Toward Gender-Inclusive Coreference Resolution

An Analysis of Gender and Bias throughout the Machine Learning Lifecycle

Yang Trista Cao**
University of Maryland

Hal Daumé III†
Microsoft Research
University of Maryland

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

1. Introduction

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

* This study extends the work in Cao and Daumé III (2020); the specific additions are highlighted in the bulleted list toward the end of the introduction.
** E-mail: ycao95@umd.edu
† E-mail: me@hal3.name

Submission received: 23 November 2020
Revised version received: 9 March 2021
Accepted for publication: 22 June 2021

© 2021 Association for Computational Linguistics
stereotypes when resolving he and she pronouns (see §2). Contemporaneously, critical work in Human-Computer Interaction has complicated discussions around gender in other fields, such as computer vision (Keyes 2018; Hamidi, Scheuerman, and Branham 2018).

Building on both lines of work, and inspired by Keyes’ (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 non-trans people. We take the normative position that while the folk model of gender is even today widespread, building systems that adhere to it implicitly or explicitly can lead to significant harms to binary and non-binary trans individuals, and that we should aim to understand and minimize those harms. We take this as particularly important not just from the perspective of potentially improving the quality of our systems when run on documents by or about trans people (as well as documents by or about non-trans people), but more pointedly to minimize the harms caused by our systems by reinforcing existing unjust social hierarchies (Lambert and Packer 2019).

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

- **Indirect: subject of query.** If a person is the subject of a web query, relevant web pages about xem may be downranked if “multiple references to the same person” is an important feature in ranking and the coreference system cannot recognize and resolve xyr pronouns. This can lead to quality of service and erasure harms.

- **Direct: by choice.** If a grammar checker uses coreference features, it may insist that an author writing hir third-person autobiography is repeatedly making errors in referring to hirself. This can lead to quality of service and stereotyping (by reinforcing the stereotype that trans identities are not “real”).

- **Direct: not by choice.** If an information extraction on job applications uses a coreference system as a preprocessor, but the coreference system relies on cisnormative assumptions, then errors may disproportionately affect those who do not fit in the gender binary. This can lead to allocative harms (for hiring) as well as erasure harms.

- **Many stakeholders** If a machine translation system needs to use discourse context to generate appropriate pronouns or gendered morphological inflections in a target language, then errors can result in directly misgendering subjects of the document being translated.

To address these (and other) potential harms in more detail, as well as where and how they arise, we need to (a) complicate what “gender” means and (b) uncover how harms can enter into

---

1 Throughout, we avoid mapping pronouns to a “gender” label, preferring to use the nominative case of the pronoun directly, including (in English) she, he, the non-binary use of singular they, and neopronouns (e.g., ze/zir, 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 (or non-trans) individuals, whose assigned sex at birth happens to correspond to their gender. Transgender individuals can either be binary (those whose gender falls in the “male/female” dichotomy) or non-binary (those for which the relationship is more complex).

3 According to Clements (2018), misgendering occurs when you intentionally or unintentionally refer to a person, relate to a person, or use language to describe a person that doesn’t align with their affirmed gender.
natural language processing (NLP) systems. Toward (a), we begin with a unifying analysis (§3) of how gender is socially constructed, and how social conditions in the world impose expectations around people’s gender. Of particular interest is how gender is reflected in language, and how that both matches and potentially mismatches the way people experience their gender in the world. This reflection is highlighted, for instance, in folk notions such as an implicitly assumed one-to-one mapping between a gender and pronouns. Then, in order to understand social biases around gender, we find it necessary to consider the different ways in which gender can be realized linguistically, breaking down what previously have been considered “gendered words” in NLP papers into finer-grained categories of lexical, referential, grammatical, and social gender. Through this deconstruction (well-established in socio-linguistics), we can begin to interrogate what forms of gender stereotyping are prevalent in coreference resolution.

Toward (b), we ground our analysis by adapting Vaughan and Wallach’s (2019) framework of how a prototypical machine learning lifecycle operates\(^4\). We analyze forms of gender bias in coreference resolution in six of the eight stages of the lifecycle in Figure 1. We conduct much of our analysis around task definition (§4.1), bias in underlying text (§4.2), model definition (§4.5), and evaluation methodologies (§4.6) by evaluating prior coreference datasets, their corresponding annotation guidelines, and through a critical read of “gender” discussions in natural language processing papers. For our analysis of bias in annotations due to annotator positionality (§4.3), and our analysis of model definition (§4.5), we construct two new coreference datasets: MAP (a similar dataset to GAP (Webster et al. 2018) but without binary gender constraints on which we can perform counterfactual manipulations; see §4.2) and GICoref (a fully annotated coreference resolution dataset annotated on written by and/or about trans people; see §4.5.3).\(^5\) In all cases, we focus largely on harms due to over- and under-representation (Kay, Matuszek, and Munson 2015), replicating stereotypes (Sweeney 2013; Caliskan, Bryson, and Narayanan 2017) (particular those

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

\(^5\) Both datasets are released under a BSD license at github.com/TristaCao/into_inclusivecoref with corresponding datasheets (Gebru et al. 2018).
that are cisnormative and/or heteronormative), and quality of service differentials (Buolamwini and Gebru 2018).

The primary contributions of this paper are: 6.

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

We conclude (§5) with a discussion of how the natural language processing community can move forward in this task in particular, and also how this case study can be generalized to other language settings. Our goal is to highlight issues in previous instantiations of coreference resolution in order to improve tomorrow’s instantiations, continuing the lifecycle of coreference resolutions’ various task definition updates from MUC7 in 2001 through ACE in the mid 2000s and up to today.

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

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

2. Other Related Work

There are four recent papers that consider gender bias in coreference resolution systems. Rudinger et al. (2018) evaluates coreference systems for evidence of occupational stereotyping, by

---

6 As said above, this work is an extension of our work published in ACL 2020 (Cao and Daumé III 2020). Contributions with ◯ are published in Cao and Daumé III (2020) and contributions with ● are the new contributions of this paper. Note that the analysis of gender concepts in the last contribution is an extended version of the analysis in the ACL paper.
constructing Winograd-esque (Levesque, Davis, and Morgenstern 2012) test examples. They find that humans can reliably resolve these examples, but systems largely fail at them, typically in a gender-stereotypical way. In contemporaneous work, Zhao et al. (2018a) proposed a very similar, also Winograd-esque scheme, also for measuring gender-based occupational stereotypes. In addition to reaching similar conclusions to Rudinger et al. (2018), this work also used a similar “counterfactual” data process as we use in §4.3.1 in order to provide additional training data to a coreference resolution system. Webster et al. (2018) produced the GAP 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 modeling coreference resolution linguistically, taking a trans-inclusive perspective on gender.

Gender bias in NLP has been considered more broadly than just in coreference resolution, including, for instance, natural language inference (Rudinger, May, and Van Durme 2017), word embeddings (e.g., Bolukbasi et al. 2016; Romanov et al. 2019; Gonen and Goldberg 2019), sentiment analysis (Kiritchenko and Mohammad 2018), machine translation (Font and Costa-jussà 2019; Prates, Avelar, and Lamb 2019), among many others (Blodgett et al. 2020, inter alia). Gender is also an object of study in gender recognition systems (Hamidi, Scheuerman, and Branham 2018). Much of this work has focused on gender bias with a (usually implicit) binary lens, an issue which was also called out recently by Larson (2017).

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

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

---

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

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

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

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

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

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

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

3.1 Sociological Gender

Many modern trans-inclusive models of gender recognize that gender encompasses many different aspects. These aspects include the experience that one has of gender (or lack thereof), the way that one expresses one’s gender to the world, and the way that normative social conditions impose gender norms, typically as a dichotomy between masculine and feminine roles or traits (Kramarae and Treichler 1985). The latter two notions are captured by the “doing gender” model from social constructionism, which views gender as something that one does and to which one is socially accountable (West and Zimmerman 1987; Butler 1989; Risman 2009). However, viewing gender purely through the lens of expression and accountability does not capture the first aspect: one’s experience of one’s own gender (Serano 2007).
Such trans-inclusive views deconflate anatomical and biological traits and the sex that a person had assigned to them at birth from one’s gendered position in society; this includes intersex people, whose anatomical/biological factors do not match the usual designational criteria for either sex. Trans-inclusive views further typically recognize that gender exists beyond the regressive “female”/“male” binary; additionally, that one’s gender may shift by time or context (often “genderfluid”), and some people do not experience gender at all (often “agender”) (Kessler and McKenna 1978; Schilt and Westbrook 2009; Darwin 2017; Richards, Bouman, and Barker 2017). These models of gender contrast with “folk” views (that are prevalent both in linguistics, sociology, and science more broadly, as well as many societies at large) which assume that ones gender is defined by one’s anatomy (and/or chromosomes), that gender is binary between “male” and “female,” and that one’s gender is immutable... all of which are inconsistent with reality as it has been known for at least two thousand years. In §4.1 we will analyze the degree to which NLP papers make assumptions that are trans-inclusive or trans-exclusive.

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

3.2 Linguistic Gender

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

Grammatical gender, similarly defined in Ackerman (2019), is nothing more than a classification of nouns based on a principle of grammatical agreement. It is useful to distinguish between “gender languages” and “noun class languages.” The former have two or three grammatical genders that have, for animate or personal references, considerable correspondence between a fem (resp. masc) grammatical gender and referents with female- (resp. male-) social gender. In comparison, “noun class languages” have no such correspondence, and typically have many more gender classes. Some languages have no grammatical gender at all; English is generally seen as one (viewing that

8 Some authors use female/male for sex and woman/man for gender; we do not need this distinction (which is itself contestable) and use female/male for gender.
9 As identified by Keeyes (2018), references appear as early as CE 189 in the Mishnah (HaNasi 189). Similar references (with various interpretations) also appear in the Kamat Sutra (Barton 1883, Chapter IX), which dates sometime between BCE 400 and CE 300. Archaeological and linguistic evidence also depicts the lives of trans individuals around 500 BCE in North America (Bruhns 2006) and around 2000 BCE in Assyria (Neill 2008).
11 One difficulty in this discussion is that linguistic gender and social gender use the terms “feminine” and “masculine” differently; to avoid confusion, when referring to the linguistic properties, we use fem and masc.
Computational Linguistics

referential agreement of personal pronouns does not count as a form of grammatical agreement, a view which we follow, but one that is contested (Nissen 2002; Baron 1971; Bjorkman 2017)).

**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 they12, but also includes cases like “policeman” when the intended referent of this noun has social gender “male”. (Note that this is different from the case when “policeman” is used exclusively non-referentially, as in “every policeman needs to hold others accountable”, in which setting this is a case of lexical gender, as follows).

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

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

### 3.3 Interplays between Social and Linguistic Gender

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

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

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

---

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

13 Dahl (2000) provides several several complications of this analysis.
“opposite” genders for “X” and their spouse (the inference of “opposite” additionally implies a gender binary assumption).

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

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

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

### 4. Sources of Bias

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

#### 4.1 Bias in: Task Definition

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

For the purposes of this study, we consider here (and elsewhere in this paper) thirteen datasets on which coreference or anaphora are annotated in English (Table 1); eleven of these are corpora distributed by the Linguistic Data Consortium (LDC)\(^\text{15}\), and two are not. According to the authors of the QB dataset (personal communication), it was annotated under the OntoNotes guidelines, with the exception that singleton mentions were also annotated. The final dataset, GAP, did not explicitly annotate full coreference, but rather annotated a binary choice of which of two names a pronoun refers to (as described in the associated paper). None of the annotation guidelines (or papers) give explicit guidance about what personal pronouns are to be considered (with the exception of GAP which explicitly limits to he/she pronouns), or otherwise what information a

---

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

\(^{15}\) See https://catalog.ldc.upenn.edu/{LDC-ID}.
To assess task definition bias, we count, for each of the annotation guidelines, how many examples use different pronominal forms. For examples that use they, we separate four different subtypes, following Conrod’s (2018b) categorization (see §3.2):

- **NH**: Plural non-human group – “The knives are put away in their carrier.”
- **PL**: Plural group of humans – “The children are friendly, and they are happy.”
- **QI**: Quantified/indefinite – “Most chefs harshly critique their own dishes.”
- **SP**: Specific singular referent – “Jun enjoys teaching their students.”

The results are shown in Table 2 (datasets for which no relevant examples were provided are not listed). Overall, we see that in total across these seven datasets, examples with he occur more than twice as frequently as all others combined. Furthermore, they is never used in a specific setting and, somewhat interestingly, is only used as an example for quantification in older datasets.
Figure 2: (Top) For each dataset under consideration, the fraction of pronouns with different forms. Only our dataset (GICoref) and the genderqueer subreddit include neopronouns. (In the case of GAP, there are some occurrences of they, but they are never considered targets for coreference and so we exclude them from these counts.) (Bottom) For three datasets, the fraction (out of 100 annotated) of each they into one of the four usage cases.

(2005 and before). Moreover, none of the annotation guidelines have examples using neopronouns. This lack does nothing to counterbalance a general societal bias that tends to erase non-binary identities. In the case of GAP, it is explicitly mentioned that only she and he examples are considered (and only in cases where the gender of two possible referents “matches”—though it is unspecified what type of gender this is and how it is determined). Even on the binary spectrum, there is also an obvious gender bias between he and she examples.

Tasks defined in these thirteen datasets only consider binary gender and are mostly male-dominated. Systems built along with such task definitions can hardly function for non-binary and female users. See §4.6 for detailed analysis of system performances on data with both binary and non-binary pronouns.

4.2 Bias in: Data Input

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

To analyze the possible impact of input data, we consider our thirteen coreference datasets, and count how many instances of different types of pronouns are used in the raw data. We focus on she, he, and they pronouns (in all their morphological forms); we additionally counted several neopronoun forms (hir, xey and ey) and found no occurrences nor their morphological variances.16

In the case of they, we again distinguish between its four usage cases: plural, singular, quantified, and non-human. To achieve this, we annotated 100 examples uniformly at random by hand from OntoNotes (Weischedel et al. 2011). Furthermore, we compare to the raw counts in a 2015 dump from some of Reddit discussion forum17, and also limited to the genderqueer subforum. We additionally include a new dataset for this study, GICoref. This new dataset is collected to evaluate current coreference resolution systems on gender-inclusive and naturally-occuring texts. Details if the dataset are described in §4.5.3.

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

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

4.3 Bias in: Data Annotation

A significant possible source of bias comes from annotations themselves, arising from a combination of (possibly) underspecified annotations guidelines and the positionality of annotators themselves. Ackerman (2019) analyzes how human cognitively encode gender in resolving coreferences through a Broad Matching Criterion, which posits “matching gender requires at least one level of the mental representation of gender [e.g. the occupation of the person or the environment context, etc.] to be identical to the candidate antecedent in order to match.” In this section, we delve into the linguistic notions of gender and study how different aspects of linguistic notions impact an annotator’s judgments of anaphora.

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

As motivation, consider (1) below, in which an annotator is likely to determine that “her” refers to “Mary” and not “John” due to assumptions on likely ways that names may map to

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

17 It is a dataset with publicly available comments from Reddit. The dataset has about 1.7 billion comments with their related fields, such as score, author, subreddits, etc. Here is the link to the dataset www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/
pronouns (or possibly by not considering that she pronouns could refer to someone named “John”). While in (2), an annotator is likely to have difficulty making a determination because both “Sue” and “Mary” suggest “her”. In (3), an annotator lacking knowledge of name stereotypes on typical Chinese and Indian names (plus the fact that given names in Chinese—especially when romanized—generally do not signal gender strongly), respectively, will likewise have difficulty.

1. John and Mary visited her mother.

2. Sue and Mary visited her mother.

3. Liang and Aditya visited her mother.

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

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

4. My brother and niece visited her mother.

5. Mr. Hashimoto and Mrs. Iwu visited her mother.

6. المطربة والمثلية شاهدة والدتي
   walidatu-ha shahadaa al-momathela w almutarab
   mother her saw actor[masc] and singer[masc]
   The singer[masc] and actor[masc] saw her mother.

7. The nurse and the actor visited her mother.

4.3.1 Ablation Methodology. In order to determine which cues annotators are using and the degree to which they use them, we construct an ablation study in which we hide various aspects of gender and evaluate how this impacts annotators’ judgments of anaphoricity. To make the task easier for crowdsourcing, we follow the methodology of Webster et al.’s (2018) GAP dataset for studying ambiguous binary gendered pronouns. In particular, we construct binary classification examples taken from Wikipedia pages, in which a single pronoun is selected, and two possible antecedent names are given, and the annotator must select which one. We cannot use the GAP dataset directly, because their data is constrained that the “gender” of the two possible antecedents is “the same”\(^\text{18}\); 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.

\[^{18}\text{It is unclear from the GAP dataset what notion of “gender” is used, nor how it was determined to be “the same.”}\]
Mrs. (d) Rebekah Johnson Bobbitt was the younger (c) sister of (b) M. Booth. Lyndon B. Johnson (b) T. Schneider, 36th President of the United States, was born in 1910 in Stonewall, Texas. (a) They worked in the cataloging department of the Library of Congress in the 1930s before (c) their brother (d) sibling entered politics.

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

Pronoun (MAP), which is similar to the GAP dataset, but where we do not restrict the two names to match gender so that we can measure the influence of different gender cues.

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

(8) Liang saw herself in the mirror... He...

Fortunately, it is possible to enumerate a high coverage list of English terms that signal lexical gender: terms of address (Mrs., Mr.) and semantically gendered nouns (mother).

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 all third person singular pronouns with a gender neutral program (they, xey, ze).

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

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

(d) ¬ Addr: Remove all terms of address (e.g., “Mrs.”, “Sir”).

See Figure 3 for an example of all substitutions.

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

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

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

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

In the backward selection case, we begin with Orig, which is the unmodified original text. To this, we can apply the pronoun filter to get ¬Pro; differences in annotation between Orig and ¬Pro give a measure of how much any sort of gender-based inference is used. Similarly, we 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.
4.4 Annotation Results

We construct examples using the methodology defined above. We then conduct annotation experiments using crowdworkers on Amazon Mechanical Turk following the methodology by which the original GAP corpus was created. 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 4 shows the human annotation results as binary classification accuracy for resolving the pronoun to the antecedent. We can see that removing pronouns leads to 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 ORK 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’ confidence follow the same trend as the accuracy: annotators have a reasonable sense of when they are unsure. We also note that accuracy score essentially are the same for Zero and ~Poo, 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.

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

4.4.1 Limitations of (approximate) counterfactual text manipulation. Any text manipulation—like we have done in this section—runs the risk of missing out on how a human author might truly have written that text under the presumed counterfactual. For example, a speaker uttering (1) may assume that the interlocutor shares, or at least recognizes, social biases that lead one to assume that the person named “John” is likely referred to as he and “Mary” as she. This speaker may use this assumption of the listener to determine that “her” is sufficiently unambiguous in this case as to be an acceptable reference (trading off brevity and specificity; see, for instance (Arnold 2008; Frank and Goodman 2012; Orita et al. 2015a)). However, if we “counterfactually” replaced the names “John” and “Mary” to “H. Martinez” and “R. Modi” (respectively), it is unlikely that the supposed speaker would make the same decision. In this case, the speaker may well have said “Modi’s mother” or some other reference that would have

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

22 In some of the examples, a worker may apply knowledge of the situation or entities involved, for instance “President of the United States” has, to date, always been referred to using he pronouns. To capture this, we additionally asked the crowdworkers if they recognized any of the entities or the situation involved. Note, however, that even though this is true in the real world, it is not hard to imagine fictional contexts in which a human annotator would have no difficulty finding “President of the United States” to corefer with she or they pronouns.
Table 3: Analysis of a corpus of 150 NLP papers that mention “gender” along the lines of what assumptions around gender are implicitly or explicitly made in the work.

4.5 Bias in: Model Definition

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

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

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

- Does the paper discuss coreference or anaphora resolution?
- Does the paper study English (and possibly other languages)?
- Does the paper deal with linguistic gender (i.e., grammatical gender or gendered pronouns)?
- Does the paper deal with social gender?
- (If yes to the previous two:) Does the paper explicitly distinguish linguistic from social gender?
- (If yes to social gender:) Does the paper explicitly recognize that social gender is not binary?
Figure 5: Coreference resolution systems results for the ablation study on MAP dataset. The y-axis is the degree of accuracy with 95% significance intervals.

- (If yes to social gender:) Does the paper explicitly or implicitly assume social gender is immutable?23
- (If yes to social gender and to English:) Does the paper explicitly consider uses of definite singular “they” or neopronouns?

The results of this coding are in Table 3 and the list of the full set of annotations is in Appendix 1. Here, we see out of the 22 coreference papers analyzed, the vast majority conform to a “folk” theory of language:
- Only 5.5% distinguish social from linguistic gender (despite it being relevant);
- Only 5.6% explicitly model gender as inclusive of non-binary identities;
- No papers treat gender as anything other than completely immutable;
- Only 7.1% (one paper) considers neopronouns and/or specific singular they.

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

23 The most common ways in which papers implicitly assume that social gender is immutable is either 1) by relying on external knowledge bases that map names to “gender”; or 2) by scraping a history of a user’s social media posts or emails and assuming that their “gender” today matches the gender of that historical record.
4.5.2 Coreference system performance on MAP. Next, we analyze the effect that our different ablation mechanisms have on existing coreference resolutions systems. In particular, we run five coreference resolution systems on our ablated data: the AI2 system (AI2; Gardner et al. 2017), hugging face (HF; Wolf 2017), which is a neural system based on spacy (Honnibal and Montani 2017), and the Stanford deterministic (SfdD; Raghunathan et al. 2010), statistical (SfdS; Clark and Manning 2015) and neural (SfdN; Clark and Manning 2016) systems. Figure 5 shows the results. We can see that the system accuracies mostly follow the same pattern as human accuracy scores, though all are significantly lower than human results. Accuracy scores for systems drop dramatically when we ablate out referential gender in pronouns. This reveals that those coreference resolution systems rely heavily on gender-based inferences. In terms of each system, HF and SfdN systems have similar results and outperform other systems in most cases. SfdD accuracy drops significantly once names are ablated.

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

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

The GICOREF dataset consists of 95 documents from three types of sources: articles on English Wikipedia about people with non-binary gender identities, articles from LGBTQ periodicals, and fan-fiction stories from Archive Of Our Own24 (with the respective author’s permission). Each author of this paper manually annotated each of these documents and then we jointly adjudicated the results25. To reduce annotation time, any article that was substantially longer than 1000 words (pre-tokenization) was trimmed at the 1000th word.26 This data includes many examples of people who use pronouns other than she or he, people who are genderfluid and whose name or pronouns changes through the article, people who are misgendered, and people in relationships that are not heteronormative. One annotation decision we made was around the specific case of people who perform drag. Following (Butler 1989), in our annotation we considered drag performance as a form of genderfluidity; as such, we annotate the performer and the drag persona as coreferent with each other (as well as the relevant pronouns), akin to how we believed a reasonable model for handling stage names (e.g., Christopher Wallace / Notorious B.I.G.) would.

24 See https://archiveofourown.org; thanks to Os Keyes for this suggestion.
25 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.”
26 There are four documents we accidentally trimmed at the 2000th word and so we keep the longer version of them with 2000 words in the dataset.
mark them as coreferent, while famous roles played by multiple actors (e.g., Carrie, played by Sissy Spacek, Angela Bettis, and Chloé Grace-Moretz), would be marked as non-coreferent. In addition, incorrect references (misgendering and deadnaming\(^\text{27}\)) are explicitly annotated.\(^\text{28}\) Two example annotated documents, one from Wikipedia, and one from Archive of Our Own, are provided in Appendix 2 and Appendix 3 respectively.

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

We run the same systems as before on this dataset. Table 4 reports results according to the standard coreference resolution evaluation metric LEA (Moosavi and Strube 2016). It is not clear how systems or evaluation metrics should handle incorrect references (misgendering and deadnaming). Taking (9) as an example, should the misgendering entities and pronouns (cluster c) be included as a coreference to the person (cluster a) or not. If the person is a real human, including the misgendering reference as a ground truth may be potentially harmful to the person. 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 (cluster a and cluster c are the same cluster); and other with incorrect references excluded (cluster a and cluster c are separate clusters). Due to the limited number of incorrect references (0.6% of total references of people) in the dataset, the difference of the results are not significant – the difference is less than 0.2% for each entry. Nonetheless, although these are rare, they constitute a significant potential harms. Here we only report the result for latter.

\(^{27}\) 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.

\(^{28}\) Thanks to an anonymous reader of a draft version of this paper for this suggestion.

<table>
<thead>
<tr>
<th>LEA</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>
<thead>
<tr>
<th>Recall</th>
<th>HE/SHE</th>
<th>THEY</th>
<th>NEO</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI2</td>
<td>96.4%</td>
<td>93.0%</td>
<td>25.4%</td>
</tr>
<tr>
<td>HF</td>
<td>95.4%</td>
<td>85.0%</td>
<td>21.1%</td>
</tr>
<tr>
<td>SfdD</td>
<td>97.6%</td>
<td>96.6%</td>
<td>0.0%</td>
</tr>
<tr>
<td>SfdS</td>
<td>96.2%</td>
<td>86.8%</td>
<td>20.4%</td>
</tr>
<tr>
<td>SfdN</td>
<td>96.6%</td>
<td>90.1%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

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

\[(9)\] Frisk, sat in the back of the classroom, silently praying that their teacher wouldn’t call on them. They were having a bad day and didn’t think they could be misgendered today. But just then, luck, their teacher, was staring straight at them. Felix, ? Do you know the answer? Chara, s hand shot up. Ms. Richards, their name is Frisk, remember? Christine, Ms. Richards, sighed, ignoring Chara’s flinch. His name is whatever is on the sheet, the same way yours is. We have had this discussion, remember?\]
The first observation is that there is still plenty room for coreference systems to improve; the best performing system achieves an F1 score of 34%, but the Stanford neural system’s F1 score on CoNLL-2012 test set reaches 60% (Moosavi 2020). Here are some examples where the huggingface and the Stanford deterministic system output erroneous resolutions:

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

**StdD:** The artwork consisted of Sulkowics, who uses **they/them pronouns**, carrying a mattress wherever **they** went on campus.

(11) **HF & StdD:** As the son of a military father, she faced many challenges to be accepted.

(12) **HF:** Soon after Vasold’s win was announced, ze spoke with the school newspaper about **zeir surprise** at winning.

**StdD:** Soon after Vasold’s win was announced, **ze** spoke with the school newspaper about **zeir surprise** at winning.

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. Results are shown in Table 4. We find that all systems achieve a recall of at least 95% for binary pronouns, a recall of around 90% on average for they, and a recall of around a paltry 13% for neopronouns (two systems—Stanford deterministic and Stanford neural—never identify any neopronouns at all).

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

4.6 Bias in: System Testing

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

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

In terms of metrics, most intrinsic evaluation is carried out using metrics like MUC (Vilain et al. 1995), ACE (Mitchell et al. 2005), B3 (Bagga and Baldwin 1998; Stoyanov et al. 2009), or CEAF (Luo 2005) (see also (Cai and Strube 2010) for additional discussion and variants). As observed by Agarwal et al. (2019), most of these metrics are rather insensitive to arguably large errors, like the inability to link pronouns to names; to address this, they introduce a new metric to focus specifically on this named entity coreference task. These metrics also generally treat all errors similarly, regardless of whether the error compounds societal injustices (e.g., ignoring an instance of *xyr*) or not (e.g., ignore an instance of *he*), despite the fact that these have vastly different implications from the perspective of justice (Fraser 2008, e.g.).

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

4.7 Bias in: Feedback Loops

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

The most straightforward way in which this can happen is through coreference resolution systems that engage in statistically biased active learning (or bootstrapping) techniques.29 Active learning for coreference has been popular since the early 2000s, perhaps largely because coreference annotation is quite costly. Considering the approaches used in dominant papers, the active learning algorithms used are not statistically unbiased (Ng and Cardie 2003; Laws, Heimerl, and Schütze 2012; Miller, Dligach, and Savova 2012; Sachan, Hovy, and Xing 2015; Guha et al. 2015).

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

More frequent is the almost ubiquitous use of “name lists” that map names (either full names or simply given names) to “gender.” And the most frequently used of these is the resource developed by Bergsma and Lin (2006) (henceforth, B+L), in which a large quantity of text was

---

29 Here, the overloading of “bias” is unfortunate. By “statistically biased,” we mean bias in the technical sense that of a learning system that even in the limit of infinite data does not infer the optimal model, a fundamentally different concept from the sort of “bias” we consider in the rest of this paper.
processed with “high precision” anaphora resolution links to associate names with “genders.” The process specifically mapped names to pronouns, from which gender (presumably an approximation of referential gender) was inferred. This leads to a resource that pairs a full name or name substring (like “Bill Clinton”) counts for identified coreference with he (8150, 97.7%), she (70, 0.8%), it (42, 0.5%) and they (82, 1%); these are referred to, respectively, as “male”, “female”, “neuter” and “plural,” and seemingly largely used as such in work that leverages this resource. We focus on this resource only because it has become ubiquitous, both in coreference resolution and in gender analysis in NLP more broadly.

The first question we ask is: what happens when this “gender” inference data is used to infer the gender of prominent non-binary individuals. To this end, we took the names of 104 non-binary people referenced on Wikipedia⁴⁰ and queried the B+L data with them. In almost all cases, the full name was unknown in the B+L data (or had counts less than five), and in such cases we backed off to simply querying on the given name. We cross-tabulated the correct (according to Wikipedia) pronouns for these 90 people with the “gender” inferred by the B+L data.

The results are shown in Table 5, where we can see that of those who use a pronoun other than she or he (exclusively) are, essentially always, misgendered. Even on binary pronouns she and he, the accuracy is only 50%. For the case of people who use they pronouns, one might ask what the ideal behavior would be given the framing of this resource—given that “Plural” is interpreted to be “coreferent with ‘they’”, we might hope that (aside from the naming issue), people who

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

<table>
<thead>
<tr>
<th>Pronoun</th>
<th>Mascl</th>
<th>Feml</th>
<th>Neutl</th>
<th>Plurl</th>
<th>Unkl</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>she</td>
<td>27</td>
<td>41</td>
<td>18</td>
<td>0</td>
<td>14</td>
<td>22</td>
</tr>
<tr>
<td>he</td>
<td>64</td>
<td>7</td>
<td>7</td>
<td>0</td>
<td>22</td>
<td>14</td>
</tr>
<tr>
<td>they</td>
<td>39</td>
<td>32</td>
<td>12</td>
<td>5</td>
<td>12</td>
<td>41</td>
</tr>
<tr>
<td>ze</td>
<td>22</td>
<td>33</td>
<td>33</td>
<td>11</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>they∨she</td>
<td>25</td>
<td>50</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>s/he</td>
<td>0</td>
<td>0</td>
<td>50</td>
<td>0</td>
<td>50</td>
<td>2</td>
</tr>
<tr>
<td>they∧she</td>
<td>50</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>50</td>
<td>2</td>
</tr>
<tr>
<td>all/any</td>
<td>50</td>
<td>50</td>
<td>0</td>
<td>0</td>
<td>50</td>
<td>2</td>
</tr>
<tr>
<td>xe</td>
<td>50</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>50</td>
<td>2</td>
</tr>
<tr>
<td>ey</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>v</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ae</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ne</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ze∧she</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

use they are considered Plural: this only happens in 5% of cases, though. Expanding on this, one might hope that people who use they vs she or they as she are mapped to either Fem or Plural, but this only happens in 2/6 cases (and always to Fem in those cases). For the neopronoun cases, the manner in which the resource is constructed nearly excludes any reasonable behavior (aside from, perhaps, a “least bad” option of simply abstaining with an output—which only happens in 3/19 cases). This approach actively misgenders individuals, is harmful, and demonstrates that assigning gender to “names” does not work: anybody can have any combination of names and pronouns.

5. Discussion and Moving Forward

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

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

Moreover, we realized that both human and coreference systems rely heavily on gender cues in resolving coreferences. Though it is natural for human, we want to emphasize that both human and systems should not overrely on the risky cues such as names, semantically gendered nouns, and terms of address, compared to relatively safe cues like syntax. In annotating the dataset, we only had about three ambiguous coreferences where both annotators agreed either reference was possible, thus demonstrating that people are able to resolve coreferences without relying extensively on the riskier cues. One cue that we explored in detail is that around names, and it is worth pointing out recent work by Agarwal et al. (2020) in the context of named entity recognition. In this paper, the authors found that state of the art systems perform poorly on documents from non-U.S. contexts, due in large part to systems’ unfamiliarity with non-Western names. We expect similar results would hold in the coreference case, where it would be particularly interesting to evaluation in the context of name-gender lists.

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

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

More broadly than in coreference resolution, we found that natural language processing papers also tend to make strong, binary assumptions around gender (typically implicitly), a practice that we hope to see change in the future. In more recent papers, we begin to see footnotes that acknowledge that the discussion omits questions around trans or non-binary, issues. We hope to see these be promoted from footnotes to objects of study in future work; mentioning the existence of non-binary people in a footnote does little to minimize the harms a system may cause them. Much inspiration here may come from third wave feminism and queer theory (De Lauretis 1990; Jagose 1996), and perhaps more closely the recent movement within Human Computer Interaction toward Queering HCI (Light 2011) and Feminist HCI (Bardzell and Churchill 2011). The goal that queer theory has of deconstructing social norms and associated taxonomies is particularly important as NLP technology addresses more and more socially relevant issues, including but not limited to issues around gender, sex and sexuality.

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

Acknowledgments

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

References


Ackerman, Lauren. 2019. Syntactic and cognitive issues in investigating gendered coreference.
Cao & Daumé

Toward Gender-Inclusive Coreference Resolution


Cao, Yang Trista and Hal Daumé III. 2020.


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

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


Cao & Daumé Toward Gender-Inclusive Coreference Resolution

Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media, pages 69–75, Association for Computational Linguistics, New Orleans, Louisiana, USA.


Cao & Daumé  

Toward Gender-Inclusive Coreference Resolution


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


feminist dictionary. Pandora Press.


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


Levitan, Sarah Ita, Angel Maredia, and Julia Hirschberg. 2018. Linguistic cues to deception and perceived deception in interview dialogues.


Cao & Daumé

Toward Gender-Inclusive Coreference Resolution


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


Cao & Daumé

Toward Gender-Inclusive Coreference Resolution


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


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


Schwartz, H. Andrew, Gregory Park, Maarten Sap,
Cao & Daumé


Tiedemann, Jörg. 2003. Building parallel text from the common crawl. In ACL.


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

Association for Computational Linguistics, Austin, Texas.


Walsh, Bill. 2015. The post drops the ‘mike’—and the hyphen in ‘e-mail’. Washington Post; permalink.


### 1. Annotation of ACL Anthology Papers

Below we list the complete set of annotations we did of the papers described in ??.

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.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.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>-</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, Francopolou, and Laroui (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>Popovich (1989)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</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, Gillick, and Roth (1994)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Genthial, Courtin, and Menez (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, Orman, and Itai (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, Kudo, and Plátek (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>-</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, McRoy, and Ali (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, Alsamaa, and Evens (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, Buntrock, and Chute (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>Navaretta (2004)</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>-</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>-</td>
<td>-</td>
</tr>
<tr>
<td>Mota, Carvalho, and Ranchhod (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>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>Smith, Smith, and Tromble (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>Streiter, Voltmer, and Goudin (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, Kambhatla, and Roukos (2007)</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>-</td>
</tr>
<tr>
<td>Badr, Zbib, and Glass (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, Zbib, and Glass (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>Y</td>
<td>N</td>
</tr>
<tr>
<td>Bergsma, Lin, and Goebel (2009)</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>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>Y</td>
<td>N</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>Y</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>Burkhart 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, Habash, and Rambow (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, Ghosh, and Srihari (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, Gajulapalli, and Choi (2011)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Li, Miller, and Schuler (2011)</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>-</td>
</tr>
<tr>
<td>Burger et al. (2011)</td>
<td>N</td>
<td>N</td>
<td>N</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, Padró, and Turmo (2011)</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>-</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>Mareček 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, Koleva, and Krieger (2012)</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Bergsma, Post, and Yarowsky (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>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Dinu, Niculae, and Şulea (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>-</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, Habash, and Rambow (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, Fraser, and Schulte im Walde (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>Cioc, Sonderegger, and Bihs (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, Wilson, and Yarowsky (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, Rosa, and Tamchyna (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>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>Glavaš, Korenčić, and Šnajder (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>N</td>
<td>N</td>
</tr>
<tr>
<td>Sap 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>-</td>
</tr>
<tr>
<td>Nguyen 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>Prabhakaran, Reid, and Rambow (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>Sidorov, Ultes, and Schmitt (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, Abdelali, and Mubarak (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>Ahmed Khan (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>Nguyen, Trieschnigg, and Meder (2014)</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>Stewart (2014)</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>Matthews 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>Vaidya, Rambow, and Palmer (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>Kokkinakis, Ighe, and Malm (2015)</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Johannsen, Hovy, and Søgaard (2015)</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>Schwartz et al. (2015)</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</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>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Preoţiuc-Pietro et al. (2015)</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Ramakrishna et al. (2015)</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</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>N</td>
<td>Y</td>
<td>Y</td>
<td>-</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>-</td>
</tr>
<tr>
<td>Qian, Qiu, and Huang (2016)</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</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>Y</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>-</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, Paredes Palacios, and Rosso (2017)</td>
<td>N</td>
<td>Y</td>
<td>N</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>Mandravickaitė and Krišliavičius (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>Verhoeven, Škrjanec, and Pollak (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 (2017)</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>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>Ljubešić, Fišer, and Erjavec (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>-</td>
</tr>
<tr>
<td>Kraus 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>-</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>Wood-Doughty 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>Moorhy 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, Maredia, and Hirschberg (2018)</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>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, Shin, and Fung (2018)</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>-</td>
</tr>
<tr>
<td>Vanmassenhove, Hardmeier, and Way (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, Mozes, and van der Veg (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, Merghani, and Eisenstein (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>-</td>
<td>Y</td>
<td>-</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>Karlekari, Niu, and Bansal (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>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Mickus, Bonami, and Paperno (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>

43
2. Example GICoref Document from Wikipedia: Dana Zzyym

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

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.
3. Example G1Coref Document from AO3: Scar Tissue

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

Despite dreading their first true series of final exams, Crona, is relieved to have a particularly absorbative 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 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'd, 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 hesitance 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, wasn't it?
"Evans... voice echoes in the arched hall and Crona's shoulders jump, their facade falls just in time to be replaced by a much more genuine grin.
"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.
"If, 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, I... does it really do well on 'em... bad test-taker, n't it.

"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, yes, 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 practicly studies in her sleep. I'm convinced she must be practicing clairvoyance the way she burns 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 think so, they begin, stuttering, but encouraged to continue by a cock of the other'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, observed 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.

"You 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 poignent 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 - lippped frown )

"They'd probably love to have you. I bet they'd love to have 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 more effort into taking care of himself, if he had someone else besides Marie, to look after.

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

But... meh, Different strokes, I can't judge.

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

"Soul, I'm sorry," the meister blurts. Having been pent-up for months, the apology comes forth without inhibition, rolling effortlessly off their tongue.

"Sorry ... For what?" Evans quirks a brow, chuckling.

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

"F - for what ...?"

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

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

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

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