# Toward Gender-Inclusive Coreference Resolution

**Yang Trista Cao**
University of Maryland
ycao95@cs.umd.edu

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

## 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 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 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 build systems that lead to many potential harms.

## 1 Introduction

Coreference resolution—the task of determining which textual references resolve to the same real-world entity—requires making inferences about those entities. Especially when those entities are people, coreference resolution systems run the risk of making unlicensed inferences, possibly resulting in harms either to individuals or groups of people. Embedded in coreference inferences are varied aspects of gender, both because gender can show up explicitly (e.g., pronouns in English, morphology in Arabic) and because societal expectations and stereotypes around gender roles may be explicitly or implicitly assumed by speakers or listeners. This can lead to significant biases in coreference resolution systems: cases where systems "systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others" (Friedman and Nissenbaum, 1996, p. 332).

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

Building on both lines of work, and inspired by Keyes's (2018) study of vision-based automatic gender recognition systems, we consider gender bias from a broader conceptual frame than the binary "folk" model. We investigate ways in which folk notions of gender—namely that there are two genders, assigned at birth, immutable, and in perfect correspondence to gendered linguistic forms—lead to the development of technology that is exclusionary and harmful of binary and non-binary trans and cis people.[2] Addressing such issues is critical not just to improve the quality of our systems, but more pointedly to minimize the harms caused by our systems by reinforcing existing unjust social hierarchies (Lambert and Packer, 2019).

There are several stakeholder groups who may easily face harms when coreference systems is used (Blodgett et al., 2020). Those harms includes several possible harms, both allocational and representation harms (Barocas et al., 2017), including quality of service, erasure, and stereotyping harms. Following Bender's (2019) taxonomy of stakehold-

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

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



*Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595
July 5 - 10, 2020. 

ers and Barocas et al.'s (2017) taxonomy of harms, there are several 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, pages about xem may be missed if "multiple mentions of query" is a ranking feature, and the system cannot resolve xyr pronouns ⇒ quality of service, erasure.
◇ *Direct: by choice.* If a grammar checker uses coreference, it may insist that an author writing hir third-person autobiography is repeatedly making errors when referring to hirself ⇒ quality of service, stereotyping, denigration.
◇ *Direct: not by choice.* If an information extraction system run on résumés relies on cisnormative assumptions, job experiences by a candidate who has transitioned and changed his pronouns may be missed ⇒ allocative, erasure.
◇ *Many stakeholders.* If a machine translation system uses discourse context to generate pronouns, then errors can results in directly misgendering subjects of the document being translated ⇒ quality of service, denigration, erasure.

To address such harms as well as understand where and how they arise, we need to complicate (a) what "gender" means and (b) how harms can enter into natural language processing (NLP) systems. Toward (a), we begin with a unifying analysis (§ 3) of how gender is socially constructed, and how social conditions in the world impose expectations around people's gender. Of particular interest is how gender is reflected in language, and how that both matches and potentially mismatches the way people experience their gender in the world. Then, in order to understand social biases around gender, we find it necessary to consider the different ways in which gender can be realized linguistically, breaking down what previously have been considered "gendered words" in NLP papers into finer-grained categories that have been identified in the sociolinguistics literature of lexical, referential, grammatical, and social gender.

Toward (b), we focus on how bias can enter into two stages of machine learning systems: data annotation (§ 4) and model definition (§ 5). We construct two new datasets: (1) MAP (a similar dataset to GAP (Webster et al., 2018) but without binary gender constraints) on which we can perform counterfactual manipulations and (2) GICoref (a fully annotated coreference resolution dataset written by and about trans people).[3] In all cases, we focus largely on harms due to over- and under-representation (Kay et al., 2015), replicating stereotypes (Sweeney, 2013; Caliskan et al., 2017) (particular those that are cisnormative and/or heteronormative), and quality of service differentials (Buolamwini and Gebru, 2018).

**The primary contributions of this paper are:** (1) Connecting existing work on gender bias in NLP to sociological and sociolinguistic conceptions of gender to provide a scaffolding for future work on analyzing "gender bias in NLP" (§3). (2) Developing an ablation technique for measuring gender bias in coreference resolution annotations, focusing on the *human* bias that can enter into annotation tasks (§4). (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 (§5).

## 2 Related Work

There are four recent papers that consider gender bias in coreference resolution systems. Rudinger et al. (2018) evaluates coreference systems for evidence of *occupational stereotyping*, by constructing Winograd-esque (Levesque et al., 2012) test examples. They find that humans can reliably resolve these examples, but systems largely fail at them, typically in a gender-stereotypical way. In contemporaneous work, Zhao et al. (2018a) proposed a very similar, also Winograd-esque scheme, also for measuring gender-based occupational stereotypes. In addition to reaching similar conclusions to Rudinger et al. (2018), this work also used a similar "counterfactual" data process as we use in § 4.1 in order to provide additional training data to a coreference resolution system. Webster et al. (2018) produced the GAP dataset for evaluating coreference systems, by specifically seeking examples where "gender" (left underspecified) could *not* be used to help coreference. They found that coreference systems struggle in these cases, also pointing to the fact that some success of current coreference systems is due to reliance on (binary) gender stereotypes. Finally, Ackerman (2019) presents an alternative breakdown of gender than we use (§ 3), and proposes matching criteria for model-

[3]Both datasets are released under a BSD license at github.com/TristaCao/into_inclusivecoref with corresponding datasheets (Gebru et al., 2018).

ing coreference resolution linguistically, taking a trans-inclusive perspective on gender.

Gender bias in NLP has been considered more broadly than just in coreference resolution, including, natural language inference (Rudinger et al., 2017), word embeddings (e.g., Bolukbasi et al., 2016; Romanov et al., 2019; Gonen and Goldberg, 2019), sentiment analysis (Kiritchenko and Mohammad, 2018), machine translation (Font and Costa-jussà, 2019; Prates et al., 2019; Dryer, 2013; Frank et al., 2004; Wandruszka, 1969; Nissen, 2002; Doleschal and Schmid, 2001), among many others (Blodgett et al., 2020, inter alia). Gender is also an object of study in gender recognition systems (Hamidi et al., 2018). 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 (2017b) and May (2019).

## 3 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. 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."

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 (Johnson et al., 2019; McLemore, 2015). This process of *misgendering* is problematic for both trans and cis individuals to the extent that transgender historian Stryker (2008) writes:

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

### 3.1 Sociological Gender

Many modern trans-inclusive models of gender recognize that *gender* encompasses many different aspects. These aspects include the experience that one has of gender (or lack thereof), the way that one expresses one's gender to the world, and the way that normative social conditions impose gender norms, typically as a dichotomy between masculine and feminine roles or traits (Kramarae and Treichler, 1985; West and Zimmerman, 1987; Butler, 1990; Risman, 2009; Serano, 2007). Gender self-determination, on the other hand, holds that each person is the "ultimate authority" on their own gender identity (Zimman, 2019; Stanley, 2014), with Zimman (2019) further arguing the importance of the role language plays in that determination.

Such trans-inclusive models deconflate anatomical and biological traits and the sex that a person had assigned to them at birth from one's gendered position in society; this includes intersex people, whose anatomical/biological factors do not match the usual designational criteria for either sex. Trans-inclusive views typically recognize that gender exists beyond the regressive "female"/"male" binary[4]; additionally, one's gender may shift by time or context (often "genderfluid"), and some people do not experience gender at all (often "agender") (Kessler and McKenna, 1978; Schilt and Westbrook, 2009; Darwin, 2017; Richards et al., 2017). In §5 we analyze the degree to which NLP papers make trans-inclusive or trans-exclusive assumptions.

**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.). Ackerman (2019) highlights a highly overlapping concept, "bio-social gender", which consists of gender role, gender expression, and gender identity. Taking gender role as an example, upon learning that a nurse is coming to their hospital room, a patient may form expectations that this person is likely to be "female," and 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

[4]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.

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.

### 3.2 Linguistic Gender

Our discussion of linguistic gender largely follows (Corbett, 1991; Ochs, 1992; Craig, 1994; Corbett, 2013; Hellinger and Motschenbacher, 2015; Fuertes-Olivera, 2007), departing from earlier characterizations that postulate a direct mapping from language to gender (Lakoff, 1975; Silverstein, 1979). Our taxonomy is related but not identical to (Ackerman, 2019), which we discuss in §2.

**Grammatical gender**, similarly defined in Ackerman (2019), is nothing more than a classification of nouns based on a principle of *grammatical agreement*. In "gender languages" there are typically 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-)[5] social gender. In comparison, "noun class languages" have no such correspondence, and typically many more classes. Some languages have no grammatical gender at all; English is generally seen as one (Nissen, 2002; Baron, 1971) (though this is contested (Bjorkman, 2017)).

**Referential gender** (similar, but not identical to Ackerman's (2019) "conceptual gender") relates linguistic expressions to extra-linguistic reality, typically identifying referents as "female," "male," or "gender-indefinite." Fundamentally, referential gender only exists when there is an entity being referred to, and their gender (or sex) is realized linguistically. The most obvious examples in English are gendered third person pronouns (SHE, HE), including neopronouns (ZE, EM) and singular THEY[6], 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" as well as gendered terms of address like "Mrs." Importantly, lexical gender is a property of the linguistic unit, *not* a property of its referent in the real world, which may or may not exist. For instance, in "Every son loves his parents", there is no real world referent of "son" (and therefore no *referential* gender), yet it still (likely) takes HIS as a pronoun anaphor because "son" has lexical gender MASC.

### 3.3 Social and Linguistic Gender Interplays

The relationship between these aspects of gender is complex, and none is one-to-one. The referential gender of an individual (e.g., pronouns in English) may or may not match their social gender 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 (for Transgender Equality, 2017)). The other linguistic forms of gender (grammatical, lexical) also need not match each other, nor match referential gender (Hellinger and Motschenbacher, 2015).

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 in English) or lexical gender. We focus on English, which has no grammatical gender, but does have lexical gender. English also marks referential gender on singular third person pronouns.

Below, we use this more nuanced notion of different types of gender to inspect how bias play out in coreference resolution systems. These biases may arise in the context of any of these notions of gender, and we encourage future work to extend care over and be explicit about what notions of gender are being utilized and when.

## 4 Bias in Human Annotation

A possible source of bias in coreference systems comes from human annotations on the data used to train them. Such biases can arise from a combination of (possibly) underspecified annotations guidelines and the positionality of annotators themselves. In this section, we study how different aspects of linguistic notions impact an annotator's

[5] 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.

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

~~Mrs.~~ $\xrightarrow{(d)} \emptyset$ ~~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.

Figure 1: Example of applying *all* ablation substitutions for an example context in the MAP corpus. Each substitution type is marked over the arrow and separately color-coded.

judgments of anaphora. This parallels Ackerman (2019) linguistic analysis, in which a Broad Matching Criterion is proposed, which posits that "matching gender requires at least one level of the mental representation of gender to be identical to the candidate antecedent in order to match."

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

### 4.1 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. We construct binary classification examples taken from Wikipedia pages, in which a single pronoun is selected, and two possible antecedent names are given, and the annotator must select which one. We cannot use Webster et al.'s GAP dataset directly, because their data is constrained that the "gender" of the two possible antecedents is "the same"[7]; 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) following Webster et al.'s approach, but we do not restrict the two names to match gender.

In ablating gender information, one challenge is that removing social gender cues (e.g., "nurse" tending female) is not possible because they can exist anywhere. Likewise, it is not possible to remove syntactic cues in a non-circular manner. For example in (1), syntactic structure strongly suggests the antecedent of "herself" is "Liang", making it less likely that "He" corefers with Liang later (though it is possible, and such cases exist in natural data due either to genderfluidity or misgendering).

(1) Liang saw herself in the mirror. . . *He*. . .

Fortunately, it *is* possible to enumerate a high coverage list of English terms that signal lexical gender: terms of address (Mrs., Mr.) and semantically gendered nouns (mother).[8] We assembled a list by taking many online lists (mostly targeted at English language learners), merging them, and manual filtering. The assembling process and the final list is published with the MAP dataset and its datasheet.

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

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

(b) ¬NAME: Replace names by random names with only a first initial and last name.

(c) ¬SEM: Replace semantically gendered nouns with gender-indefinite variants.

(d) ¬ADDR: Remove terms of address.[9]

See Figure 1 for an example of all substitutions.

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

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

[8] 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.

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

stereotypes; while backward selection gives a sense of how much impact each type of gender cue has in the context of all the others.

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

### 4.2 Annotation Results

We construct examples using the methodology defined above. We then conduct annotation experiments using crowdworkers on Amazon Mechanical Turk following the methodology by which the original GAP corpus was created[10]. Because we wanted to also capture uncertainty, we ask the crowdworkers how sure they are in their choices, between "definitely" sure, "probably" sure and "unsure."

Figure 2 shows the human annotation results as binary classification accuracy for resolving the pronoun to the antecedent. We can see that removing pronouns leads to significant drop in accuracy. This indicates that gender-based inferences, especially social gender stereotypes, play the most significant role when annotators resolve coreferences. This confirms the findings of Rudinger et al. (2018) and Zhao et al. (2018a) that human annotated data incorporates bias from stereotypes.

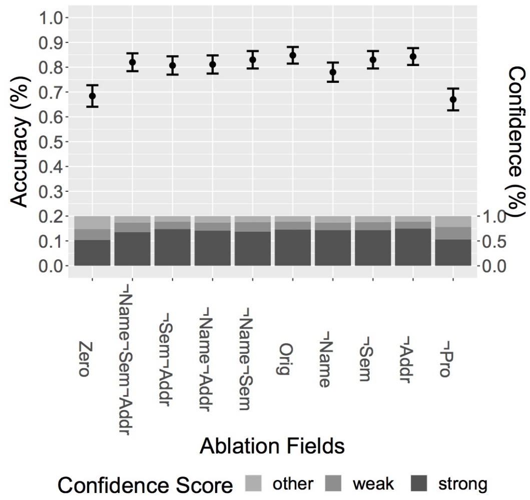


Figure 2: Human annotation results for the ablation study on MAP dataset. Each column is a different ablation, and the y-axis is the degree of *accuracy* with 95% significance intervals. Bottom bar plots are annotator certainties as how sure they are in their choices.

Moreover, if we compare ORIG with columns left to it, we see that name is another significant cue for annotator judgments, while lexical gender cues do not have significant impacts on human annotation accuracies. This is likely in part due to the low appearance frequency of lexical gender cues in our dataset. Every example has pronouns and names, whereas 49% of the examples have semantically gendered nouns but only 3% of the examples include terms of address. We also note that if we compare ¬NAME ¬SEM ¬ADDR to ¬SEM ¬ADDR and ¬NAME ¬ADDR, accuracy drops when removing gender cues. Though the differences are not statistically significant, we did not expect the accuracy drop.

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

## 5 Bias in Model Specifications

In addition to biases that can arise from the data that a system is trained on, as studied in the previ-

[10]Our study was approved by the Microsoft Research Ethics Board. Workers were paid $1 to annotate ten contexts (the average annotation time was seven minutes).

ous section, bias can also come from how models are structured. For instance, a system may fail to recognize anything other than a dictionary of fixed pronouns as possible referents to entities. Here, 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. Second, similar to Zhao et al. (2018a) and Rudinger et al. (2018), we evaluate five freely available systems on the ablated data from §4. Third, we evaluate these systems on the dataset we created: Gender Inclusive Coreference (GICOREF).

### 5.1 Cis-normativity in published NLP papers

In our first study, we adapt the approach Keyes (2018) took for analyzing the degree to which computer vision papers encoded trans-exclusive models of gender. In particular, we 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 resolution?
- Does the paper study English?
- **L.G**: Does the paper deal with linguistic gender (grammatical gender or gendered pronouns)?
- **S.G**: Does the paper deal with social gender?
- **L.G≠S.G**: (If yes to L.G and S.G:) Does the paper distinguish linguistic from social gender?
- **S.G Binary**: (If yes to S.G:) Does the paper explicitly or implicitly assume that social gender is binary?
- **S.G Immutable**: (If yes to S.G:) Does the paper explicitly or implicitly assume social gender is immutable?
- **They/Neo**: (If yes to S.G and to English:) Does the paper explicitly consider uses of definite singular "they" or neopronouns?

The results of this coding are in Table 1 (the full annotation is in Appendix B). 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;[11]

◇ Only 7.1% (one paper!) considers neopronouns and/or specific singular THEY.

| | All Papers | | Coref Papers | |
|---|---|---|---|---|
| L.G? | 52.6% | (of 150) | 95.4% | (of 22) |
| S.G? | 58.0% | (of 150) | 86.3% | (of 22) |
| L.G≠S.G? | 11.1% | (of 27) | 5.5% | (of 18) |
| S.G Binary? | 92.8% | (of 84) | 94.4% | (of 18) |
| S.G Immutable? | 94.5% | (of 74) | 100.0% | (of 14) |
| They/Neo? | 3.5% | (of 56) | 7.1% | (of 14) |

Table 1: Analysis of a corpus of 150 NLP papers that mention "gender" along the lines of what assumptions around gender are implicitly or explicitly made.

The situation for papers not specifically about coreference is similar (the majority of these papers are either purely linguistic papers about grammatical gender in languages other than English, or papers that do "gender recognition" of authors based on their writing; May (2019) discusses the (re)production of gender in automated gender recognition in NLP in much more detail). Overall, the situation more broadly is equally troubling, and generally also fails to escape from the folk theory of gender. In particular, none of the differences 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 is that although we do not know exactly what decisions are baked in to all systems, the vast majority in our study (including two papers by one of the authors (Daumé and Marcu, 2005; Orita et al., 2015)) come with strong gender binary assumptions, and exist within a broader sphere of literature which erases non-binary and binary trans identities.

### 5.2 System performance on MAP

Next, we analyze the effect that our different ablation mechanisms have on existing coreference resolutions systems. In particular, we run five coreference resolution systems on our ablated data: the AI2 system (AI2; Gardner et al., 2017), hugging face (HF; Wolf, 2017), which is a neural system based on spacy, and the Stanford deterministic (SfdD; Raghunathan et al., 2010), statistical (SfdS; Clark and Manning, 2015) and neural (SfdN; Clark and Manning, 2016) systems. Figure 3 shows the results. We can see that the system accuracies mostly follow the same pattern as human accuracy scores, though all are significantly lower than human results. Accuracy scores for systems drop

[11]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.

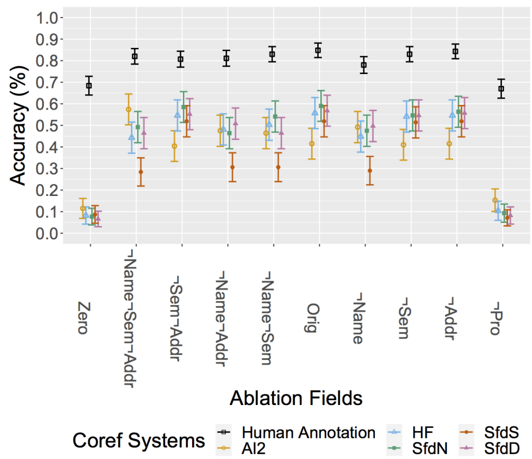


Figure 3: Coreference resolution systems results for the ablation study on MAP dataset. The y-axis is the degree of *accuracy* with 95% significance intervals.

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, HF and SfdN systems have similar results and outperform other systems in most cases. SfdD accuracy drops significantly once names are ablated.

These results echo and extend previous observations made by Zhao et al. (2018a), who focus on detecting stereotypes within occupations. They detect gender bias by checking if the system accuracies are the same for cases that can be resolved by syntactic cues and cases that cannot, with original data and reversed-gender data. Similarly, Rudinger et al. (2018) focus on detecting stereotypes within occupations as well. They construct 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.

### 5.3 System behavior on gender-inclusive data

Finally, in order to evaluate current coreference resolution models in gender inclusive contexts we introduce a new dataset, GICOREF. Here 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.

Our new GICOREF dataset consists of 95 documents from three types of sources: articles from 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)[12]. These documents were each annotated by both of the authors and adjudicated.[13] This data includes many examples of people who use pronouns other than SHE or HE (the dataset contains 27% HE, 20% SHE, 35% THEY, and 18% neopronouns, people who are genderfluid and whose names or pronouns change through the article, people who are misgendered, and people in relationships that are not heteronormative. In addition, *incorrect references* (misgendering and deadnaming[14]) are explicitly annotated.[15] Two example annotated documents, one from Wikipedia, and one from Archive of Our Own, are provided in Appendix C and Appendix D.

| | Precision | Recall | F1 |
|---|---|---|---|
| **AI2** | 40.4% | **29.2%** | 33.9% |
| **HF** | **68.8%** | 22.3% | 33.6% |
| **SfdD** | 50.8% | 23.9% | 32.5% |
| **SfdS** | 59.8% | 24.1% | **34.3%** |
| **SfdN** | 59.4% | 24.0% | 34.2% |

Table 2: LEA scores on GICOREF (incorrect reference excluded) with various coreference resolution systems. Rows are different systems while columns are precision, recall, and F1 scores. When evaluate, we only count exact matches of pronouns and name entities.

We run the same systems as before on this dataset. Table 2 reports results according the standard coreference resolution evaluation metric LEA (Moosavi and Strube, 2016). Since no systems are implemented to explicitly mark incorrect references, and no current evaluation metrics address this case, we perform the same evaluation twice. One with incorrect references included as regular references in the ground truth; and other with incorrect references excluded. Due to the limited number of incorrect references in the dataset, the

[12]See https://archiveofourown.org; thanks to Os Keyes for this suggestion.

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

[14]According to Clements (2017) deadnaming occurs when someone, intentionally or not, refers to a person who’s transgender by the name they used before they transitioned.

[15]Thanks to an anonymous reader of a draft version of this paper for this suggestion.

difference of the results are not significant. Here we only report the latter.

The first observation is that there is still plenty room for coreference systems to improve; the best performing system achieves as F1 score of 34%, but the Stanford neural system's F1 score on CoNLL-2012 test set reaches 60% (Moosavi, 2020). Additionally, we can see 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. We find that all systems achieve a recall of at least 95% for binary pronouns, a recall of around 90% on average for THEY, and a recall of around a paltry 13% for neopronouns (two systems—Stanford deterministic and Stanford neural—never identify any neopronouns at all).

## 6 Discussion and Moving Forward

Our goal in this paper was to analyze how gender bias exist in coreference resolution annotations and models, with a particular focus on how it may fail to adequately process text involving binary and non-binary trans referents. We thus created two datasets: MAP and GICOREF. Both datasets show significant gaps in system performance, but perhaps moreso, show that taking crowdworker judgments as "gold standard" can be problematic. 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).

Moreover, although we studied crowdworkers on Mechanical Turk (because they are often employed as annotators for NLP resources), if other populations are used for annotation, it becomes important to consider their positionality and how that may impact annotations. This echoes a related finding in annotation of hate-speech that annotator positionality matters (Olteanu et al., 2019). More broadly, we found that trans-exclusionary assumptions around *gender* in NLP papers is made commonly (and implicitly), a practice that we hope to see change in the future because it fundamentally limits the applicability of NLP systems.

The primary limitation of our study and analysis is that it is limited to English. This is particularly limiting because English lacks a grammatical gender system, and some extensions of our work to languages with grammatical gender are non-trivial. We also emphasize that while we endeavored to be inclusive, 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 and also in terms of the data we annotated. 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[16], but the dataset nonetheless remains 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. This led us to counterfactual text manipulation, which, while useful, is essentially impossible to do flawlessly. Moreover, our ability to evaluate coreference systems with data that includes *incorrect references* was limited as well, because current systems do not mark any forms of misgendering or deadnaming explicitly, and current metrics do not take this into account. Finally, because the social construct of gender is fundamentally contested, some of our results may apply only under some frameworks.

We hope this paper can serve as a roadmap for future studies. 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 can use some of our analysis as inspiration for how one can attempt to measure—and then root out—different forms of bias in coreference resolution systems and NLP systems more broadly.

## Acknowledgments

The authors are grateful to a number of people who have provided pointers, edits, suggestions, and annotation facilities to improve this work: Lauren Ackerman, Cassidy Henry, Os Keyes, Chandler May, Hanyu Wang, and Marion Zepf, 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.

[16]We endeavored to represent some non-Western gender identies that do not fall into the male/female binary, including 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).

## References


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

Lauren Ackerman. 2019. Syntactic and cognitive issues in investigating gendered coreference. *Glossa: a journal of general linguistics*, 4.

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

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

Sarah Alkuhlani and Nizar Habash. 2011. A corpus for modeling morpho-syntactic agreement in Arabic: Gender, number and rationality. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 357–362, Portland, Oregon, USA. Association for Computational Linguistics.

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

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

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

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

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

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

Janet Baker, Larry Gillick, and Robert Roth. 1994. Research in large vocabulary continuous speech recognition. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

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

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

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*.

Naomi S. Baron. 1971. A reanalysis of english grammatical gender. *Lingua*, 27:113–140.

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.

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

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

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

Bronwyn M. Bjorkman. 2017. Singular they and the syntactic representation of gender in english. *Glossa: A Journal of General Linguistics*, 2(1):80.

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

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

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*.

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

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

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.

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

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

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

Judith Butler. 1990. *Gender Trouble*. Routledge.

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

Michael Carl, Sandrine Garnier, Johann Haller, Anne Altmayer, and Bärbel Miemietz. 2004. Controlling gender equality with shallow NLP techniques. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 820–826, Geneva, Switzerland. COLING.

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

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

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

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

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

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*.

KC Clements. 2017. What is deadnaming? Blog post.

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

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.

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

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

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

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

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

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

Kareem Darwish, Ahmed Abdelali, and Hamdy Mubarak. 2014. Using stem-templates to improve Arabic POS and gender/number tagging. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2926–2931, Reykjavik, Iceland. European Language Resources Association (ELRA).

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.

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

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

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

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

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

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.

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

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

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

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

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

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*.

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*.

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

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

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

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

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

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

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

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

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*.

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

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

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.

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

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

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

Levi C. R. Hord. 2016. Bucking the linguistic binary: Gender neutral language in english, swedish, french, and german.

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

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

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

Kelly Johnson, Colette Auerswald, Allen J. LeBlanc, and Walter O. Bockting. 2019. 7. invalidation experiences and protective factors among non-binary adolescents. *Journal of Adolescent Health*, 64(2, Supplement):S4.

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

Sweta Karlekar, Tong Niu, and Mohit Bansal. 2018. Detecting linguistic characteristics of Alzheimer's dementia by interpreting neural models. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 701–707, New Orleans, Louisiana. Association for Computational Linguistics.

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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Verónica López-Ludeña, Rubén San-Segundo, Syaheerah Lufti, Juan Manuel Lucas-Cuesta, Julián David Echevarry, and Beatriz Martínez-González. 2011. Source language categorization for improving a speech into sign language translation system. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, pages 84–93, Edinburgh, Scotland, UK. Association for Computational Linguistics.

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

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

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

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

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

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

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

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

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

Chandler May. 2019. Neurips2019.

Kevin A. McLemore. 2015. Experiences with misgendering: Identity misclassification of transgender spectrum individuals. *Self and Identity*, 14(1):51–74.

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

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

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

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

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

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

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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Grusha Prasad and Joanna Morris. 2020. The p600 for singular "they": How the brain reacts when john decides to treat themselves to sushi. PsyArXiv.

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*.

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

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

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

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

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

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*.

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

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

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).

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

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

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 *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

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

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

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

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

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

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

H. Andrew Schwartz, Gregory Park, Maarten Sap, Evan Weingarten, Johannes Eichstaedt, Margaret Kern, David Stillwell, Michal Kosinski, Jonah Berger, Martin Seligman, and Lyle Ungar. 2015. Extracting human temporal orientation from Facebook language. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 409–419, Denver, Colorado. Association for Computational Linguistics.

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

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

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

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

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

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

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

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

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

E. Stanley. 2014. Gender self-determination. *TSQ: Transgender Studies Quarterly*, 1:89–91.

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

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

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

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

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

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

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

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

National Center for Transgender Equality. 2017. Understanding drag. Blog post.

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

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

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

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

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

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

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

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

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

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

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

Thomas Wolf. 2017. State-of-the-art neural coreference resolution for chatbots. Blog post.

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

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

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*.

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

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

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

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

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

Lal Zimman. 2019. Trans self-identification and the language of neoliberal selfhood: Agency, power, and the limits of monologic discourse. *International Journal of the Sociology of Language*, 2019:147–175.

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*.

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

## A Examples of Possible Bias in Data Annotation

Bias can enter coreference resolution datasets, which we use to train our systems, through annotation phase. Annotators may use linguistic notions to infer social gender. For instance, consider (2) 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 (3), an annotator is likely to have difficulty making a determination because both "Sue" and "Mary" suggest "her". In (4), 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.

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

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

(4) Liang and Aditya visited *her* mother.

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 (5) and terms of address (6)), grammatical gender references (in gender languages like Arabic (7)), and social gender references (8). The last of these ((8)) is the case in which the correct referent is likely to be least clear to most annotators, and also the case studied by Rudinger et al. (2018) and Zhao et al. (2018a).

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

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

(7) 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.

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

## B Annotation of ACL Anthology Papers

Below we list the complete set of annotations we did of the papers described in §5.1. For each of the papers considered, we annotate the following items:

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

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

| Citation | Coref | L.G | S.G | Eng | L≠S | 0/1 | Imm | Neo |
|---|---|---|---|---|---|---|---|---|
| Sidner (1981) | Y | Y | Y | Y | N | - | - | - |
| Bainbridge (1985) | Y | Y | N | Y | - | - | - | - |
| Kameyama (1986) | Y | Y | Y | Y | N | Y | Y | N |
| Mellish (1988) | N | Y | N | Y | - | - | - | - |
| Danlos and Namer (1988) | N | Y | N | N | - | - | - | - |
| Yoshimoto (1988) | N | Y | N | N | - | - | - | - |
| Zock et al. (1988) | N | Y | N | N | - | - | - | - |
| Popowich (1989) | N | Y | N | Y | - | - | - | - |
| Mani et al. (1993) | Y | N | Y | Y | - | Y | - | - |
| Narayanan and Hashem (1993) | N | Y | N | N | - | - | - | - |
| Soloman and Wood (1994) | N | Y | N | Y | - | - | - | - |
| Quantz (1994) | N | Y | N | Y | - | - | - | - |
| Baker et al. (1994) | - | - | - | - | - | - | - | - |
| Genthial et al. (1994) | N | Y | N | N | - | - | - | - |
| Levinger et al. (1995) | N | Y | N | N | - | - | - | - |
| Holan et al. (1997) | N | Y | N | N | - | - | - | - |
| Dorna et al. (1998) | N | N | N | Y | - | - | - | - |
| Harabagiu and Maiorano (1999) | Y | Y | Y | Y | N | Y | Y | N |
| Avgustinova and Uszkoreit (2000) | N | Y | N | N | - | - | - | - |
| Channarukul et al. (2000) | N | Y | N | Y | - | - | - | - |
| Abuleil et al. (2002) | N | Y | N | N | - | - | - | - |
| Cucerzan and Yarowsky (2003) | N | Y | N | N | - | - | - | - |
| Pakhomov et al. (2003) | N | N | Y | Y | - | - | - | - |
| Tadić and Fulgosi (2003) | N | Y | N | N | - | - | - | - |
| Debowski (2003) | N | Y | N | N | - | - | - | - |
| Navarretta (2004) | Y | Y | Y | N | N | Y | Y | - |
| Carl et al. (2004) | Y | Y | Y | N | N | Y | Y | - |
| Mota et al. (2004) | N | Y | N | Y | - | - | - | - |
| Eisner and Karakos (2005) | N | Y | N | Y | - | - | - | - |
| Boulis and Ostendorf (2005) | N | N | Y | Y | - | Y | Y | N |
| Smith et al. (2005) | N | Y | N | N | - | - | - | - |
| Bergsma and Lin (2006) | Y | Y | Y | Y | N | Y | Y | N |
| Vogt and André (2006) | N | N | Y | N | - | Y | Y | - |
| Quirk and Corston-Oliver (2006) | N | Y | N | Y | - | - | - | - |
| Dada (2007) | N | Y | N | N | - | - | - | - |

| Citation | Coref | L.G | S.G | Eng | L≠S | 0/1 | Imm | Neo |
|---|---|---|---|---|---|---|---|---|
| Streiter et al. (2007) | N | N | Y | N | - | - | - | - |
| Jing et al. (2007) | Y | Y | Y | Y | N | Y | - | N |
| Badr et al. (2008) | N | Y | N | N | - | - | - | - |
| Marchal et al. (2008) | N | Y | N | N | - | - | - | - |
| van Peursen (2009) | N | Y | N | N | - | - | - | - |
| Badr et al. (2009) | N | Y | N | N | - | - | - | - |
| Garera and Yarowsky (2009) | N | Y | Y | Y | N | Y | Y | N |
| Bergsma et al. (2009) | Y | Y | Y | Y | N | Y | Y | N |
| Nastase and Popescu (2009) | N | Y | N | N | - | - | - | - |
| Nanba et al. (2009) | N | N | N | Y | - | - | - | - |
| Robaldo and Di Carlo (2009) | N | N | N | Y | - | - | - | - |
| Mukherjee and Liu (2010) | N | N | Y | Y | - | Y | Y | - |
| Ng (2010) | Y | Y | Y | Y | N | Y | Y | N |
| Burkhardt et al. (2010) | N | N | Y | N | - | Y | Y | - |
| Marton et al. (2010) | N | Y | N | N | - | - | - | - |
| Le Nagard and Koehn (2010) | Y | Y | Y | Y | N | Y | Y | N |
| Rojas-Barahona et al. (2011) | N | Y | N | N | - | - | - | - |
| Mukund et al. (2011) | N | Y | N | N | - | - | - | - |
| Sarawgi et al. (2011) | N | N | Y | Y | - | Y | Y | N |
| Li et al. (2011) | Y | Y | Y | Y | N | Y | Y | N |
| Burger et al. (2011) | N | N | Y | Y | - | Y | Y | N |
| Mohammad and Yang (2011) | N | N | Y | Y | - | Y | Y | N |
| Sapena et al. (2011) | Y | Y | Y | Y | N | Y | Y | N |
| Charton and Gagnon (2011) | Y | Y | Y | Y | N | Y | Y | N |
| Alkuhlani and Habash (2011) | N | Y | N | N | - | - | - | - |
| Mareček et al. (2011) | N | Y | N | N | - | - | - | - |
| López-Ludeña et al. (2011) | N | Y | N | N | - | - | - | - |
| Declerck et al. (2012) | Y | Y | N | Y | - | - | - | - |
| Bergsma et al. (2012) | N | N | Y | Y | - | Y | Y | N |
| Alkuhlani and Habash (2012) | N | Y | N | N | - | - | - | - |
| Filippova (2012) | N | N | Y | Y | - | Y | - | - |
| Dinu et al. (2012) | N | Y | N | N | - | - | - | - |
| El Kholy and Habash (2012) | N | Y | N | N | - | - | - | - |
| Yu (2012) | N | N | N | N | - | - | - | - |
| Guillou (2012) | Y | Y | Y | Y | Y | Y | - | - |
| Vogel and Jurafsky (2012) | N | N | Y | Y | - | Y | Y | N |
| Goldberg and Elhadad (2013) | N | Y | N | N | - | - | - | - |
| Marton et al. (2013) | N | Y | N | N | - | - | - | - |
| Weller et al. (2013) | N | Y | N | Y | - | - | - | - |
| Ciot et al. (2013) | N | N | Y | N | - | Y | Y | - |
| Volkova et al. (2013) | N | N | Y | Y | - | Y | Y | N |
| Levitan (2013) | N | N | Y | Y | - | N | N | N |
| Bojar et al. (2013) | N | Y | N | N | - | - | - | - |
| Glavaš et al. (2013) | N | Y | N | N | - | - | - | - |
| Liu et al. (2013) | N | N | N | N | - | - | - | - |
| Kestemont (2014) | N | N | N | Y | - | - | - | - |
| Novák and Žabokrtský (2014) | Y | Y | N | Y | - | - | - | - |
| Babych et al. (2014) | N | Y | N | N | - | - | - | - |
| Soler-Company and Wanner (2014) | N | N | Y | Y | - | Y | Y | N |
| Chen and Ng (2014) | Y | Y | Y | Y | N | Y | Y | N |

| Citation | Coref | L.G | S.G | Eng | L≠S | 0/1 | Imm | Neo |
|---|---|---|---|---|---|---|---|---|
| Sap et al. (2014) | N | N | Y | Y | - | Y | Y | - |
| Nguyen et al. (2014a) | N | N | Y | Y | - | Y | Y | N |
| Prabhakaran et al. (2014) | N | N | Y | Y | - | Y | Y | N |
| Sidorov et al. (2014) | N | N | Y | Y | - | Y | Y | N |
| Darwish et al. (2014) | N | Y | N | N | - | - | - | - |
| Ahmed Khan (2014) | N | Y | N | N | - | - | - | - |
| Nguyen et al. (2014b) | N | N | Y | N | - | Y | Y | - |
| Stewart (2014) | N | N | Y | Y | - | Y | Y | - |
| Matthews et al. (2014) | N | Y | N | N | - | - | - | - |
| Vaidya et al. (2014) | N | Y | N | N | - | - | - | - |
| Kokkinakis et al. (2015) | N | Y | Y | N | N | Y | - | - |
| Johannsen et al. (2015) | N | N | Y | Y | - | Y | Y | - |
| Schwartz et al. (2015) | N | N | N | Y | - | - | - | - |
| Hovy (2015) | N | N | Y | Y | - | Y | Y | N |
| Agarwal et al. (2015) | N | Y | Y | Y | N | Y | Y | N |
| Preoţiuc-Pietro et al. (2015) | N | N | Y | Y | N | Y | Y | - |
| Ramakrishna et al. (2015) | N | Y | Y | Y | N | Y | Y | N |
| Taniguchi et al. (2015) | N | N | Y | Y | - | N | Y | N |
| Schofield and Mehr (2016) | N | N | Y | Y | - | Y | Y | N |
| Levitan et al. (2016) | N | N | Y | Y | - | Y | Y | N |
| Flekova et al. (2016) | N | N | Y | Y | - | Y | Y | N |
| Tran and Ostendorf (2016) | N | N | N | Y | - | - | - | - |
| Qian et al. (2016) | N | Y | N | Y | - | - | - | - |
| Li et al. (2016) | N | N | Y | Y | - | Y | Y | N |
| Zhang et al. (2016) | N | N | Y | Y | - | Y | Y | N |
| Garimella and Mihalcea (2016) | N | N | Y | Y | - | Y | Y | N |
| Reddy and Knight (2016) | N | N | Y | Y | - | Y | Y | N |
| Li and Dickinson (2017) | N | N | Y | N | - | Y | Y | - |
| Pérez Estruch et al. (2017) | N | N | Y | Y | - | Y | Y | N |
| Pérez-Rosas et al. (2017) | N | N | Y | Y | - | Y | Y | N |
| Rabinovich et al. (2017) | N | N | Y | N | - | Y | Y | - |
| Costa-jussà (2017) | N | Y | N | N | - | - | - | - |
| Sap et al. (2017) | N | N | Y | Y | - | Y | - | - |
| Zhao et al. (2017) | N | N | Y | Y | - | Y | Y | N |
| Mandravickaitė and Krilavičius (2017) | N | N | Y | Y | - | Y | Y | N |
| Verhoeven et al. (2017) | N | N | Y | Y | - | Y | Y | N |
| Larson (2017a) | N | Y | Y | Y | Y | N | N | Y |
| Koolen and van Cranenburgh (2017) | N | N | Y | N | - | N | Y | - |
| Tatman (2017) | N | N | Y | Y | - | Y | Y | N |
| Soler-Company and Wanner (2017) | N | N | Y | Y | - | Y | Y | N |
| Ljubešić et al. (2017) | N | N | Y | N | - | Y | Y | - |
| Litvinova et al. (2017) | N | N | Y | N | - | Y | Y | - |
| Mohammad et al. (2018) | N | N | Y | Y | - | Y | - | - |
| Wang and Jurgens (2018) | N | Y | Y | Y | Y | N | N | N |
| Kraus et al. (2018) | N | N | Y | Y | - | Y | - | - |
| Martinc and Pollak (2018) | N | N | Y | Y | - | Y | Y | N |
| Chan and Fyshe (2018) | N | N | Y | Y | - | Y | Y | N |
| Durmus and Cardie (2018) | N | N | N | Y | - | - | - | - |
| Zaghouani and Charfi (2018) | N | Y | Y | N | N | Y | Y | - |
| Plank (2018) | N | N | Y | Y | - | Y | Y | N |

| Citation | Coref | L.G | S.G | Eng | L≠S | 0/1 | Imm | Neo |
|---|---|---|---|---|---|---|---|---|
| Wood-Doughty et al. (2018) | N | N | Y | Y | - | Y | Y | N |
| Moorthy et al. (2018) | N | N | Y | Y | - | Y | - | - |
| Levitan et al. (2018) | N | N | Y | Y | - | Y | Y | N |
| Webster et al. (2018) | Y | Y | Y | Y | N | Y | Y | N |
| Park et al. (2018) | N | Y | Y | Y | N | Y | Y | N |
| Vanmassenhove et al. (2018) | N | Y | Y | N | N | Y | Y | - |
| Kleinberg et al. (2018) | N | N | Y | Y | - | Y | Y | N |
| Zhao et al. (2018b) | N | N | Y | Y | - | Y | Y | N |
| Balusu et al. (2018) | N | N | N | Y | - | - | - | - |
| Rudinger et al. (2018) | Y | Y | Y | Y | N | N | - | Y |
| Zhao et al. (2018a) | Y | Y | Y | Y | N | Y | Y | N |
| Kiritchenko and Mohammad (2018) | - | - | - | - | - | - | - | - |
| Barbieri and Camacho-Collados (2018) | N | N | Y | Y | - | Y | N | - |
| van der Goot et al. (2018) | N | N | Y | N | - | Y | Y | - |
| Karlekar et al. (2018) | N | N | Y | Y | - | Y | Y | N |
| de Gibert et al. (2018) | N | N | N | Y | - | - | - | - |
| Mickus et al. (2019) | N | Y | N | N | - | - | - | - |

## C 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 .

## D Example GICoref Document from AO3: Scar Tissue

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

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

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

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

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

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

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

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

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

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

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

" That ... makes sense . "

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

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

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

" Oh , huh , Stein$_H$ and Marie$_B$ $_G$ ? Nice , " his$_C$ brows lift , clearly some varying degree of happy for the other$_A$ .

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

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

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

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

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

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

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

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

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

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

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

His$_C$ comments comfort them$_A$ , an for a moment they$_A$ forget how this came to be . The cathedral in Italy , Lady Medusa$_I$ 's wrath , and the black blood that infected him$_C$ . Every moment they$_A$ spent in the presence of Soul Evans$_C$ builds always up to this ; fixation on the memories of their$_J$ first encounters and all the pain they$_A$ 've caused him$_C$ , the pain they$_A$ 've caused he$_C$ and Maka$_E$ $_K$ both . As quickly as Soul$_C$ had lifted the swordsman$_A$ 's spirits , they$_A$ 'd weighed themselves$_A$ down once more . It seemed so normal , though . Soul$_C$ could n't bring himself$_C$ to feel any sense of accomplishment in the coaxing - out of Crona$_A$ 's smile when the return of their$_A$ self doubt was as certain as the sun in the sky . His$_C$ own stubbornness could n't let his$_C$ diminished self worth lie .

With another encouraging smile , rows of sharpened incisors appearing oddly charismatic , he$_{C}$ opens his$_{C}$ mouth to speak – but finds himself$_{C}$ cut off before he$_{C}$ can even squeeze a word in .

“ Soul$_{C}$ , I$_{A}$ ’m sorry , ” the meister$_{A}$ blurts .

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

“ Sorry ... ? For what ? ” Evans$_{C}$ quirks a brow , chuckling .

He$_{C}$ adjusts his$_{C}$ stance to face Crona$_{A}$ with the whole of his$_{C}$ body , maintaining his$_{C}$ positive demeanor .

“ F - for what ... ? ”

They$_{A}$ stammer , shaking their$_{A}$ head . For all their$_{A}$ remorse , they$_{A}$ thought this would have been obvious .

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

Soul$_{C}$ seems stunned for a moment before harnessing his$_{C}$ quick wit .

“ Hey , now , you$_{A}$ ca n’t take all the credit like that , Ragnarok$_{L}$ did most of the damage , ” he$_{C}$ . . .