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 pronouns1 (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-

1Throughout, 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).

2Following 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).
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 them may be missed if “multiple mentions of query” is a ranking feature, and the system cannot resolve the pronouns ⇒ quality of service, erasure.
- **Direct**: by choice. If a grammar checker uses coreference, it may insist that an author writing his third-person autobiography repeatedly makes errors when referring to himself ⇒ 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 result 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). 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) (particularly 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-
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 (Zimmerman, 2019; Stanley, 2014), with Zimmerman (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; 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 refersents 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 THEY⁶, 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

---

⁵One 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.

⁶People’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).
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”;

...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). 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, KEY, 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.

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\textsuperscript{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.

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-

\textsuperscript{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;\(^{11}\)

<table>
<thead>
<tr>
<th></th>
<th>All Papers</th>
<th>Coref Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>L.G?</td>
<td>52.6%</td>
<td>95.4%</td>
</tr>
<tr>
<td>S.G?</td>
<td>58.0%</td>
<td>86.3%</td>
</tr>
<tr>
<td>L.G≠S.G?</td>
<td>11.1%</td>
<td>5.5%</td>
</tr>
<tr>
<td>S.G Binary?</td>
<td>92.8%</td>
<td>94.4%</td>
</tr>
<tr>
<td>S.G Immutable?</td>
<td>94.5%</td>
<td>100.0%</td>
</tr>
<tr>
<td>They/Neo?</td>
<td>3.5%</td>
<td>7.1%</td>
</tr>
</tbody>
</table>

\(^{11}\)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.

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 A12 system (A12; 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
dramatically when we ablate out referential gender in pronouns. This reveals that those coreference resolution systems rely 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). These documents were each annotated by both of the authors and adjudicated. 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) are explicitly annotated. Two example annotated documents, one from Wikipedia, and one from Archive of Our Own, are provided in Appendix C and Appendix D.

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

<table>
<thead>
<tr>
<th>System</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI2</td>
<td>40.4%</td>
<td>29.2%</td>
<td>33.9%</td>
</tr>
<tr>
<td>HF</td>
<td>68.8%</td>
<td>22.3%</td>
<td>33.6%</td>
</tr>
<tr>
<td>SfdD</td>
<td>50.8%</td>
<td>23.9%</td>
<td>32.5%</td>
</tr>
<tr>
<td>SfdS</td>
<td>59.8%</td>
<td>24.1%</td>
<td>34.3%</td>
</tr>
<tr>
<td>SfdN</td>
<td>59.4%</td>
<td>24.0%</td>
<td>34.2%</td>
</tr>
</tbody>
</table>

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.
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 dataset16, 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.

16We endeavored to represent some non-Western gender identities 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 mâhā (Hawaii).
References


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


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


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.


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


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


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


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.

المطربة والممثلة شاهدا والدتها

walidatu ha shahidanaa walidatu ha 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\(\neq\)S: (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).

<table>
<thead>
<tr>
<th>Citation</th>
<th>Coref</th>
<th>L.G</th>
<th>S.G</th>
<th>Eng</th>
<th>L(\neq)S</th>
<th>0/1</th>
<th>Imm</th>
<th>Neo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sidner (1981)</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Bainbridge (1985)</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Kameyama (1986)</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Mellish (1988)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Danlos and Namer (1988)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Yoshimoto (1988)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Zock et al. (1988)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Popowich (1989)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Mani et al. (1993)</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>Y</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Narayanan and Hashem (1993)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Soloman and Wood (1994)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Quantz (1994)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Baker et al. (1994)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Genthal et al. (1994)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Levinger et al. (1995)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Holan et al. (1997)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Dorna et al. (1998)</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Harabagiu and Maiorano (1999)</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Avgustinova and Uszkoreit (2000)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Channarukul et al. (2000)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Abuleil et al. (2002)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Cucerzan and Yarowsky (2003)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Pakhomov et al. (2003)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Tadić and Fulgosi (2003)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Debowski (2003)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Navarretta (2004)</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
</tr>
<tr>
<td>Carl et al. (2004)</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
</tr>
<tr>
<td>Mota et al. (2004)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Eisner and Karakos (2005)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Boulis and Ostendorf (2005)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Smith et al. (2005)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Bergsma and Lin (2006)</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Vogt and André (2006)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
</tr>
<tr>
<td>Quirk and Corston-Oliver (2006)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Dada (2007)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Citation</td>
<td>Coref</td>
<td>L.G</td>
<td>S.G</td>
<td>Eng</td>
<td>L≠S</td>
<td>0/1</td>
<td>Imm</td>
<td>Neo</td>
</tr>
<tr>
<td>----------</td>
<td>-------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Streiter et al. (2007)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Jing et al. (2007)</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>-</td>
<td>N</td>
</tr>
<tr>
<td>Badr et al. (2008)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Marchal et al. (2008)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>van Peursen (2009)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Badr et al. (2009)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Garera and Yarowsky (2009)</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>Bergsma et al. (2009)</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>Nastase and Popescu (2009)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Namba et al. (2009)</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Robaldo and Di Carlo (2009)</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Mukherjee and Liu (2010)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>Y</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Ng (2010)</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Burkhardt et al. (2010)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
</tr>
<tr>
<td>Marton et al. (2010)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Le Nagard and Koehn (2010)</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Rojas-Barahona et al. (2011)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Mukund et al. (2011)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Sarawgi et al. (2011)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Li et al. (2011)</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Burger et al. (2011)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Mohammad and Yang (2011)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Sapena et al. (2011)</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Charton and Gagnon (2011)</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Alkuhlani and Habash (2011)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Marcˇek et al. (2011)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>López-Ludeña et al. (2011)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Declerck et al. (2012)</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Bergsma et al. (2012)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Alkuhlani and Habash (2012)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Filippova (2012)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>Y</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Dinu et al. (2012)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>El Kholy and Habash (2012)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Yu (2012)</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Guillou (2012)</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Vogel and Jurafsky (2012)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Goldberg and Elhadad (2013)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Marton et al. (2013)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Weller et al. (2013)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Ciot et al. (2013)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
</tr>
<tr>
<td>Volkova et al. (2013)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Levitan (2013)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Bojar et al. (2013)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Glavaš et al. (2013)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Liu et al. (2013)</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Kestemont (2014)</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Novák and Žabokrtský (2014)</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Babych et al. (2014)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Soler-Company and Wanner (2014)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Chen and Ng (2014)</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Citation</td>
<td>Coref</td>
<td>L.G</td>
<td>S.G</td>
<td>Eng</td>
<td>L≠S</td>
<td>0/1</td>
<td>Imm</td>
<td>Neo</td>
</tr>
<tr>
<td>----------------------------------</td>
<td>-------</td>
<td>------</td>
<td>------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Sap et al. (2014)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Nguyen et al. (2014a)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Prabhakaran et al. (2014)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Sidory et al. (2014)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Darwish et al. (2014)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>N</td>
</tr>
<tr>
<td>Ahmed Khan (2014)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>N</td>
</tr>
<tr>
<td>Nguyen et al. (2014b)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Stewart (2014)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>N</td>
</tr>
<tr>
<td>Matthews et al. (2014)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>N</td>
</tr>
<tr>
<td>Vaidya et al. (2014)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>N</td>
</tr>
<tr>
<td>Kokkinakis et al. (2015)</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>-</td>
<td>N</td>
</tr>
<tr>
<td>Johanssen et al. (2015)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>N</td>
</tr>
<tr>
<td>Schwartz et al. (2015)</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>N</td>
</tr>
<tr>
<td>Hovy (2015)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Agarwal et al. (2015)</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Preotiuc-Pietro et al. (2015)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
</tr>
<tr>
<td>Ramakrishna et al. (2015)</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Taniguchi et al. (2015)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>N</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Schofield and Mehr (2016)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Levitan et al. (2016)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Flekova et al. (2016)</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Tran and Ostendorf (2016)</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>N</td>
</tr>
<tr>
<td>Qian et al. (2016)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>N</td>
</tr>
<tr>
<td>Li et al. (2016)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Zhang et al. (2016)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Garimella and Mihalcea (2016)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Reddy and Knight (2016)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Li and Dickinson (2017)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
</tr>
<tr>
<td>Pérez Estruch et al. (2017)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Pérez-Rosas et al. (2017)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Rabinovich et al. (2017)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
</tr>
<tr>
<td>Costa-jussà (2017)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Sap et al. (2017)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
</tr>
<tr>
<td>Zhao et al. (2017)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Mandrovickaitė and Krilavičius (2017)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Verbheeven et al. (2017)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Larson (2017a)</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Koolen and van Cranenburgh (2017)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>-</td>
<td>N</td>
<td>Y</td>
<td>-</td>
</tr>
<tr>
<td>Tatman (2017)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Soler-Company and Wanner (2017)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Ljubešić et al. (2017)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
</tr>
<tr>
<td>Litvinova et al. (2017)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
</tr>
<tr>
<td>Mohammad et al. (2018)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>Y</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Wang and Jurgens (2018)</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Krauss et al. (2018)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>Y</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Martinc and Pollak (2018)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Chan and Fyshe (2018)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Durmus and Cardie (2018)</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>N</td>
</tr>
<tr>
<td>Zaghouni and Charfi (2018)</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
</tr>
<tr>
<td>Plank (2018)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Citation</td>
<td>Coref</td>
<td>L.G</td>
<td>S.G</td>
<td>Eng</td>
<td>L≠S</td>
<td>0/1</td>
<td>Imm</td>
<td>Neo</td>
</tr>
<tr>
<td>----------------------------------</td>
<td>-------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Wood-Doughty et al. (2018)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Moorthy et al. (2018)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>Y</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Levitan et al. (2018)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Webster et al. (2018)</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Park et al. (2018)</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Vanmassenhove et al. (2018)</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
</tr>
<tr>
<td>Kleinberg et al. (2018)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Zhao et al. (2018b)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Balusu et al. (2018)</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Rudinger et al. (2018)</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>Y</td>
</tr>
<tr>
<td>Zhao et al. (2018a)</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Kiritchenko and Mohammad (2018)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Barbieri and Camacho-Collados (2018)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>Y</td>
<td>N</td>
<td>-</td>
</tr>
<tr>
<td>van der Goot et al. (2018)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
</tr>
<tr>
<td>Karlekar et al. (2018)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>de Gibert et al. (2018)</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Mickus et al. (2019)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Dana Alix Zzyym 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 v. Pompeo.

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

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

Legal case
Zzyym 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.
Despite dreading their first true series of final exams, Crona, 's relieved to have a particularly absorptive memory, lucky to recall all the material they'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 arrived at the essay portion. Thankful it was easy enough to answer, however, their subtle eye-roll stemmed entirely from just how much writing it asked of them, 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 neck, wrist, back, and ribs ached from the strain of their typical hunched posture—a habit they defaulted to, and Miss Marie, silently wished they'd be more mindful of. It was a relief, at least to them, not to be the last one out of the lecture hall. Booklet turned in, they left the room as quietly as possible and lingered just outside, an air of hesitancy settling upon them, as they 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 it?" Evans, voice echoes in the arched hall and Crona, 's shoulders jump, their frame still a tense and anxious mess.

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

Smiling with the assumed purpose of making Soul comfortable with the interaction, A defense mechanism, "I... I guess, for a final, it was easier than I expected... everyone... made it sound like it'd be difficult."

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

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

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

"That... makes sense."

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

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

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

"Oh, huh, Stein, and Marie...? Nice," his brows lift, clearly some varying degree of happy for the other. The optimism is short-lived, observing as Crona, 's expression falls back to its characteristic expressionless gaze.

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

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

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

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

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

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

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

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

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

"But... meh. Different strokes," if he can't judge.

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

D Example GICoref Document from AO3: Scar Tissue

[[Source: https://archiveofourown.org/works/14476524]]

[[Author: cornheck]]
With another encouraging smile, rows of sharpened incisors appearing oddly charismatic, he opens mouth to speak – but finds cut off before he can even squeeze a word in.

“Soul, I’m sorry,” the meister, blurs.

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

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

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

“For what...?”

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

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

Soul seems stunned for a moment before harnessing quick wit.

“Hey, now. You can’t take all the credit like that. Ragnarok did most of the damage,” he...