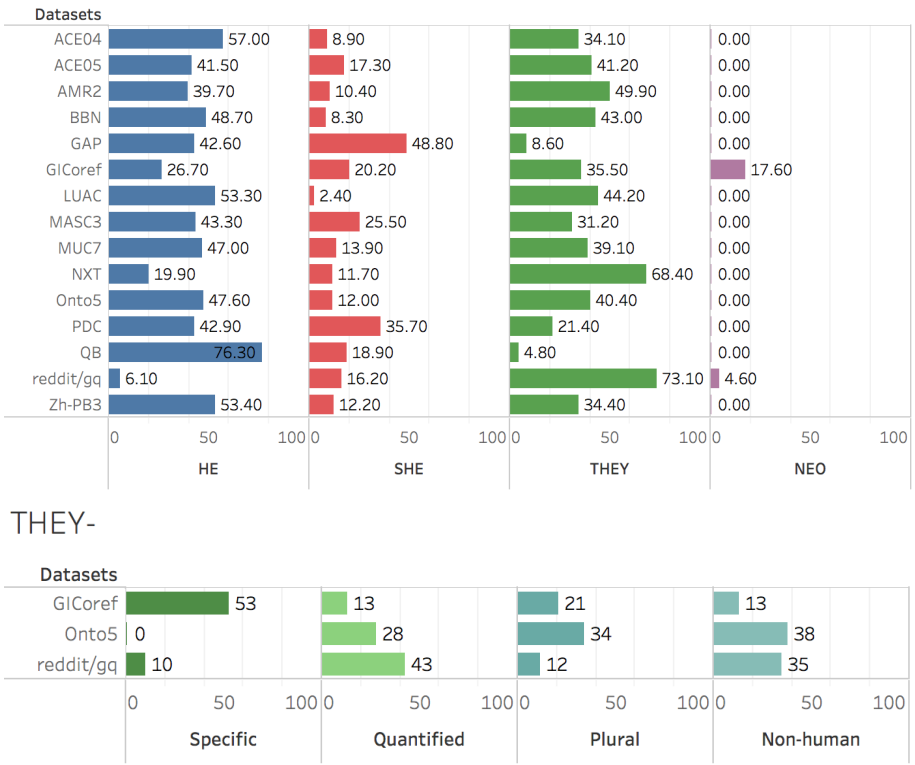


Fig. 2. (Top) For each dataset under consideration, the fraction of pronouns with different forms. Only our dataset (GICoref) and the genderqueer subreddit include neopronouns. (In the case of GAP, there *are* some occurrences of THEY, but they are never considered targets for coreference and so we exclude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

To analyze the possible impact of input data, we consider our thirteen coreference datasets, 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.[11] In the case of THEY, we distinguish between four uses of THEY: plural, singular, quantified (e.g., "Every student handed in their homework"), and non-human (referring, for instance, to a set of boxes). To achieve this, we annotated 100 examples uniformly at random by hand from OntoNotes [Weischedel et al. 2011]. Furthermore, we compare to the raw counts in a 2015 dump from some of Reddit discussion forum[12], and also limited to the `genderqueer` subforum. We additionally a new dataset for study, GICoref, described below in §3.2.1.

The results of this analysis are in Figure 2, including results on our new dataset, GICOREF. Overall, the examples used in the documentation of each of these datasets focuses entirely on binary gendered pronouns, generally with many more HE examples than SHE examples. Only the older datasets (MUC7 and Zh-PB3) include any examples of THEY, some of which are in a quantified form. Although the examples in annotation guidelines (or papers) are insufficient to fully tell what

[11]There was one instance of "hir" but that was a almost certainly typo for "his" (given the context), and several instances of "ey" used a contractions for plural THEY in transcripts of spoken English.
[12]From https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/

annotations were intended by the authors, they do provide a sense of what may have been top of mind in dataset construction.

3.2.1 New Dataset: Gender Inclusive Coreference. In order to make progress on gender inclusive coreference resolution, we introduce a new dataset, GICoref. This dataset is collected for the purpose to evaluate current coreference resolution models 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. 2018] and [Rudinger et al. 2018] (as well as in contrast to our MAP dataset), 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 dataset 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 (with the respective author's permission). Each author manually annotated each of these documents (additional detail of the dataset can be found in the accompanying datasheet), and then we jointly adjudicated the results. To reduce annotation time, any article that was substantially longer than 1000 words (pre-tokenization) was trimmed at the 1000th word.[13] 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. Two example annotated documents, one from Wikipedia, and one from Archive of Our Own, are provided in Appendix A and Appendix B respectively.

Although the majority of the examples in the dataset 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.

3.3 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. In order to evaluate this bias, we start with naturally occurring text in which we can cast the coreference task as a binary classification problem ("which of these two names does this pronoun refer to?"). We then generate counterfactual "augmentations" of this dataset by ablating the various notions of linguistic gender described in §2.2, similar to [Zmigrod et al. 2019]. We finally evaluate the impact of these ablations on human annotator agreement to answer the question: which forms of linguistic knowledge are most essentially 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"). While 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.

[13] There are four documents we accidentally trimmed at the 2000th word and so we keep the longer version of them with 2000 words in the dataset.

In all these cases, the plausible rough inference is that a reader takes a name, uses it to infer the social 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. [2018].

(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* shahidanaan walidatuha 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.

3.3.1 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 (note: agreement, not "accuracy"). To make the task easier for crowdsourcing, we follow the methodology of Webster et al.'s [2018] GAP dataset 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 us the GAP dataset directly, because their data is constrained that the "gender" of the two possible antecedents in "the same"[14]; for us, we are specifically interested in how annotators make decisions even when additional gender information is available. Thus, we construct a dataset called *Maybe Ambiguous Pronoun* (MAP), which is similar to the GAP dataset, but where we do not restrict the two names to match gender so that we can measure the influence of different gender cues. Details of the MAP dataset can be found in the accompanying datasheet.

One challenge in ablating gender information 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 (like (8)) 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 both to genderfluidity and misgendering).

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

Fortunately, it *is* possible to enumerate a reasonably high coverage list of English terms that signal lexical gender: terms of address (Mrs., Mr.) and lexically gendered nouns (mother).[15] The full list of terms of address we use and lexical gender words can be found in the accompanying datasheet; this was assembled by taking many online lists (mostly targeted at English language learners), merging them, and manual filtering.

To execute the "hiding" of various aspects of gender, we use the following substitutions:

(a) Replace all third person singular pronouns with a gender neutral program (THEY, XEY, ZE).
(b) Replace all names (e.g., "Aditya Modi") by a random name with only a first initial and last name (e.g., "B. Hernandez").

[14] It is unclear from the GAP dataset what notion of "gender" is used, nor how it was determined to be "the same."
[15] 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.

~~Mrs.~~ $\xrightarrow{(d)}$ ∅ ~~Rebekah Johnson Bobbitt~~ $\xrightarrow{(b)}$ **M. Booth** was the younger ~~sister~~ $\xrightarrow{(c)}$ **sibling** of ~~Lyndon B. Johnson~~ $\xrightarrow{(b)}$ **T. Schneider**, 36th President of the United States. Born in 1910 in Stonewall, Texas, ~~she~~ $\xrightarrow{(a)}$ **they** worked in the cataloging department of the Library of Congress in the 1930s before ~~her~~ $\xrightarrow{(a)}$ **their** ~~brother~~ $\xrightarrow{(c)}$ **sibling** entered politics.

Table 3. Example of applying *all* ablation substitutions for an example context in the GAP corpus. Each substitution type is marked.

(c) Replace all semantically gendered nouns (e.g., "mother") with a gender-indefinite variant (e.g., "parent").
(d) Remove all terms of address (e.g., "Mrs.", "Sir"); an alternative we did not explore would be to replace all with Mx. or Dr.

See Table 3 for an example that contains all of these substitutions.[16]

We perform two sets of experiments, one following a "forward selection" type ablation (start with everything removed and add them 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; while backward selection gives a sense of how much impact each type of gender cue has in the context of all the others.

The baseline setting we begin with is Zero, in which we apply all four substitutions. Because this also removes gender cues from the pronouns themselves, an annotator cannot even rely on social gender to perform these resolutions. We next consider adding back in the original pronouns (always either he or she in this data), yielding ¬Name ¬Sem ¬Addr. Any difference in annotation behavior between Zero and ¬Name ¬Sem ¬Addr can only be due to social gender stereotypes. 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 get ¬Name by only removing names, which gives a measure of how much names are used (in the context of all other cues); we get ¬Sem by only removing semantically gendered words; and ¬Addr by only removing terms of address.

3.3.2 Annotation Results. We construct examples using the methodology defined above and conduct annotation experiments using crowdworkers on Amazon Mechanical Turk (this study was approved by both the Microsoft Research Ethics Board and the University of Maryland IRB). Following the methodology by which the original GAP corpus was created, workers were shown ten contexts, with two names highlighted, and a pronoun highlighted, and were asked to determine which name the pronoun referred two. They were paid $1 to annotate these ten (the average annotation time

[16] We use NLTK [Loper and Bird 2002] to identify entities of type person for the name substitution. For name replacement, we use a list of 1000 most common surnames in the United States and use first initials that are not vowels (mostly to avoid "A" and "I"). We use spacy (https://spacy.io) to do part of speech tagging and dependency parsing, which are necessary to select the proper form of the substituted pronouns due to morphological ambiguities in English third person pronouns.

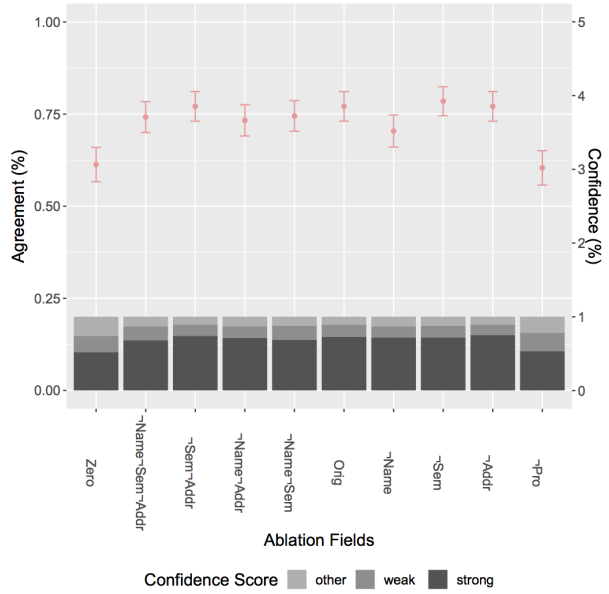


Fig. 3. Human annotation results for the ablation study on MAP dataset. Each column is a different ablation, and the y-axis is the degree of *agreement* with annotations performed on the original data, measured as accuracy. Red intervals are 95% significance intervals of human annotation agreement. Bottom bar plots are annotator certainties. Darkest areas indicate percentages of annotators who had high confidence in their answer ("definitely"); mid-grey areas and light grey areas indicate those with middling certainty ("probably") and unsure, respectively.

was seven minutes). Because we wanted to also capture uncertainty, we ask the crowdworkers how sure they are in their choice, between "definitely" sure, "probably" sure and "unsure."[17]

Figure 3 shows the human annotation results as binary classification agreement levels for resolving the pronoun to the antecedent annotated MAP data. If we compare Zero with ¬Name ¬Sem ¬Addr, and Orig with ¬Pro, we can see that removing pronouns leads to significant drop in agreement. This indicates that gender-based inferences, especially social gender stereotypes, play the most significant role when annotators resolve coreferences. This verifies the conclusion from [Rudinger et al. 2018] and [Zhao et al. 2018] that human annotated data incorporates bias from stereotypes.

Moreover if we compare ¬Name ¬Sem ¬Addr with ¬Sem ¬Addr and Orig with ¬Name, we see that name is another significant cue that annotators rely on. On the other hand, lexical gender and term of address cues do not have significant impacts on human annotation accuracies in our results. This is likely in part due to the low appearance frequency of those cues in our dataset. Every example has pronouns and names, whereas 48.6% of the examples have lexical gender cues but only 2.7% of the examples includes terms of address. Finally, we can see that annotators certainty values basically follows the same trend as the agreement, meaning that annotators have a reasonable sense of when they are unsure.

Finally, we note that agreement levels are essentially the same for Zero and ¬Pro. The difference between these is substantial: Zero has all removable forms of explicit gender removed, while ¬Pro only has pronouns replaced. In particular, in Table 3, the Zero has all of the rewrites allied, while

[17] 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.

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

Table 4. 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.

¬PRO only has "she"/"her" replaced with "they"/"their." This suggests that once explicit binary gender is gone from pronouns, the impact of any other form of linguistic gender in annotator's decisions is also removed.

3.3.3 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 aer's 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. 2015]). 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.

3.4 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 systems 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 models 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. [2018] and Rudinger et al. [2018], we evaluate a handful of freely available systems on the ablated data from §3.3. Third, we evaluate these systems on the dataset we created: Gender Inclusive Coreference (GICOREF).

3.4.1 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 began 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?[18]
- (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 4. 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;
- ◇ No papers treat gender as anything other than completely immutable;
- ◇ Only 7.1% (one paper!) considers neopronouns and/or specific singular THEY.

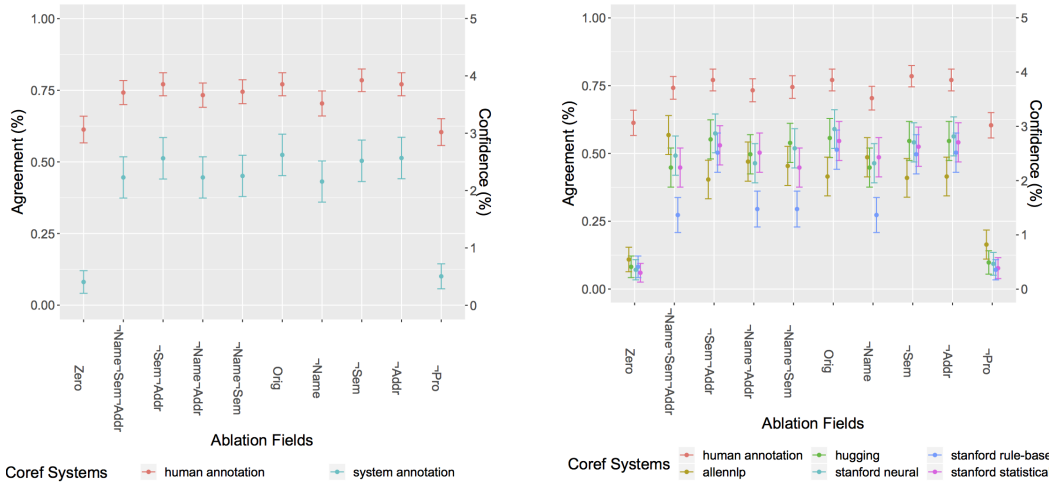
The situation for papers not specifically about coreference is similar (the majority of these papers are either purely linguistic papers about grammatical gender in languages other than English, or papers that do "gender recognition" of authors based on their writing). Overall, the situation more broadly is equally troubling, and generally also fails to escape from the folk theory of gender. In particular, none of the differences 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 knew exactly what decisions are baked in to all systems, the vast majority in our study come with strong gender binary assumptions, and exist within a broader sphere of literature which erases non-binary identities.

3.4.2 Coreference system performance on MAP. Next, we analyzed 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, which is a neural system based on spacy, and the Stanford deterministic [SfdD; Raghunathan et al. 2010], statistical [SfdS; Clark and Manning 2015] and neural [SfdN; Clark and Manning 2016] systems. Figure 4 shows the results. We can see that the average system accuracies mostly follow the same pattern as human agreement levels, though all are significantly lower than human results. In particular, lack of access to name information hurts many systems.

These results echo and extend previous observations made by Zhao et al. [2018], 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 dataset 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.

3.4.3 Coreference system performance on GICoref. Finally, we evaluate the same set of systems on our GICOREF dataset. Table 5 (left) reports results according the standard coreference resolution evaluation metric LEA [Moosavi and Strube 2016]. The first observation is that these systems hardly work at all; the best performing system achieves as F1 score of 14%. Additionally, we can see

[18]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.



(a) Human agreement in comparison to average system performance accuracy results for the ablation study on MAP dataset.

(b) Coreference resolution systems results for the ablation study on MAP dataset. Each color represents a system, while red represents human annotations.

Fig. 4. Accuracy scores for systems drop dramatically when we ablate out referential gender in pronouns. This reveals that those coreference resolution systems reply heavily on gender-based inferences. In terms of each systems, Huggingface and Stanford Neural systems have similar results and outperform other systems in most cases. Stanford Rule-based accuracy drops significantly once names are ablated.

LEA	Precision	Recall	F1
AI2	14.08	10.30	11.90
HF	27.74	7.53	11.85
SfdD	20.97	10.17	13.70
SfdS	19.82	5.45	8.55
SfdN	19.47	5.41	8.46

Recall	HE/SHE	THEY	NEO
AI2	96.41	92.96	25.35
HF	95.38	85.02	21.13
SfdD	97.61	96.57	0.00
SfdS	96.24	86.82	20.42
SfdN	96.58	90.07	0.00

Table 5. (Left) LEA scores on GICoref dataset with various coreference resolution systems. Rows are different systems while columns are precision, recall, and F1 scores. When evaluate, 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.

from the results that Stanford deterministic system outperforms the other four systems and that, in general, 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; see Table 5 (right). Here, we see 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 (including two systems that never identify any of them).

3.5 Bias in: Model 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 datasets [Buolamwini and Gebru 2018]). Alternatively, evaluation metrics may weight different errors in a way that is incongruous with their harm. For example, while misgendering via the use of an incorrect pronoun is 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 (§3.2 and §3.3) 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; Edens et al. 2003; Pirkola and Järvelin 1996]. In machine translation, producing correct gendered forms in gender languages often requires coreference to have been solved [Guillou 2012; Hardmeier and Federico 2010; Hardmeier and Guillou 2018; Mitkov 1999]. In the translation case, this then begs the question: which data are being used and how biased are they? 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 [Fraser 2008, e.g.,].

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, folk lore has it that getting pronouns "correct" does not really impact Bleu scores, which was recently confirmed (and alternatives proposed) by Guillou and Hardmeier [2018]. To quantify this bias, 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 datasets 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% (±0.22), which is barely statistically significant. Evaluated only on the ≈ 17% of the sentences in these datasets containing either HE or SHE, the degradation is about 3.1%. While this degradation is noticable, it perhaps does not reflect the real cost of such translation errors due to the high societal cost of misgendering.

3.6 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.[19] Active learning for coreference has been popular since the early 2000s, perhaps largely because coreference annotation is quite costly. Considering the approaches used in dominant papers, the active learning

[19]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.

	"Gender" % Likelihood					
	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

Table 6. 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).

algorithms used are not statistically unbiased [Guha et al. 2015; Laws et al. 2012; Miller et al. 2012; Ng and Cardie 2003; Sachan 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 III 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 (8150, 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[20] and queried the B+L data with them. In almost all cases, the full

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

name was unknown in the B+L data (or had counts less than five), and in which 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 6, where we can see that of those who use a pronoun other than SHE or HE (exclusively) are, essentially always, misgendered. Even in the best cases, the accuracy of this approach is only about 24%, and the errors occur much more frequently on pronouns other than HE. 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.

4 OTHER RELATED WORK

There are three primary papers that consider gender bias in coreference resolution, which have largely already been discussed. Rudinger et al. [2018] evaluates coreference systems for evidence of *occupational stereotypes*, by constructing Winograd-esque [Levesque et al. 2012] binary classification test examples like "*The paramedic* performed CPR on *the passenger* even though *[she/he/they]* knew it was too late." 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. [2018] proposed a very similar, also Winograd-esque scheme, also for measuring gender-based occupational stereotypes. In addition to reaching similar conclusions to Rudinger et al. [2018], this work also used a similar "counterfactual" data process as we used in order to provide additional training data to a coreference resolution system. Finally, Webster et al. [2018] produced the "Gender Ambiguous Pronoun" (GAP) dataset for evaluating coreference systems, by specifically seeking examples where "gender" (of some unspecified sort) could *not* be used to help coreference, because all entities in the surrounding context have the same "gender." They found that current 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.

Gender bias in NLP has been considered more broadly than just in coreference resolution, including, for instance, natural language inference [Rudinger et al. 2017], word embeddings [e.g., Bolukbasi et al. 2016; Gonen and Goldberg 2019; Romanov et al. 2019], sentiment analysis [Kiritchenko and Mohammad 2018], machine translation [Font and Costa-jussà 2019; Prates et al. 2019], among many others [Blodgett et al. 2019, inter alia]. Gender is also an object of study in gender recognition systems [Hamidi et al. 2018, e.g.,]. Much of this work has focused on gender bias with a (usually implicit) binary lens, an issue which 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 et al. [2014] and Reali et al. [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 et al. [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[21], or because different languages mark for gender in different ways: e.g., 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].

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 that 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 datasets for studying gender inclusion in coreference resolution. The MAP dataset we created counterfactually (and therefore it is subject to general concerns about counterfactual data construction), which allowed us to very precisely control how different types gender information. The GI Coref dataset 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 datasets show significant gaps in system performance, but perhaps moreso, 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 datasets and systems, we need to hire or consult experiential experts [Patton et al. 2019; Young et al. 2019].

More broadly, we found that assumptions around *gender* in 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 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, while 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. We also hope that developers of datasets, 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.

[21] 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 [Nissen 2002; Wandruszka 1969].

ACKNOWLEDGMENTS

The authors are grateful to a number of people who have provided pointers, edits, and suggestions to improve this work: Cassidy Henry, Marion Zepf, and Os Keyes all contributed to various aspects of this work, including suggestions for data sources for the GI Coref dataset. We also thank the CLIP lab at the University of Maryland for comments on previous drafts.

REFERENCES

Oshin Agarwal, 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.

Travis M. Andrews. 2017. The singular, gender-neutral 'they' added to the Associated Press Stylebook. Washington Post; permalink.

Jennifer E Arnold. 2008. Reference production: Production-internal and addressee-oriented processes. *Language and cognitive processes*, 23(4):495–527.

Amit Bagga 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.

Mohit Bansal 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. Association for Computational Linguistics.

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

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The Problem With Bias: Allocative Versus Representational Harms in Machine Learning. In *Proceedings of SIGCIS*.

Emily M. Bender. 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.

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

Shane Bergsma and Dekang Lin. 2006. Bootstrapping path-based pronoun resolution. In *COLING/ACL*.

Su Lin Blodgett, Hal Daumé, III, Hanna Wallach, and Solon Barocas. 2019. Debunking debiasing. In *Text as Data*.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Proceedings of NeurIPS*.

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

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

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

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

Maria Bustillos. 2011. Our desperate, 250-year-long search for a gender-neutral pronoun. permalink.

Judith Butler. 1990. Gender trouble.

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

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

Manuel Carreiras, 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).

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

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

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

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

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

Colette G Craig. 1994. Classifier languages. *The encyclopedia of language and linguistics*, 2:565–569.

Osten Dahl. 2000. Animacy and the notion of semantic gender. *Trends in linguistics studies and monographs*, 124:99–116.

Helana Darwin. 2017. Doing gender beyond the binary: A virtual ethnography. *Symbolic Interaction*, 40(3):317–334.

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

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

Ursula Doleschal and Sonja Schmid. 2001. Doing gender in Russian. *Gender Across Languages. The linguistic representation of women and men*, 1:253–282.

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

Elizabeth Du and Ross Liddy. 1990. Anaphora in natural language processing and information retrieval. *Inf. Process. Manage.*, 26.

Richard J. Edens, 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*.

Yulia Esaulova, 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).

Joel Escudé Font 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*.

Anke Frank, Chr Hoffmann, Maria Strobel, et al. 2004. Gender issues in machine translation. *Univ. Bremen*.

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

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

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

Pedro A Fuertes-Olivera. 2007. A corpus-based view of lexical gender in written business english. *English for Specific Purposes*, 26(2):219–234.

Matt Gardner, 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*.

Alan Garnham, 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).

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

GLAAD. 2007. Media reference guide–transgender. permalink.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. In *Proceedings of NAACL-HLT*.

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

Liane Guillou. 2012. Improving pronoun translation for statistical machine translation. In *Student Research Workshop at EACL*.

Liane Guillou and Christian Hardmeier. 2018. Automatic reference-based evaluation of pronoun translation misses the point. In *EMNLP*.

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

Foad Hamidi, 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.

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

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

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

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

Annamarie Jagose. 1996. *Queer theory: An introduction*. NYU Press.

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

Suzanne J. Kessler and Wendy McKenna. 1978. *Gender: An ethnomethodological approach*. University of Chicago Press.

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

Svetlana Kiritchenko and Saif M. Mohammad. 2018. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In *Proceedings of *SEM*.

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

Cheris Kramarae and Paula A Treichler. 1985. *A feminist dictionary*. Pandora Press.

Robin Lakoff. 1975. Language and woman's place. *New York ao: Harper and Row*.

Max Lambert and Melina Packer. 2019. How gendered language leads scientists astray. New York Times.

Brian N Larson. 2017. Gender as a variable in natural-language processing: Ethical considerations. In *ACL Workshop on Ethics in NLP*.

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

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

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

Edward Loper and Steven Bird. 2002. NLTK: the natural language toolkit. *arXiv preprint cs/0205028*.

Xiaoqiang Luo. 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. Association for Computational Linguistics.

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

Merriam-Webster. 2016. Words we're watching: Singular 'they'. permalink.

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

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

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

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

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

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

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

Elinor Ochs. 1992. Indexing gender. *Rethinking context: Language as an interactive phenomenon*, 11:335.

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

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

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

Desmond Upton Patton, Philipp Blandfort, William R Frey, Michael B Gaskell, and Svebor Karaman. 2019. Annotating twitter data from vulnerable populations: Evaluating disagreement between domain experts and graduate student annotators.

Ari Pirkola and Kalervo Järvelin. 1996. The effect of anaphor and ellipsis resolution on proximity searching in a text database. *Inf. Process. Manage.*, 32.

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

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

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. A multi-pass sieve for coreference resolution. In *EMNLP*.

Altaf Rahman 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. Association for Computational Linguistics.

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

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

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

Alexey Romanov, Maria De-Arteaga, Hanna Wallach, 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*.

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

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *NAACL*.

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

Kristen Schilt and Laurel Westbrook. 2009. Doing gender, doing heteronormativity. *Gender & Society*, 23(4).

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

Michael Silverstein. 1979. Language structure and linguistic ideology. *The elements: A parasession on linguistic units and levels*, pages 193–247.

Jason R Smith, 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 *ACL*.

Veselin Stoyanov, Nathan Gilbert, Claire Cardie, and Ellen Riloff. 2009. Conundrums in noun phrase coreference resolution: Making sense of the state-of-the-art. In *ACL*.

Susan Stryker. 2008. *Transgender history*. Seal Press.

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

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

Jennifer Wortman Vaughan and Hanna Wallach. 2019. Microsoft research webinar: Machine learning and fairness. permalink.

Marc Vilain, 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. Association for Computational Linguistics.

Bill Walsh. 2015. The post drops the 'mike'–and the hyphen in 'e-mail'. Washington Post; permalink.

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

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *TACL*.

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

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

Meg Young, Lassana Magassa, and Batya Friedman. 2019. Toward inclusive tech policy design: A method for underrepresented voices to strengthen tech policy documents. *Ethics and Information Technology*, pages 1–15.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicenté Ordoñez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *NAACL*.

Ran Zmigrod, 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*.

A 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 .

B 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 ... I think so , ” they begin , stuttering , but encouraged to continue by a cock of Soul 's head ; a social cue even they could read , “ The professor ... and Miss Marie asked if I 'd like to come and stay with them for the time being . ”

“ Oh , huh , Stein and Marie ? Nice , ” his brows lift , clearly some varying degree of happy for the other .

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

“ It seems like you 've got a good thing going with those two . ”

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

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

“ Y 'know , I think it 's less of an invitation and more of an extended welcome . ”

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

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

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

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

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

“ But ... meh . Different strokes . I ca n't judge . ”

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

“ Soul , I 'm sorry , ” the meister blurts .

Having been pent - up for months , the apology comes forth without inhibition , rolling effortlessly off their tongue .

“ Sorry ... ? For what ? ” Evans quirks a brow , chuckling .

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

“ F - for what ... ? ”

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

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

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

“ Hey , now , you ca n't take all the credit like that , Ragnarok did most of the damage , ” he ...