Gender Bias in Machine Translation

Beatrice Savoldi[1,2], **Marco Gaido**[1,2], **Luisa Bentivogli**[2], **Matteo Negri**[2], **Marco Turchi**[2]
[1]University of Trento, Italy
[2]Fondazione Bruno Kessler, Italy
{bsavoldi,mgaido,bentivo,negri,turchi}@fbk.eu

Abstract

Machine translation (MT) technology has facilitated our daily tasks by providing accessible shortcuts for gathering, processing, and communicating information. However, it can suffer from biases that harm users and society at large. As a relatively new field of inquiry, studies of gender bias in MT still lack cohesion. This advocates for a unified framework to ease future research. To this end, we: *i)* critically review current conceptualizations of bias in light of theoretical insights from related disciplines, *ii)* summarize previous analyses aimed at assessing gender bias in MT, *iii)* discuss the mitigating strategies proposed so far, and *iv)* point toward potential directions for future work.

1 Introduction

Interest in understanding, assessing, and mitigating gender bias is steadily growing within the natural language processing (NLP) community, with recent studies showing how gender disparities affect language technologies. Sometimes, for example, coreference resolution systems fail to recognize women doctors (Zhao et al., 2017; Rudinger et al., 2018), image captioning models do not detect women sitting next to a computer (Hendricks et al., 2018), and automatic speech recognition works better with male voices (Tatman, 2017). Despite a prior disregard for such phenomena within research agendas (Cislak et al., 2018), it is now widely recognized that NLP tools encode and reflect controversial social asymmetries for many seemingly neutral tasks, machine translation (MT) included. Admittedly, the problem is not new (Frank et al., 2004). A few years ago, Schiebinger (2014) criticized the phenomenon of ''masculine default'' in MT after running one of her interviews through a commercial translation system. In spite of several feminine mentions in the text, she was repeatedly referred to by masculine pronouns. Gender-related concerns have also been voiced by online MT users, who noticed how commercial systems entrench social gender expectations, for example, translating engineers as masculine and nurses as feminine (Olson, 2018).

With language technologies entering widespread use and being deployed at a massive scale, their societal impact has raised concern both within (Hovy and Spruit, 2016; Bender et al., 2021) and outside (Dastin, 2018) the scientific community. To take stock of the situation, Sun et al. (2019) reviewed NLP studies on the topic. However, their survey is based on monolingual applications, whose underlying assumptions and solutions may not be directly applicable to languages other than English (Zhou et al., 2019; Zhao et al., 2020; Takeshita et al., 2020) and cross-lingual settings. Moreover, MT is a multifaceted task, which requires resolving multiple gender-related subtasks at the same time (e.g., coreference resolution, named entity recognition). Hence, depending on the languages involved and the factors accounted for, gender bias has been conceptualized differently across studies. To date, gender bias in MT has been tackled by means of a narrow, problem-solving oriented approach. While technical countermeasures are needed, failing to adopt a wider perspective and engage with related literature outside of NLP can be detrimental to the advancement of the field (Blodgett et al., 2020).

In this paper, we intend to put such literature to use for the study of gender bias in MT. We go beyond surveys restricted to monolingual NLP (Sun et al., 2019) or that are more limited in scope (Costa-jussà, 2019; Monti, 2020), and present the first comprehensive review of gender bias in MT. In particular, we **1)** offer a unified framework that introduces the concepts, sources, and effects of bias in MT, clarified in light of

Transactions of the Association for Computational Linguistics, vol. 9, pp. 845–874, 2021. https://doi.org/10.1162/tacl_a_00401
Action Editor: Emily M. Bender. Submission batch: 3/2021; Revision batch: 4/2021; Published 8/2021.

relevant notions on the relation between gender and different languages; and **2)** critically discuss the state of the research by identifying blind spots and key challenges.

2 Bias Statement

Bias is a fraught term with partially overlapping, or even competing, definitions (Campolo et al., 2017). In cognitive science, bias refers to the possible outcome of heuristics, that is, mental shortcuts that can be critical to support prompt reactions (Tversky and Kahneman, 1973, 1974). AI research borrowed from such a tradition (Rich and Gureckis, 2019; Rahwan et al., 2019) and conceived bias as the divergence from an ideal or expected value (Glymour and Herington, 2019; Shah et al., 2020), which can occur if models rely on spurious cues and unintended shortcut strategies to predict outputs (Schuster et al., 2019; McCoy et al., 2019; Geirhos et al., 2020). Since this can lead to systematic errors and/or adverse social effects, bias investigation is not only a scientific and technical endeavor but also an ethical one, given the growing societal role of NLP applications (Bender and Friedman, 2018). As Blodgett et al. (2020) recently called out, and has been endorsed in other venues (Hardmeier et al., 2021), analyzing bias is an inherently normative process that requires identifying *what* is deemed as harmful behavior, *how*, and to *whom*. Here, we stress a human-centered, sociolinguistically motivated framing of bias. By drawing on the definition by Friedman and Nissenbaum (1996), we consider as biased an MT model that *systematically* and *unfairly* discriminates against certain individuals or groups in favor of others. We identify bias per a specific model's behaviors, which are assessed by envisaging their potential risks when the model is deployed (Bender et al., 2021) and the harms that could ensue (Crawford, 2017), with people in focus (Bender, 2019). Since MT systems are used daily by millions of individuals, they could impact a wide array of people in different ways.

As a guide, we rely on Crawford (2017), who defines two main categories of harms produced by a biased system: *i)* **Representational** harms (R) (i.e., detraction from the representation of social groups and their identity, which, in turn, affects attitudes and beliefs); and *ii)* **Allocational** harms (A) (i.e., a system allocates or withholds opportunities or resources to certain groups). Considering the so-far reported real-world instances of gender bias (Schiebinger, 2014; Olson, 2018) and those addressed in the MT literature reviewed in this paper, **(R)** can be further distinguished into *under-representation* and *stereotyping*.

Under-representation refers to the reduction of the visibility of certain social groups through language by *i)* producing a disproportionately low representation of women (e.g., most feminine entities in a text are misrepresented as male in translation); or *ii)* not recognizing the existence of non-binary individuals (e.g., when a system does not account for gender neutral forms). For such cases, the misrepresentation occurs in the language employed to talk ''about'' such groups.[1] Also, this harm can imply the reduced visibility of the language used ''by'' speakers of such groups by *iii)* failing to reflect their identity and communicative repertoires. In these cases, an MT flattens their communication and produces an output that indexes unwanted gender identities and social meanings (e.g., women and non-binary speakers are not referred to by their preferred linguistic expressions of gender).

Stereotyping regards the propagation of negative generalizations of a social group, for example, belittling feminine representation to less prestigious occupations (teacher (Feminine) vs. lecturer (Masculine)), or in association with attractiveness judgments (pretty lecturer (Feminine)).

Such behaviors are harmful as they can directly affect the self-esteem of members of the target group (Bourguignon et al., 2015). Additionally, they can propagate to indirect stakeholders. For instance, if a system fosters the visibility of the way of speaking of the dominant group, MT users can presume that such a language represents the most appropriate or prestigious variant[2]—at the expense of other groups and communicative repertoires. These harms can aggregate, and the ubiquitous embedding of MT in Web applications provides us with paradigmatic examples of how the two types of **(R)** can interplay. For example, if women or non-binary[3] scientists are the subjects of a query, automatically translated pages run the

[1] See also the classifications by Dinan et al. (2020).

[2] For an analogy on how technology shaped the perception of feminine voices as shrill and immature, see Tallon (2019).

[3] Throughout the paper, we use non-binary as an umbrella term for referring to all gender identities between or outside the masculine/feminine binary categories.

risk of referring to them via masculine-inflected job qualifications. Such misrepresentations can lead readers to the experience of feelings of identity invalidation (Zimman et al., 2017). Also, users may not be aware of being exposed to MT mistakes due to the deceptively fluent output of a system (Martindale and Carpuat, 2018). In the long run, stereotypical assumptions and prejudices (e.g., only men are qualified for high-level positions) will be reinforced (Levesque, 2011; Régner et al., 2019).

Regarding **(A)**, MT services are consumed by the general public and can thus be regarded as resources in their own right. Hence, **(R)** can directly imply **(A)** as a performance disparity across users in the *quality of service*, namely, the overall efficiency of the service. Accordingly, a woman attempting to translate her biography by relying on an MT system requires additional energy and time to revise incorrect masculine references. If such disparities are not accounted for, the MT field runs the risk of producing systems that prevent certain groups from fully benefiting from such technological resources.

In the following, we operationalize such categories to map studies on gender bias to their motivations and societal implications (Tables 1 and 2).

3 Understanding Bias

To confront bias in MT, it is vital to reach out to other disciplines that foregrounded how the socio-cultural notions of gender interact with language(s), translation, and implicit biases. Only then can we discuss the multiple factors that concur to encode and amplify gender inequalities in language technology. Note that, except for Saunders et al. (2020), current studies on gender bias in MT have assumed an (often implicit) binary vision of gender. As such, our discussion is largely forced into this classification. Although we also describe bimodal feminine/masculine linguistic forms and social categories, we emphasize that gender encompasses multiple biosocial elements not to be conflated with sex (Risman, 2018; Fausto-Sterling, 2019), and that some individuals do not experience gender, at all, or in binary terms (Glen and Hurrell, 2012).

3.1 Gender and Language

The relation between language and gender is not straightforward. First, the linguistic structures used to refer to the extra-linguistic reality of gender vary across languages (§3.1.1). Moreover, how gender is assigned and perceived in our verbal practices depends on contextual factors as well as assumptions about social roles, traits, and attributes (§3.1.2). Lastly, language is conceived as a tool for articulating and constructing personal identities (§3.1.3).

3.1.1 Linguistic Encoding of Gender

Drawing on linguistic work (Corbett, 1991; Craig, 1994; Comrie, 1999; Hellinger and Bußman, 2001, 2002, 2003; Corbett, 2013; Gygax et al., 2019) we describe the linguistic forms (lexical, pronominal, grammatical) that bear a relation with the extra-linguistic reality of gender. Following Stahlberg et al. (2007), we identify three language groups:

Genderless languages (e.g., Finnish, Turkish). In such languages, the gender-specific repertoire is at its minimum, only expressed for basic lexical pairs, usually kinship or address terms (e.g., in Finnish *sisko*/sister vs. *veli*/brother).

Notional gender languages[4] (e.g., Danish, English). On top of lexical gender (*mom*/*dad*), such languages display a system of pronominal gender (*she/he*, *her/him*). English also hosts some marked derivative nouns (*actor*/*actress*) and compounds (*chairman*/*chairwoman*).

Grammatical gender languages (e.g., Arabic, Spanish). In these languages, each noun pertains to a class such as masculine, feminine, and neuter (if present). Although for most inanimate objects gender assignment is only formal,[5] for human referents masculine/feminine markings are assigned on a semantic basis. Grammatical gender is defined by a system of morphosyntactic agreement, where several parts of speech beside the noun (e.g., verbs, determiners, adjectives) carry gender inflections.

In light of this, the English sentence ''*He/She* is a good friend'' has no overt expression of gender in a genderless language like Turkish (''*O* iyi bir arkadaş''), whereas Spanish spreads several masculine or feminine markings (''*El/la*

[4]Also referred to as *natural* gender languages. Following McConnell-Ginet (2013), we prefer notional to avoid terminological overlapping with ''natural'', i.e., biological/anatomical sexual categories. For a wider discussion on the topic, see Nevalainen and Raumolin-Brunberg (1993); Curzan (2003).

[5]E.g., ''moon'' is masculine in German, feminine in French, and neuter in Greek.

es *un/a buen/a amig<u>o</u>/a*''). Although general, such macro-categories allow us to highlight typological differences across languages. These are crucial to frame gender issues in both human and machine translation. Also, they exhibit to what extent speakers of each group are led to think and communicate via binary distinctions,[6] as well as underline the relative complexity in carving out a space for lexical innovations that encode non-binary gender (Hord, 2016; Conrod, 2020). In this sense, while English is bringing the singular *they* in common use and developing neo-pronouns (Bradley et al., 2019), for grammatical gender languages like Spanish neutrality requires the development of neo-morphemes (''*Ell<u>e</u>* es *un<u>e</u> buen<u>e</u> amig<u>ue</u>''*).

3.1.2 Social Gender Connotations

To understand gender bias, we have to grasp not only the structure of different languages, but also how linguistic expressions are connoted, deployed, and perceived (Hellinger and Motschenbacher, 2015). In grammatical gender languages, feminine forms are often subject to a so-called semantic derogation (Schulz, 1975), for example, in French, *couturier* (fashion designer) vs. *couturière* (seamstress). English is no exception (e.g., *governor*/*governess*).

Moreover, bias can lurk underneath seemingly neutral forms. Such is the case of epicene (i.e., gender neutral) nouns where gender is not grammatically marked. Here, gender assignment is linked to (typically binary) social gender, that is, ''the socially imposed dichotomy of masculine and feminine role and character traits'' (Kramarae and Treichler, 1985). As an illustration, Danish speakers tend to pronominalize *dommer* (judge) with *han* (he) when referring to the whole occupational category (Gomard, 1995; Nissen, 2002). Social gender assignment varies across time and space (Lyons, 1977; Romaine, 1999; Cameron, 2003) and regards stereotypical assumptions about what is typical or appropriate for men and women. Such assumptions impact our perceptions (Hamilton, 1988; Gygax et al., 2008; Kreiner et al., 2008) and influence our behavior (e.g., leading individuals to identify with and fulfill stereotypical expectations; Wolter and Hannover, 2016; Sczesny et al., 2018) and verbal communication (e.g., women are often misquoted in the academic community; Krawczyk, 2017).

Translation studies highlight how social gender assignment influences translation choices (Jakobson, 1959; Chamberlain, 1988; Comrie, 1999; Di Sabato and Perri, 2020). Primarily, the problem arises from typological differences across languages and their gender systems. Nonetheless, socio-cultural factors also influence how translators deal with such differences. Consider the character of the cook in Daphne du Maurier's *Rebecca*, whose gender is never explicitly stated in the whole book. In the lack of any available information, translators of five grammatical gender languages represented the character as either a man or a woman (Wandruszka, 1969; Nissen, 2002). Although extreme, this case can illustrate the situation of uncertainty faced by MT: the mapping of one-to-many forms in gender prediction. But, as discussed in §4.1, mistranslations occur when contextual gender information is available as well.

3.1.3 Gender and Language Use

Language use varies between demographic groups and reflects their backgrounds, personalities, and social identities (Labov, 1972; Trudgill, 2000; Pennebaker and Stone, 2003). In this light, the study of gender and language variation has received much attention in socio- and corpus linguistics (Holmes and Meyerhoff, 2003; Eckert and McConnell-Ginet, 2013). Research conducted in speech and text analysis highlighted several gender differences, which are exhibited at the phonological and lexical-syntactic level. For example, women rely more on hedging strategies (''it seems that''), purpose clauses (''in order to''), first-person pronouns, and prosodic exclamations (Mulac et al., 2001; Mondorf, 2002; Brownlow et al., 2003). Although some correspondences between gender and linguistic features hold across cultures and languages (Smith, 2003; Johannsen et al., 2015), it should be kept in mind that they are far from universal[7] and should not be intended in a stereotyped and oversimplified

[6]Outside of the Western paradigm, there are cultures whose languages traditionally encode gender outside of the binary (Epple, 1998; Murray, 2003; Hall and O'Donovan, 2014).

[7]It has been largely debated whether gender-related differences are inherently biological or cultural and social products (Mulac et al., 2001). Currently, the idea that they depend on biological reasons is largely rejected (Hyde, 2005) in favor of a socio-cultural or performative perspective (Butler, 1990).

manner (Bergvall et al., 1996; Nguyen et al., 2016; Koolen and van Cranenburgh, 2017).

Drawing on gender-related features proved useful for building demographically informed NLP tools (Garimella et al., 2019) and personalized MT models (Mirkin et al., 2015; Bawden et al., 2016; Rabinovich et al., 2017). However, using personal gender as a variable requires a prior understanding of which categories may be salient, and a critical reflection on how gender is intended and ascribed (Larson, 2017). Otherwise, if we assume that the only relevant categories are ''male'' and ''female'', our models will inevitably fulfill such a reductionist expectation (Bamman et al., 2014).

3.2 Gender Bias in MT

To date, an overview of how several factors may contribute to gender bias in MT does not exist. We identify and clarify concurring problematic causes, accounting for the context in which systems are developed and used (§2). To this aim, we rely on the three overarching categories of bias described by Friedman and Nissenbaum (1996), which foreground different sources that can lead to machine bias. These are: pre-existing bias—rooted in our institutions, practices and attitudes (§3.2.1); technical bias—due to technical constraints and decisions (§3.2.2); and emergent bias —arising from the interaction between systems and users (§3.2.3). We consider such categories as placed along a continuum, rather than being discrete.

3.2.1 Pre-existing Bias

MT models are known to reflect gender disparities present in the data. However, reflections on such generally invoked disparities are often overlooked. Treating data as an abstract, monolithic entity (Gitelman, 2013)—or relying on ''overly broad/overloaded terms like *training data bias*''[8] (Suresh and Guttag, 2019)—do not encourage reasoning on the many factors of which data are the product: first and foremost, the historical, socio-cultural context in which they are generated.

A starting point to tackle these issues is the Europarl corpus (Koehn, 2005), where only 30% of sentences are uttered by women (Vanmassenhove et al., 2018). Such an imbalance is a direct window into the glass ceiling that has hampered women's access to parliamentary positions. This case exemplifies how data might be ''tainted with historical bias'', mirroring an ''unequal ground truth'' (Hacker, 2018). Other gender variables are harder to spot and quantify.

Empirical linguistics research pointed out that subtle gender asymmetries are rooted in languages' use and structure. For instance, an important aspect regards how women are referred to. Femaleness is often explicitly invoked when there is no textual need to do so, even in languages that do not require overt gender marking. A case in point regards Turkish, which differentiates *cocuk* (child) and *kiz cocugu* (female child) (Braun, 2000). Similarly, in a corpus search, Romaine (2001) found 155 explicit female markings for *doctor* (female, woman, or lady doctor), compared with only 14 *male doctor*. Feminist language critique provided extensive analysis of such a phenomenon by highlighting how referents in discourse are considered men by default unless explicitly stated (Silveira, 1980; Hamilton, 1991). Finally, prescriptive top–down guidelines limit the linguistic visibility of gender diversity, for example, the Real Academia de la Lengua Española recently discarded the official use of nonbinary innovations and claimed the functionality of masculine generics (Mundo, 2018; López et al., 2020).

By stressing such issues, we are not condoning the reproduction of pre-existing bias in MT. Rather, the above-mentioned concerns are the starting point to account for when dealing with gender bias.

3.2.2 Technical Bias

Technical bias comprises aspects related to data creation, model design, and training and testing procedures. If present in training and testing samples, asymmetries in the semantics of language use and gender distribution are respectively learned by MT systems and rewarded in their evaluation. However, as just discussed, biased representations are not merely quantitative, but also qualitative. Accordingly, straightforward procedures (e.g., balancing the number of speakers in existing datasets) do not ensure a fairer representation of gender in MT outputs. Since datasets are a crucial source of bias, it is also crucial to advocate for a careful data curation (Mehrabi et al., 2019; Paullada et al., 2020; Hanna et al., 2021; Bender et al., 2021), guided by pragmatically

[8] See Johnson (2020a) and Samar (2020) for a discussion on how such narrative can be counterproductive for tackling bias.

and socially informed analyses (Hitti et al., 2019; Sap et al., 2020; Devinney et al., 2020) and annotation practices (Gaido et al., 2020).

Overall, while data can mirror gender inequalities and offer adverse shortcut learning opportunities, it is ''quite clear that data alone rarely constrain a model sufficiently'' (Geirhos et al., 2020) nor explain the fact that models *overamplify* (Shah et al., 2020) such inequalities in their outputs. Focusing on models' components, Costa-jussà et al. (2020b) demonstrate that architectural choices in multilingual MT impact the systems' behavior: Shared encoder-decoders retain less gender information in the source embeddings and less diversion in the attention than language-specific encoder-decoders (Escolano et al., 2021), thus disfavoring the generation of feminine forms. While discussing the loss and decay of certain words in translation, Vanmassenhove et al. (2019, 2021) attest to the existence of an algorithmic bias that leads underrepresented forms in the training data (as it may be the case for feminine references) to further decrease in the MT output. Specifically, Roberts et al. (2020) prove that beam search (unlike sampling) is skewed toward the generation of more frequent (masculine) pronouns, as it leads models to an extreme operating point that exhibits zero variability.

Thus, efforts towards understanding and mitigating gender bias should also account for the model and its algorithmic implications. To date, this remains largely unexplored.

3.2.3 Emergent Bias

Emergent bias may arise when a system is used in a different context than the one it was designed for—for example, when it is applied to another demographic group. From car crash dummies to clinical trials, we have evidence of how not accounting for gender differences brings to the creation of male-grounded products with dire consequences (Liu and Dipietro Mager, 2016; Criado-Perez, 2019), such as higher death and injury risks in vehicle crashes and less effective medical treatments for women. Similarly, unbeknownst to their creators, MT systems that are not intentionally envisioned for a diverse range of users will not generalize for the feminine segment of the population. Hence, in the interaction with an MT system, a woman will likely be misgendered or not have her linguistic style preserved (Hovy et al., 2020). Other conditions of user/system mismatch may be the result of changing societal knowledge and values. A case in point regards Google Translate's historical decision to adjust its system for instances of gender ambiguity. Since its launch twenty years ago, Google had provided only one translation for single-word gender-ambiguous queries (e.g., *professor* translated in Italian with the masculine *professore*). In a community increasingly conscious of the power of language to hardwire stereotypical beliefs and women's invisibility (Lindqvist et al., 2019; Beukeboom and Burgers, 2019), the bias exhibited by the system was confronted with a new sensitivity. The service's decision (Kuczmarski, 2018) to provide a double feminine/masculine output (*professor→professoressa|professore*) stems from current demands for gender-inclusive resolutions. For the recognition of non-binary groups (Richards et al., 2016), we invite studies on how such modeling could be integrated with neutral strategies (§6).

4 Assessing Bias

First accounts on gender bias in MT date back to Frank et al. (2004). Their manual analysis pointed out how English-German MT suffers from a dearth of linguistic competence, as it shows severe difficulties in recovering syntactic and semantic information to correctly produce gender agreement.

Similar inquiries were conducted on other target grammatical gender languages for several commercial MT systems (Abu-Ayyash, 2017; Monti, 2017; Rescigno et al., 2020). While these studies focused on contrastive phenomena, Schiebinger (2014)[9] went beyond linguistic insights, calling for a deeper understanding of gender bias. Her article on Google Translate's ''masculine default'' behavior emphasized how such a phenomenon is related to the larger issue of gender inequalities, also perpetuated by socio-technical artifacts (Selbst et al., 2019). All in all, these qualitative analyses demonstrated that gender problems encompass all three MT paradigms (neural, statistical, and rule-based), preparing the ground for quantitative work.

To attest the existence and scale of gender bias across several languages, dedicated benchmarks, evaluations, and experiments have been designed.

[9]See also Schiebinger's project *Gendered Innovations*: `http://genderedinnovations.stanford.edu/case-studies/nlp.html`.

Study	Benchmark	Gender	Harms
(Prates et al., 2018)	Synthetic, U.S. Bureau of Labor Statistics	b	R: under-rep, stereotyping
(Cho et al., 2019)	Synthetic equity evaluation corpus (EEC)	b	R: under-rep, stereotyping
(Gonen and Webster, 2020)	BERT-based perturbations on natural sentences	b	R: under-rep, stereotyping
(Stanovsky et al., 2019)	WinoMT	b	R: under-rep, stereotyping
(Vanmassenhove et al., 2018)	Europarl (generic)	b	A: quality
(Hovy et al., 2020)	Trustpilot (reviews with gender and age)	b	R: under-rep

Table 1: For each **Study**, the Table shows on which **Benchmark** gender bias is assessed, how **Gender** is intended (here only in binary (b) terms). Finally, we indicate which (R)epresentational—*under-representation* and *stereotyping*—or (A)llocational **Harm**—as reduced *quality* of service—is addressed in the study.

We first discuss large scale analyses aimed at assessing gender bias in MT, grouped according to two main conceptualizations: *i)* works focusing on the weight of prejudices and stereotypes in MT (§4.1); and *ii)* studies assessing whether gender is properly preserved in translation (§4.2). In accordance with the human-centered approach embraced in this survey, in Table 1 we map each work to the harms (see §2) ensuing from the biased behaviors they assess. Finally, we review existing benchmarks for comparing MT performance across genders (§4.3).

4.1 MT and Gender Stereotypes

In MT, we record prior studies concerned with pronoun translation and coreference resolution across typologically different languages accounting for both animate and inanimate referents (Hardmeier and Federico, 2010; Le Nagard and Koehn, 2010; Guillou, 2012). For the specific analysis on gender bias, instead, such tasks are exclusively studied in relation to human entities.

Prates et al. (2018) and Cho et al. (2019) design a similar setting to assess gender bias. Prates et al. (2018) investigate pronoun translation from 12 genderless languages into English. Retrieving ∼1,000 job positions from the U.S. Bureau of Labor Statistics, they build simple constructions like the Hungarian ''*ö* egy *mérnök*'' (''*he/she* is an *engineer*''). Following the same template, Cho et al. (2019) extend the analysis to Korean-English including both occupations and sentiment words (e.g., *kind*). As their samples are ambiguous by design, the observed predictions of he/she pronouns should be random, yet they show a strong masculine skew.[10]

To further analyze the under-representation of *she* pronouns, Prates et al. (2018) focus on 22 macro-categories of occupation areas and compare the proportion of pronoun predictions against the real-world proportion of men and women employed in such sectors. In this way, they find that MT not only yields a masculine default, but it also underestimates feminine frequency at a greater rate than occupation data alone suggest. Such an analysis starts by acknowledging pre-existing bias (see §3.2.1)—for example, low rates of women in STEM—to attest the existence of machine bias, and defines it as the exacerbation of actual gender disparities.

Going beyond word lists and simple synthetic constructions, Gonen and Webster (2020) inspect the translation into Russian, Spanish, German, and French of natural yet ambiguous English sentences. Their analysis on the ratio and type of generated masculine/feminine job titles consistently exhibits social asymmetries for target grammatical gender languages (e.g., *lecturer* masculine vs. *teacher* feminine). Finally, Stanovsky et al. (2019) assess that MT is skewed to the point of actually ignoring explicit feminine gender information in source English sentences. For instance, MT systems yield a wrong masculine translation of the job title *baker*, although it is referred to by the pronoun *she*. Aside from overlooking of overt gender mentions, the model's reliance on unintended (and irrelevant) cues for gender assignment is further confirmed by the

[10]Cho et al. (2019) highlight that a higher frequency of feminine references in the MT output does not necessarily imply a bias reduction. Rather, it may reflect gender stereotypes, as for *hairdresser* that is skewed toward feminine. This observation points to the tension between frequency count, suitable for testing under-representation, and qualitative-oriented analysis on bias conceptualized in terms of stereotyping.

fact that adding a socially connoted (but formally epicene) adjective (the *pretty* baker) pushes models toward feminine inflections in translation.

We observe that the propagation of stereotypes is a widely researched form of gender asymmetries in MT, one that so far has been largely narrowed down to occupational stereotyping. After all, occupational stereotyping has been studied by different disciplines (Greenwald et al., 1998) attested across cultures (Lewis and Lupyan, 2020), and it can be easily detected in MT across multiple language directions with consistent results. Current research should not neglect other stereotyping dynamics, as in the case of Stanovsky et al. (2019) and Cho et al. (2019), who include associations to physical characteristics or psychological traits. Also, the intrinsically contextual nature of societal expectations advocates for the study of culture-specific dimensions of bias. Finally, we signal that the BERT-based perturbation method by Webster et al. (2019) identifies other bias-susceptible nouns that tend to be assigned to a specific gender (e.g., *fighter* as masculine). As Blodgett (2021) underscores, however, ''the existence of these undesirable correlations is not sufficient to identify them as normatively undesirable''. It should thus be investigated whether such statistical preferences can cause harms (e.g., by checking if they map to existing harmful associations or quality of service disparities).

4.2 MT and Gender Preservation

Vanmassenhove et al. (2018) and Hovy et al. (2020) investigate whether speakers' gender[11] is properly reflected in MT. This line of research is preceded by findings on gender personalization of statistical MT (Mirkin et al., 2015; Bawden et al., 2016; Rabinovich et al., 2017), which claim that gender ''signals'' are weakened in translation.

Hovy et al. (2020) conjecture the existence of age and gender stylistic bias due to models' under-exposure to the writings of women and younger segments of the population. To test this hypothesis, they automatically translate a corpus of online reviews with available metadata about users (Hovy et al., 2015). Then, they compare such demographic information with the prediction of age and gender classifiers run on the MT output. Results indicate that different commercial MT models systematically make authors ''sound'' older and male. Their study thus concerns the under-representation of the language used ''by'' certain speakers and how it is perceived (Blodgett, 2021). However, the authors do not inspect which linguistic choices MT overproduces, nor which stylistic features may characterize different socio-demographic groups.

Still starting from the assumption that demographic factors influence language use, Vanmassenhove et al. (2018) probe MT's ability to preserve speaker's gender translating from English into ten languages. To this aim, they develop gender-informed MT models (see § 5.1) whose outputs are compared with those obtained by their baseline counterparts. Tested on a set for spoken language translation (Koehn, 2005), their enhanced models show consistent gains in terms of overall quality when translating into grammatical gender languages, where speaker's references are often marked. For instance, the French translation of ''I'm *happy*'' is either ''Je suis *heureuse*'' or ''Je suis *hereux*'' for a female/male speaker, respectively. Through a focused cross-gender analysis (carried out by splitting their English-French test set into 1st person male vs. female data) they assess that the largest margin of improvement for their gender-informed approach concerns sentences uttered by women, since the results of their baseline disclose a quality of service disparity in favor of male speakers. As well as morphological agreement, they also attribute such improvement to the fact that their enhanced model produces gendered preferences in other word choices. For instance, it opts for *think* rather than *believe*, which is in concordance with corpus studies claiming a tendency for women to use less assertive speech (Newman et al., 2008). Note that the authors rely on manual analysis to ascribe performance differences to gender-related features. In fact, global evaluations on generic test sets alone are inadequate to pointedly measure gender bias.

4.3 Existing Benchmarks

MT outputs are typically evaluated against reference translations employing standard metrics such as BLEU (Papineni et al., 2002) or TER (Snover et al., 2006). This procedure poses two challenges. First, these metrics provide coarse-grained scores for translation quality, as they treat

[11] Note that these studies distinguish speakers into female/male. As discussed in §3.1.3, we invite a reflection on the appropriateness and use of such categories.

all errors equally and are rather insensitive to specific linguistic phenomena (Sennrich, 2017). Second, generic test sets containing the same gender imbalance present in the training data can reward biased predictions. Here, we describe the publicly available MT *Gender Bias Evaluation Testsets* (GBETs) (Sun et al., 2019), that is, benchmarks designed to probe gender bias by isolating the impact of gender from other factors that may affect systems' performance. Note that different benchmarks and metrics respond to different conceptualizations of bias (Barocas et al., 2019). Common to them all in MT, however, is that biased behaviors are formalized by using some variants of averaged performance[12] disparities across gender groups, comparing the accuracy of gender predictions on an equal number of masculine, feminine, and neutral references.

Escudé Font and Costa-jussà (2019) developed the bilingual English-Spanish **Occupations test set.** It consists of 1,000 sentences equally distributed across genders. The phrasal structure envisioned for their sentences is ''I've known {*her*|*him*|<*proper noun*>} for a long time, my friend works as {a|an} <occupation>''. The evaluation focuses on the translation of the noun *friend* into Spanish (*amigo/a*). Since gender information is present in the source context and sentences are the same for both masculine/feminine participants, an MT system exhibits gender bias if it disregards relevant context and cannot provide the correct translation of *friend* at the same rate across genders.

Stanovsky et al. (2019) created **WinoMT** by concatenating two existing English GBETs for coreference resolution (Rudinger et al., 2018; Zhao et al., 2018a). The corpus consists of 3,888 Winogradesque sentences presenting two human entities defined by their role and a subsequent pronoun that needs to be correctly resolved to one of the entities (e.g., ''The *lawyer* yelled at the *hairdresser* because *he* did a bad job''). For each sentence, there are two variants with either *he* or *she* pronouns, so as to cast the referred annotated entity (*hairdresser*) into a proto- or anti-stereotypical gender role. By translating WinoMT into grammatical gender languages, one can thus measure systems' ability to resolve the anaphoric relation and pick the correct feminine/masculine inflection for the occupational noun. On top of quantifying under-representation as the difference between the total amount of translated feminine and masculine references, the subdivision of the corpus into proto- and anti-stereotypical sets also allows verifying if MT predictions correlate with occupational stereotyping.

Finally, Saunders et al. (2020) enriched the original version of WinoMT in two different ways. First, they included a third gender-neutral case based on the singular *they* pronoun, thus paving the way to account for non-binary referents. Second, they labeled the entity in the sentence which is not coreferent with the pronoun (*lawyer*). The latter annotation is used to verify the shortcomings of some mitigating approaches as discussed in §5.

The above-mentioned corpora are known as *challenge sets*, consisting of sentences created *ad hoc* for diagnostic purposes. In this way, they can be used to quantify bias related to stereotyping and under-representation in a controlled environment. However, since they consist of a limited variety of synthetic gender-related phenomena, they hardly address the variety of challenges posed by real-world language and are relatively easy to overfit. As recognized by Rudinger et al. (2018) ''they may demonstrate the presence of gender bias in a system, but not prove its absence''.

The Arabic Parallel Gender Corpus (Habash et al., 2019) includes an English-Arabic test set[13] retrieved from OpenSubtitles natural language data (Lison and Tiedemann, 2016). Each of the 2,448 sentences in the set exhibits a first person singular reference to the speaker (e.g., ''I'm *rich*''). Among them, ∼200 English sentences require gender agreement to be assigned in translation. These were translated into Arabic in both gender forms, obtaining a quantitatively and qualitatively equal amount of sentence pairs with annotated masculine/feminine references. This natural corpus thus allows for cross-gender evaluations on MT production of correct speaker's gender agreement.

MuST-SHE (Bentivogli et al., 2020) is a natural benchmark for three language pairs (English-French/Italian/Spanish). Built on TED talks data (Cattoni et al., 2021), for each language pair it comprises ∼1,000 (*audio*, *transcript*, *transla-*

[12]This is a value-laden option (Birhane et al., 2020), and not the only possible one (Mitchell et al., 2020). For a broader discussion on measurement and bias we refer the reader also to Jacobs (2021); Jacobs et al. (2020).

[13]Overall, the corpus comprises over 12,000 annotated sentences and 200,000 synthetic sentences.

Approach	Authors	Benchmark	Gender	Harms
Gender tagging (sentence-level)	Vanmassenhove et al.	Europarl (generic)	b	R: under-rep, A: quality
	Elaraby et al.	Open subtitles (generic)	b	R: under-rep, A: quality
Gender tagging (word-level)	Saunders et al.	expanded WinoMT	nb	R: under-rep, stereotyping
	Stafanovičs et al.	WinoMT	b	R: under-rep, stereotyping
Adding context	Basta et al.	WinoMT	b	R: under-rep, stereotyping
Word-embeddings	Escudé Font and Costa-jussà	Occupation test set	b	R: under-rep
Fine-tuning	Costa-jussà and de Jorge	WinoMT	b	R: under-rep, stereotyping
Black-box injection	Moryossef et al.	Open subtitles (selected sample)	b	R: under-rep, A: quality
Lattice-rescoring	Saunders and Byrne	WinoMT	b	R: under-rep, steretoyping
Re-inflection	Habash et al.; Alhafni et al.	Arabic Parallel Gender Corpus	b	R: under-rep, A: quality

Table 2: For each **Approach** and related **Authors**, the Table shows on which **Benchmark** it is tested, if **Gender** is intended in binary terms (b), or including non-binary (nb) identities. Finally, we indicate which (R)epresentational—*under-representation* and *stereotyping*—or (A)llocational **Harm**—as reduced *quality* of service—the approach attempts to mitigate.

tion) triplets, thus allowing evaluation for both MT and speech translation (ST). Its samples are balanced between masculine and feminine phenomena, and incorporate two types of constructions: *i)* sentences referring to the speaker (e.g., ''*I* was *born* in Mumbai''), and *ii)* sentences that present contextual information to disambiguate gender (e.g., ''My *mum* was *born* in Mumbai''). Since every gender-marked word in the target language is annotated in the corpus, MuST-SHE grants the advantage of complementing BLEU- and accuracy-based evaluations on gender translation for a great variety of phenomena.

Unlike challenge sets, natural corpora quantify whether MT yields reduced feminine representation in authentic conditions and whether the quality of service varies across speakers of different genders. However, as they treat all gender-marked words equally, it is not possible to identify if the model is propagating stereotypical representations.

All in all, we stress that each test set and metric is only a proxy for framing a phenomenon or an ability (e.g., anaphora resolution), and an approximation of what we truly intend to gauge. Thus, as we discuss in §6, advances in MT should account for the observation of gender bias in real-world conditions to avoid a situation in which achieving high scores on a mathematically formalized estimation could lead to a false sense of security. Still, benchmarks remain valuable tools to monitor models' behavior. As such, we remark that evaluation procedures ought to cover both models' general performance and gender-related issues. This is crucial to establish the capabilities and limits of mitigating strategies.

5 Mitigating Bias

To attenuate gender bias in MT, different strategies dealing with input data, learning algorithms, and model outputs have been proposed. As attested by Birhane et al. (2020), since advancements are oftentimes exclusively reported in terms of values internal to the machine learning field (e.g., efficiency, performance), it is not clear how such strategies are meeting societal needs by reducing MT-related harms. In order to conciliate technical perspectives with the intended social purpose, in Table 2 we map each mitigating approach to the harms (see §2) they are meant to alleviate, as well as to the benchmark their effectiveness is evaluated against. Complementarily, we hereby describe each approach by means of two categories: model debiasing (§5.1) and debiasing through external components (§5.2).

5.1 Model Debiasing

This line of work focuses on mitigating gender bias through architectural changes of general-purpose MT models or via dedicated training procedures.

Gender Tagging. To improve the generation of speaker's referential markings, Vanmassenhove et al. (2018) prepend a gender tag (M or F) to each source sentence, both at training and inference time. As their model is able to leverage this additional information, the approach proves useful to handle morphological agreement when translating from English into French. However, this solution requires additional metadata regarding the speakers' gender that might not always be feasible to acquire. Automatic annotation of speakers' gender (e.g., based on first names) is not advisable,

as it runs the risk of introducing additional bias by making unlicensed assumptions about one's identity.

Elaraby et al. (2018) bypass this risk by defining a comprehensive set of cross-lingual gender agreement rules based on POS tagging. In this way, they identify speakers' and listeners' gender references in an English-Arabic parallel corpus, which is consequently labeled and used for training. The idea, originally developed for spoken language translation in a two-way conversational setting, can be adapted for other languages and scenarios by creating new dedicated rules. However, in realistic deployment conditions where reference translations are not available, gender information still has to be externally supplied as metadata at inference time.

Stafanovičs et al. (2020) and Saunders et al. (2020) explore the use of word-level gender tags. While Stafanovičs et al. (2020) just report a gender translation improvement, Saunders et al. (2020) rely on the expanded version of WinoMT to identify a problem concerning gender tagging: It introduces noise if applied to sentences with references to multiple participants, as it pushes their translation toward the same gender. Saunders et al. (2020) also include a first non-binary exploration of neutral translation by exploiting an artificial dataset, where neutral tags are added and gendered inflections are replaced by placeholders. The results are inconclusive, however, most likely due to the small size and synthetic nature of their dataset.

Adding Context. Without further information needed for training or inference, Basta et al. (2020) adopt a generic approach and concatenate each sentence with its preceding one. By providing more context, they attest a slight improvement in gender translations requiring anaphoric coreference to be solved in English-Spanish. This finding motivates exploration at the document level, but it should be validated with manual (Castilho et al., 2020) and interpretability analyses since the added context can be beneficial for gender-unrelated reasons, such as acting as a regularization factor (Kim et al., 2019).

Debiased Word Embeddings. The two abovementioned mitigations share the same intent: supply the model with additional gender knowledge. Instead, Escudé Font and Costa-jussà (2019) leverage pre-trained word embeddings, which are debiased by using the hard-debiasing method proposed by Bolukbasi et al. (2016) or the GN-GloVe algorithm (Zhao et al., 2018b). These methods respectively remove gender associations or isolate them from the representations of English gender-neutral words. Escudé Font and Costa-jussà (2019) employ such embeddings on the decoder side, the encoder side, and both sides of an English-Spanish model. The best results are obtained by leveraging GN-GloVe embeddings on both encoder and decoder sides, increasing BLEU scores and gender accuracy. The authors generically apply debiasing methods developed for English also to their target language. However, with being Spanish a grammatical gender language, other language-specific approaches should be considered to preserve the quality of the original embeddings (Zhou et al., 2019; Zhao et al., 2020). We also stress that it is debatable whether depriving systems of some knowledge and diminish their perceptions is the right path toward fairer language models (Dwork et al., 2012; Caliskan et al., 2017; Gonen and Goldberg, 2019; Nissim and van der Goot, 2020). Also, Goldfarb-Tarrant et al. (2020) find that there is no reliable correlation between intrinsic evaluations of bias in word-embeddings and cascaded effects on MT models' biased behavior.

Balanced Fine-tuning. Costa-jussà and de Jorge (2020) rely on Gebiotoolkit (Costa-jussà et al., 2020c) to build gender-balanced datasets (i.e., featuring an equal amount of masculine/feminine references) based on Wikipedia biographies. By fine-tuning their models on such natural and more even data, the generation of feminine forms is overall improved. However, the approach is not as effective for gender translation on the anti-stereotypical WinoMT set. As discussed in §3.2.2, they employ a straightforward method that aims to increase the number of Wikipedia pages covering women in their training data. However, such coverage increase does not mitigate stereotyping harms, as it does not account for the qualitative different ways in which men and women are portrayed (Wagner et al., 2015).

5.2 Debiasing through External Components

Instead of directly debiasing the MT model, these mitigating strategies intervene in the inference phase with external dedicated components. Such approaches do not imply retraining, but introduce

the additional cost of maintaining separate modules and handling their integration with the MT model.

Black-box Injection. Moryossef et al. (2019) attempt to control the production of feminine references to the speaker and numeral inflections (plural or singular) for the listener(s) in an English-Hebrew spoken language setting. To this aim, they rely on a short construction, such as ''*she* said to *them*'', which is prepended to the source sentence and then removed from the MT output. Their approach is simple, it can handle two types of information (gender and number) for multiple entities (speaker and listener), and improves systems' ability to generate feminine target forms. However, as in the case of Vanmassenhove et al., 2018 and Elaraby et al. (2018), it requires metadata about speakers and listeners.

Lattice Re-scoring. Saunders and Byrne (2020) propose to post-process the MT output with a lattice re-scoring module. This module exploits a transducer to create a lattice by mapping gender marked words in the MT output to all their possible inflectional variants. Developed for German, Spanish, and Hebrew, all the sentences corresponding to the paths in the lattice are re-scored with another model, which has been gender-debiased but at the cost of lower generic translation quality. Then, the sentence with the highest probability is picked as the final output. When tested on WinoMT, such an approach leads to an increase in the accuracy of gender forms selection. Note that the gender-debiased system is created by fine-tuning the model on an *ad hoc* built tiny set containing a balanced number of masculine/feminine forms. Such an approach, also known as *counterfactual data augmentation* (Lu et al., 2020), requires one to create identical pairs of sentences differing only in terms of gender references. In fact, Saunders and Byrne (2020) compile English sentences following this schema: ''The <profession> finished <*his*|*her*> work''. Then, the sentences are automatically translated and manually checked. In this way, they obtain gender-balanced parallel corpus. Thus, to implement their method for other language pairs, the generation of new data is necessary. For the fine-tuning set, the effort required is limited as the goal is to alleviate stereotypes by focusing on a pre-defined occupational lexicon. However, data augmentation is very demanding for complex sentences that represent a rich variety of gender agreement phenomena[14] such as those occurring in natural language scenarios.

Gender Re-inflection. Habash et al. (2019) and Alhafni et al. (2020) confront the problem of speaker's gender agreement in Arabic with a post-processing component that re-inflects first person references into masculine/feminine forms. In Alhafni et al. (2020), the preferred gender of the speaker and the translated Arabic sentence are fed to the component, which re-inflects the sentence in the desired form. In Habash et al. (2019) the component can be: *i*) a two-step system that first identifies the gender of first person references in an MT output, and then re-inflects them in the opposite form; or *ii*) a single-step system that always produces both forms from an MT output. Their method does not necessarily require speakers' gender information: If metadata are supplied, the MT output is re-inflected accordingly; otherwise, both feminine/masculine inflections are offered (leaving to the user the choice of the appropriate one). The implementation of the re-inflection component was made possible by the Arabic Parallel Gender Corpus (see §4.3), which demanded an expensive work of manual data creation. However, such corpus grants research on English-Arabic the benefits of a wealth of gender-informed natural language data that have been curated to avoid hetero-centrist interpretations and preconceptions (e.g., proper names and speakers of sentences like ''that's my wife'' are flagged as gender-ambiguous). Along the same line, Google Translate also delivers two outputs for short gender-ambiguous queries (Johnson, 2020b). Among languages with grammatical gender, the service is currently available only for English-Spanish.

In light of the above, we remark that there is no conclusive state-of-the-art method for mitigating bias. The discussed interventions in MT tend to respond to specific aspects of the problem with modular solutions, but if and how they can be integrated within the same MT system remains unexplored. As we have discussed through the survey, the umbrella term ''gender bias'' refers to a wide array of undesirable phenomena. Thus, it is unlikely that a one-size-fits-all solution will be

[14]Zmigrod et al. (2019) proposed an automatic approach for augmenting data into morphologically rich languages, but it is only viable for simple constructions with one single entity.

able tackle problems that differ from one another, as they depend on, for example, how bias is conceptualized, the language combinations, the kinds of corpora used. As a result, we believe that generalization and scalability should not be the only criteria against which mitigating strategies are valued. Conversely, we should make room for openly context-aware interventions. Finally, gender bias in MT is a socio-technical problem. We thus highlight that engineering interventions alone are not a panacea (Chang, 2019) and should be integrated with long-term multidisciplinary commitment and practices (D'Ignazio and Klein, 2020; Gebru, 2020) necessary to address bias in our community, hence in its artifacts, too.

6 Conclusion and Key Challenges

Studies confronting gender bias in MT are rapidly emerging; in this paper we presented them within a unified framework to critically overview current conceptualizations and approaches to the problem. Since gender bias is a multifaceted and interdisciplinary issue, in our discussion we integrated knowledge from related disciplines, which can be instrumental to guide future research and make it thrive. We conclude by suggesting several directions that can help this field going forward.

Model De-biasing. Neural networks rely on easy-to-learn shortcuts or ''cheap tricks'' (Levesque, 2014), as picking up on spurious correlations offered by training data can be easier for machines than learning to actually solve a specific task. What is ''easy to learn'' for a model depends on the *inductive bias* (Sinz et al., 2019; Geirhos et al., 2020) resulting from architectural choices, training data and learning rules. We think that explainability techniques (Belinkov et al., 2020) represent a useful tool to identify spurious cues (features) exploited by the model during inference. Discerning them can provide the research community with guidance on how to improve models' generalization by working on data, architectures, loss functions and optimizations. For instance, data responsible for spurious features (e.g., stereotypical correlations) might be recognized and their weight at training time might be lowered (Karimi Mahabadi et al., 2020). Additionally, state-of-the-art architectural choices and algorithms in MT have mostly been studied in terms of overall translation quality without specific analyses regarding gender translation. For instance, current systems segment text into subword units with statistical methods that can break the morphological structure of words, thus losing relevant semantic and syntactic information in morphologically rich languages (Niehues et al., 2016; Ataman et al., 2017). Several languages show complex feminine forms, typically derivative and created by adding a suffix to the masculine form, such as *Lehrer/Lehrerin* (de), *studente/studentessa* (it). It would be relevant to investigate whether, compared to other segmentation techniques, statistical approaches disadvantage (rarer and more complex) feminine forms. The MT community should not overlook focused hypotheses of such kind, as they can deepen our comprehension of the gender bias conundrum.

Non-textual Modalities. Gender bias for non-textual automatic translations (e.g., audiovisual) has been largely neglected. In this sense, ST represents a small niche (Costa-jussà et al., 2020a). For the translation of speaker-related gender phenomena, Bentivogli et al. (2020) prove that direct ST systems exploit speaker's vocal characteristics as a gender cue to improve feminine translation. However, as addressed by Gaido et al. (2020), relying on physical gender cues (e.g., pitch) for such task implies reductionist gender classifications (Zimman, 2020) making systems potentially harmful for a diverse range of users. Similarly, although image-guided translation has been claimed useful for gender translation since it relies on visual inputs for disambiguation (Frank et al., 2018; Ive et al., 2019), it could bend toward stereotypical assumptions about appearance. Further research should explore such directions to identify potential challenges and risks, by drawing on bias in image captioning (van Miltenburg, 2019) and consolidated studies from the fields of automatic gender recognition and human–computer interaction (HCI) (Hamidi et al., 2018; Keyes, 2018; May, 2019).

Beyond Dichotomies. Besides a few notable exceptions for English NLP tasks (Manzini et al., 2019; Cao and Daumé III, 2020; Sun et al., 2021) and one in MT (Saunders et al., 2020), the discussion around gender bias has been reduced to the binary masculine/feminine dichotomy. Although research in this direction is currently hampered by the absence of data, we invite considering inclusive solutions and exploring nuanced dimensions of gender. Starting from language practices, Indirect Non-binary Language (INL) overcomes

gender specifications (e.g., using *service, humankind* rather than *waiter/waitress* or *mankind*).[15] Although more challenging, INL can be achieved also for grammatical gender languages (Motschenbacher, 2014; Lindqvist et al., 2019), and it is endorsed for official EU documents (Papadimoulis, 2018). Accordingly, MT models could be brought to avoid binary forms and move toward gender-unspecified solutions, for example, adversarial networks including a discriminator that classifies speaker's linguistic expression of gender (masculine or feminine) could be employed to ''neutralize'' speaker-related forms (Li et al., 2018; Delobelle et al., 2020). Conversely, Direct Non-binary Language (DNL) aims at increasing the visibility of non-binary individuals via neologisms and neomorphemes (Bradley et al., 2019; Papadopoulos, 2019; Knisely, 2020). With DNL starting to circulate (Shroy, 2016; Santiago, 2018; López, 2019), the community is presented with the opportunity to promote the creation of inclusive data.

Finally, as already highlighted in legal and social science theory, discrimination can arise from the intersection of multiple identity categories (e.g., race and gender) (Crenshaw, 1989) which are not additive and cannot always be detected in isolation (Schlesinger et al., 2017). Following the MT work by Hovy et al. (2020), as well as other intersectional analyses from NLP (Herbelot et al., 2012; Jiang and Fellbaum, 2020) and AI-related fields (Buolamwini and Gebru, 2018), future studies may account for the interaction of gender attributes with other sociodemographic classes.

Human-in-the-Loop. Research on gender bias in MT is still restricted to lab tests. As such, unlike other studies that rely on participatory design (Turner et al., 2015; Cercas Curry et al., 2020; Liebling et al., 2020), the advancement of the field is not measured with people's experience in focus or in relation to specific deployment contexts. However, these are fundamental considerations to guide the field forward and, as HCI studies show (Vorvoreanu et al., 2019), to propel the creation of gender-inclusive technology. In particular, representational harms are intrinsically difficult to estimate and available benchmarks only provide a rough idea of their extent. This is an argument in favor of focused studies[16] on their individual or aggregate effects in everyday life. Also, we invite the whole development process to be paired with bias-aware research methodology (Havens et al., 2020) and HCI approaches (Stumpf et al., 2020), which can help to operationalize sensitive attributes like gender (Keyes et al., 2021). Finally, MT is not only built for people, but also by people. Thus, it is vital to reflect on the implicit biases and backgrounds of the people involved in MT pipelines at all stages and how they could be reflected in the model. This means starting from bottom-level countermeasures, engaging with translators (De Marco and Toto, 2019; Lessinger, 2020) and annotators (Waseem, 2016; Geva et al., 2019), and considering everyone's subjective positionality—and, crucially, also the lack of diversity within technology teams (Schluter, 2018; Waseem et al., 2020).

[15]INL suggestions have also been recently implemented within Microsoft text editors (Langston, 2020).

[16]To the best of our knowledge, the *Gender-Inclusive Language Models Survey* is the first project of this kind that includes MT. At time of writing it is available at: https://docs.google.com/forms/d/e/1FAIpQLSfKenp4RKtDhKA0WLqPflGSBV2VdBA9h3F8MwqRex_4kiCf9Q/viewform.

Acknowledgments

We would like to thank the anonymous reviewers and the TACL Action Editors. Their insightful comments helped us improve on the current version of the paper.

References

Emad A. S. Abu-Ayyash. 2017. Errors and non-errors in english-arabic machine translation of gender-bound constructs in technical texts. *Procedia Computer Science*, 117:73–80. https://doi.org/10.1016/j.procs.2017.10.095

Bashar Alhafni, Nizar Habash, and Houda Bouamor. 2020. Gender-aware reinflection using linguistically enhanced neural models. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 139–150, Online. Association for Computational Linguistics.

Duygu Ataman, Matteo Negri, Marco Turchi, and Marcello Federico. 2017. Linguistically

motivated vocabulary reduction for neural machine translation from Turkish to English. *The Prague Bulletin of Mathematical Linguistics*, 108(1)331–342. https://doi.org/10.1515/pralin-2017-0031

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160. https://doi.org/10.1111/josl.12080

Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook:org.

Christine Basta, Marta R. Costa-jussà, and José A. R. Fonollosa. 2020. Towards mitigating gender bias in a decoder-based neural machine translation model by adding contextual information. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop*, pages 99–102, Seattle, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.winlp-1.25

Rachel Bawden, Guillaume Wisniewski, and Hélène Maynard. 2016. Investigating gender adaptation for speech translation. In *Proceedings of the 23ème Conférence sur le Traitement Automatique des Langues Naturelles*, volume 2, pages 490–497, Paris, FR.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2020. On the linguistic representational power of neural machine translation models. *Computational Linguistics*, 46(1):1–52. https://doi.org/10.1162/coli_a_00367

Emily M. Bender. 2019. A typology of ethical risks in language technology with an eye towards where transparent documentation might help. In *CRAASH. The future of Artificial Intelligence: Language, Ethics, Technology*. Cambridge, UK.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604. https://doi.org/10.1162/tacl_a_00041

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT '21)*, pages 610–623, Online. ACM. https://doi.org/10.1145/3442188.3445922

Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia A. Di Gangi, Roldano Cattoni, and Marco Turchi. 2020. Gender in danger? Evaluating speech translation technology on the MuST-SHE Corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6923–6933, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.619

Victoria L. Bergvall, Janet M. Bing, and Alice F. Freed. 1996. *Rethinking Language and Gender Research: Theory and Practice*. London, UK. Addison Wesley Longman.

Camiel J. Beukeboom and Christian Burgers. 2019. How stereotypes are shared through language: A Review and Introduction of the Social Categories and Stereotypes Communication (SCSC) framework. *Review of Communication Research*, 7:1–37.

Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2020. The underlying values of machine learning research. In *Resistance AI Workshop @ NeurIPS*, Online.

Su Lin Blodgett. 2021. *Sociolinguistically Driven Approaches for Just Natural Language Processing*. Doctoral Dissertation.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of ''bias'' in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2016)*, volume 29, pages 4349–4357, Barcelona, ES. Curran Associates, Inc.

David Bourguignon, Vincent Y. Yzerbyt, Catia P. Teixeira, and Ginette Herman. 2015. When does it hurt? Intergroup permeability moderates the link between discrimination and self-esteem. *European Journal of Social Psychology*, 45(1):3–9. https://doi.org/10.1002/ejsp.2083

Evan D. Bradley, Julia Salkind, Ally Moore, and Sofi Teitsort. 2019. Singular 'they' and novel pronouns: gender-neutral, nonbinary, or both? *Proceedings of the Linguistic Society of America*, 4(1):36–1. https://doi.org/10.3765/plsa.v4i1.4542

Friederike Braun. 2000. *Geschlecht im Türkischen: Untersuchungen zum sprachlichen Umgang mit einer sozialen Kategorie*, Turcologica Series. Otto Harrassowitz Verlag, Wiesbaden, DE.

Sheila Brownlow, Julie A. Rosamond, and Jennifer A. Parker. 2003. Gender-linked linguistic behavior in television interviews. *Sex Roles*, 49(3–4):121–132. https://doi.org/10.1023/A:1024404812972

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, New York, USA. PMLR.

Judith Butler. 1990. *Gender Trouble: Feminism and the Subversion of Identity*. Routledge, New York, USA.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186. https://doi.org/10.1126/science.aal4230, PubMed: 28408601

Deborah Cameron. 2003. Gender issues in language change. *Annual Review of Applied Linguistics*, 23:187–201. https://doi.org/10.1017/S0267190503000266

Alex Campolo, Madelyn R. Sanfilippo, Meredith Whittaker, and Kate Crawford. 2017. AI Now Report 2017. *New York: AI Now Institute*.

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

Sheila Castilho, Maja Popović, and Andy Way. 2020. On context span needed for machine translation evaluation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3735–3742, Marseille, FR. European Language Resources Association.

Roldano Cattoni, Mattia A. Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. MuST-C: A multilingual corpus for end-to-end speech translation. *Computer Speech & Language*, 66:101155. https://doi.org/10.1016/j.csl.2020.101155

Amanda Cercas Curry, Judy Robertson, and Verena Rieser. 2020. Conversational assistants and gender stereotypes: Public perceptions and desiderata for voice personas. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 72–78, Online. Association for Computational Linguistics.

Lori Chamberlain. 1988. Gender and the metaphorics of translation. *Signs: Journal of Women in Culture and Society*, 13(3):454–472. https://doi.org/10.1086/494428

Kai-Wei Chang. 2019. Bias and fairness in natural language processing. Tutorial at the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. On measuring gender bias in translation of gender-neutral pronouns. In *Proceedings of the First Workshop on Gender Bias in Natural Language*

Processing, pages 173–181, Florence, IT. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3824

Aleksandra Cislak, Magdalena Formanowicz, and Tamar Saguy. 2018. Bias against research on gender bias. *Scientometrics*, 115(1):189–200.

Bernard Comrie. 1999. Grammatical gender systems: A linguist's assessment. *Journal of Psycholinguistic Research*, 28:457–466. https://doi.org/10.1023/A:1023212225540

Kirby Conrod. 2020. Pronouns and gender in language. *The Oxford Handbook of Language and Sexuality*. https://doi.org/10.1093/oxfordhb/9780190212926.013.63

Greville G. Corbett. 1991. *Gender*. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge, UK.

Greville G. Corbett. 2013. *The Expression of Gender*. De Gruyter Mouton, Berlin, DE.

Marta R. Costa-jussà. 2019. An analysis of gender bias studies in natural language processing. *Nature Machine Intelligence*, 1:495–496. https://doi.org/10.1038/s42256-019-0105-5

Marta R. Costa-jussà, Christine Basta, and Gerard I. Gállego. 2020a. Evaluating gender bias in speech translation. *arXiv preprint arXiv:2010.14465*.

Marta R. Costa-jussà and Adrià de Jorge. 2020. Fine-tuning neural machine translation on gender-balanced datasets. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 26–34, Online. Association for Computational Linguistics.

Marta R. Costa-jussà, Carlos Escolano, Christine Basta, Javier Ferrando, Roser Batlle, and Ksenia Kharitonova. 2020b. Gender bias in multilingual neural machine translation: The architecture matters. *arXiv preprint arXiv:2012.13176*.

Marta R. Costa-jussà, Pau Li Lin, and Cristina España-Bonet. 2020c. GeBioToolkit: Automatic extraction of gender-balanced multilingual corpus of Wikipedia biographies. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4081–4088, Marseille, FR. European Language Resources Association.

Colette G. Craig. 1994, Classifier languages. In Ronald E. Asher & James M. Y. Simpson, editor, *The Encyclopedia of Language and Linguistics*, volume 2, pages 565–569. Pergamon Press, Oxford, UK.

Kate Crawford. 2017. The trouble with bias. In *Conference on Neural Information Processing Systems (NIPS) – Keynote*, Long Beach, USA.

Kimberlé Crenshaw. 1989. Demarginalizing the intersection of race and sex: A Black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum*, 1989:139–167.

Caroline Criado-Perez. 2019. *Invisible Women: Exposing Data Bias in a World Designed for Men*. Chatto & Windus, London, UK.

Anne Curzan. 2003. *Gender Shifts in the History of English*. Cambridge University Press, Cambridge, UK. https://doi.org/10.1017/CBO9780511486913

Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G. Accessed: 2021-02-25.

Marcella De Marco and Piero Toto. 2019. Introduction: The potential of gender training in the translation classroom. In *Gender Approaches in the Translation Classroom: Training the Doers*, pages 1–7, Palgrave Macmillan, Cham, CH. https://doi.org/10.1007/978-3-030-04390-2_1

Pieter Delobelle, Paul Temple, Gilles Perrouin, Benoît Frénay, Patrick Heymans, and Bettina Berendt. 2020. Ethical adversaries: Towards mitigating unfairness with adversarial machine learning. In *Informal Proceedings of the Bias and Fairness in AI Workshop at ECML-PKDD (BIAS 2020)*. BIAS 2020.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2020. Semi-supervised topic

modeling for gender bias discovery in English and Swedish. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 79–92, Online. Association for Computational Linguistics.

Bruna Di Sabato and Antonio Perri. 2020. Grammatical gender and translation: A cross-linguistic overview. In Luise von Flotow and Hala Kamal, editors, *The Routledge Handbook of Translation, Feminism and Gender*. Routledge, New York, USA. https://doi.org/10.4324/9781315158938-32

Catherine D'Ignazio and Lauren F. Klein. 2020. *Data Feminism*. MIT Press, London, UK.

Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020. Multi-dimensional gender bias classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 314–331, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.23

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, pages 214–226, New York, USA. Association for Computing Machinery. https://doi.org/10.1145/2090236.2090255

Penelope Eckert and Sally McConnell-Ginet. 2013. *Language and Gender*. Cambridge University Press, Cambridge, UK. https://doi.org/10.1017/CBO9781139245883

Mostafa Elaraby, Ahmed Y. Tawfik, Mahmoud Khaled, Hany Hassan, and Aly Osama. 2018. Gender aware spoken language translation applied to English-Arabic. In *Proceedings of the 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, pages 1–6, Algiers, DZ. https://doi.org/10.1109/ICNLSP.2018.8374387

Carolyn Epple. 1998. Coming to terms with Navajo nádleehí: A critique of berdache, ''gay'', ''alternate gender'', and ''two-spirit''. *American Ethnologist*, 25(2):267–290.

Carlos Escolano, Marta R. Costa-jussà, José A. R. Fonollosa, and Mikel Artetxe. 2021. Multilingual machine translation: Closing the gap between shared and language-specific encoder-decoders. In *Proceedings of the 16th conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Online. https://doi.org/10.1525/ae.1998.25.2.267

Joel Escudé Font and Marta R. Costa-jussà. 2019. Equalizing gender bias in neural machine translation with word embeddings techniques. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 147–154, Florence, IT. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3821

Anne Fausto-Sterling. 2019. Gender/sex, sexual orientation, and identity are in the body: How did they get there? *The Journal of Sex Research*, 56(4–5):529–555.

Anke Frank, Christiane Hoffmann, and Maria Strobel. 2004. Gender issues in machine translation. *University of Bremen*.

Stella Frank, Desmond Elliott, and Lucia Specia. 2018. Assessing multilingual multimodal image description: Studies of native speaker preferences and translator choices. *Natural Language Engineering*, 24(3):393–413. https://doi.org/10.1017/S1351324918000074

Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, 14(3):330–347. https://doi.org/10.1145/230538.230561

Marco Gaido, Beatrice Savoldi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2020. Breeding gender-aware direct speech translation systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3951–3964, Online. International Committee on Computational Linguistics.

Aparna Garimella, Carmen Banea, Dirk Hovy, and Rada Mihalcea. 2019. Women's syntactic resilience and men's grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing. In *Proceedings of the 57th Annual*

Meeting of the Association for Computational Linguistics, pages 3493–3498, Florence, IT. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1339

Timnit Gebru. 2020. Race and gender. In Markus D. Dubber, Frank Pasquale, and Sunit Das, editors, *The Oxford Handbook of Ethics of AI*. Oxford Handbook Online. https://doi.org/10.1093/oxfordhb/9780190067397.013.16

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673. https://doi.org/10.1038/s42256-020-00257-z

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? An investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, CN. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1107

Lisa Gitelman. 2013. *Raw Data is an Oxymoron*. MIT Press.

Fiona Glen and Karen Hurrell. 2012. Measuring gender identity. https://www.equalityhumanrights.com/sites/default/files/technical_note_final.pdf. Accessed: 2021-02-25.

Bruce Glymour and Jonathan Herington. 2019. Measuring the biases that matter: The ethical and casual foundations for measures of fairness in algorithms. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 269–278, New York, USA. Association for Computing Machinery. https://doi.org/10.1145/3287560.3287573

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sanchez, Mugdha Pandya, and Adam Lopez. 2020. Intrinsic bias metrics do not correlate with application bias. *arXiv preprint arXiv:2012.15859*.

Kirsten Gomard. 1995. The (un)equal treatment of women in language: A comparative study of Danish, English, and German. *Working Papers on Language, Gender and Sexism*, 5(1):5–25.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Hila Gonen and Kellie Webster. 2020. Automatically identifying gender issues in machine translation using perturbations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1991–1995, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.180

Anthony G. Greenwald, Debbie E. McGhee, and Jordan L. K. Schwartz. 1998. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74(6):1464. https://doi.org/10.1037/0022-3514.74.6.1464, PubMed: 9654756

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

Pascal M. Gygax, Daniel Elmiger, Sandrine Zufferey, Alan Garnham, Sabine Sczesny, Lisa von Stockhausen, Friederike Braun, and Jane Oakhill. 2019. A language index of grammatical gender dimensions to study the impact of grammatical gender on the way we perceive women and men. *Frontiers in Psychology*, 10:1604. https://doi.org/10.3389/fpsyg.2019.01604, PubMed: 31379661

Pascal M. Gygax, Ute Gabriel, Oriane Sarrasin, Jane Oakhill, and Alan Garnham. 2008. Generically intended, but specifically interpreted: When beauticians, musicians and mechanics are all men. *Language and Cognitive Processes*, 23:464–485. https://doi.org/10.1080/01690960701702035

Nizar Habash, Houda Bouamor, and Christine Chung. 2019. Automatic gender identification and reinflection in Arabic. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 155–165, Florence, IT. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3822

Philipp Hacker. 2018. Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU law. *Common market law review*, 55(4):1143–1185.

Kira Hall and Veronica O'Donovan. 2014. Shifting gender positions among Hindi-speaking hijras. *Rethinking language and gender research: Theory and practice*, pages 228–266.

Foad Hamidi, Morgan K. Scheuerman, and Stacy M. Branham. 2018. Gender recognition or gender reductionism? The social implications of embedded gender recognition systems. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 1–13, New York, USA. Association for Computing Machinery. https://doi.org/10.1145/3173574.3173582

Mykol C. Hamilton. 1988. Using masculine generics: Does generic he increase male bias in the user's imagery? *Sex roles*, 19(11–12):785–799. https://doi.org/10.1007/BF00288993

Mykol C. Hamilton. 1991. Masculine bias in the attribution of personhood: People = male, male = people. *Psychology of Women Quarterly*, 15(3):393–402.

Alex Hanna, Andrew Smart, Ben Hutchinson, Christina Greer, Emily Denton, Margaret Mitchell, Oddur Kjartansson, and Parker Barnes. 2021. Towards accountability for machine learning datasets. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT '21)*, pages 560–575, Online. ACM.

Christian Hardmeier, Marta R. Costa-jussà, Kellie Webster, Will Radford, and Su Lin Blodgett. 2021. How to write a bias statement: Recommendations for submissions to the workshop on gender bias in NLP. *arXiv preprint arXiv:2104.03026*.

Christian Hardmeier and Marcello Federico. 2010. Modelling pronominal anaphora in statistical machine translation. In *Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT)*, pages 283–289, Paris, FR.

Lucy Havens, Melissa Terras, Benjamin Bach, and Beatrice Alex. 2020. Situated data, situated systems: A methodology to engage with power relations in natural language processing research. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 107–124, Online. Association for Computational Linguistics.

Marlis Hellinger and Hadumond Bußman. 2001. *Gender across Languages: The Linguistic Representation of Women and Men*, volume 1. John Benjamins Publishing, Amsterdam, NL. https://doi.org/10.1075/impact.9.05hel

Marlis Hellinger and Hadumond Bußman. 2002. *Gender across Languages: The Linguistic Representation of Women and Men*, volume 2. John Benjamins Publishing, Amsterdam, NL. https://doi.org/10.1075/impact.10.05hel

Marlis Hellinger and Hadumond Bußman. 2003. *Gender across Languages: The Linguistic Representation of Women and Men*, volume 3. John Benjamins Publishing, Amsterdam, NL. https://doi.org/10.1075/impact.05hel

Marlis Hellinger and Heiko Motschenbacher. 2015. *Gender Across Languages: The Linguistic Representation of Women and Men*, volume 4. John Benjamins, Amsterdam, NL. https://doi.org/10.1075/impact.36.01hel

Lisa A. Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755. Munich, DE. https://doi.org/10.1007/978-3-030-01219-9_47

Aurélie Herbelot, Eva von Redecker, and Johanna Müller. 2012. Distributional techniques for philosophical enquiry. In *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 45–54, Avignon, FR. Association for Computational Linguistics.

Yasmeen Hitti, Eunbee Jang, Ines Moreno, and Carolyne Pelletier. 2019. Proposed taxonomy for gender bias in text: A filtering methodology for the gender generalization subtype. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 8–17, Florence, IT. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3802

Janet Holmes and Miriam Meyerhoff. 2003. *The Handbook of Language and Gender*. Blackwell Publishing Ltd, Malden, USA. https://doi.org/10.1002/9780470756942

Levi C. R. Hord. 2016. Bucking the linguistic binary: Gender neutral language in English, Swedish, French, and German. *Western Papers in Linguistics / Cahiers linguistiques de Western*, 3(1):4.

Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. 2020. ''You sound just like your father'': Commercial machine translation systems include stylistic biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1686–1690, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.154

Dirk Hovy, Anders Johannsen, and Anders Søgaard. 2015. User review sites as a resource for large-scale sociolinguistic studies. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 452–461, Geneva, CH. International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/2736277.2741141

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, DE. Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-2096

Janet S. Hyde. 2005. The gender similarities hypothesis. *American Psychologist*, 60(6): 581–592. https://doi.org/10.1037/0003-066X.60.6.581, PubMed: 16173891

Julia Ive, Pranava Madhyastha, and Lucia Specia. 2019. Distilling translations with visual awareness. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6525–6538, Florence, IT. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1653

Abigail Z. Jacobs. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 375–385, New York, USA. Association for Computing Machinery. https://doi.org/10.1145/3442188.3445901

Abigail Z. Jacobs, Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. The meaning and measurement of bias: Lessons from natural language processing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 706, New York, USA. Association for Computing Machinery. https://doi.org/10.1145/3351095.3375671

Roman Jakobson. 1959. On Linguistic Aspects of Translation. In Reuben A. Brower, editor, *On translation*, pages 232–239. Cambridge, USA. Harvard University Press. https://doi.org/10.4159/harvard.9780674731615.c18

May Jiang and Christiane Fellbaum. 2020. Interdependencies of gender and race in contextualized word embeddings. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 17–25, Online. Association for Computational Linguistics.

Anders Johannsen, Dirk Hovy, and Anders Søgaard. 2015. Cross-lingual syntactic variation over age and gender. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 103–112, Beijing, CN. https://doi.org/10.18653/v1/K15-1011

Kari Johnson. 2020a. AI weekly: A deep learning pioneer's teachable moment on AI bias. https://venturebeat.com/2020/06/26/ai-weekly-a-deep-learning-pioneers-teachable-moment-on-ai-bias/. Accessed: 2021-02-25.

Melvin Johnson. 2020b. A scalable approach to reducing gender bias in Google Translate. https://ai.googleblog.com/2020/40/04/a-scalable-approach-to-reducing-gender.html. Accessed: 2021-02-25.

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2020. End-to-end bias mitigation by modelling biases in corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8706–8716, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.769

Os Keyes. 2018. The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW). https://doi.org/10.1145/3274357

Os Keyes, Chandler May, and Annabelle Carrell. 2021. You keep using that word: Ways of thinking about gender in computing research. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW). https://doi.org/10.1145/3449113

Yunsu Kim, Duc Thanh Tran, and Hermann Ney. 2019. When and why is document-level context useful in neural machine translation? In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 24–34, Hong Kong, CN. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-6503

Kris Aric Knisely. 2020. Le français non-binaire: Linguistic forms used by non-binary speakers of French. *Foreign Language Annals*, 53(4):850–876. https://doi.org/10.1111/flan.12500

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the tenth Machine Translation Summit*, pages 79–86, Phuket, TH. AAMT.

Corina Koolen and Andreas van Cranenburgh. 2017. These are not the stereotypes you are looking for: Bias and fairness in authorial gender attribution. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 12–22, Valencia, ES. Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1602

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

Michał Krawczyk. 2017. Are all researchers male? Gender misattributions in citations. *Scientometrics*, 110(3):1397–1402. https://doi.org/10.1007/s11192-016-2192-y, PubMed: 28255187

Hamutal Kreiner, Patrick Sturt, and Simon Garrod. 2008. Processing definitional and stereotypical gender in reference resolution: Evidence from eye-movements. *Journal of Memory and Language*, 58:239–261. https://doi.org/10.1016/j.jml.2007.09.003

James Kuczmarski. 2018. Reducing gender bias in Google Translate. https://www.blog.google/products/translate/reducing-gender-bias-google-translate/. Accessed: 2021-02-25.

William Labov. 1972. *Sociolinguistic Patterns*. 4. University of Pennsylvania Press.

Jennifer Langston. 2020. New AI tools help writers be more clear, concise and inclusive in

Office and across the Web. https://blogs.microsoft.com/ai/microsoft-365-ai-tools/. Accessed: 2021-02-25.

Brian Larson. 2017. Gender as a variable in natural-language processing: Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, ES. Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1601

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

Enora Lessinger. 2020. Le président est une femme: The challenges of translating gender in UN texts. In Luise von Flotow and Hala Kamal, editors, *The Routledge Handbook of Translation, Feminism and Gender*. New York, USA. Routledge. https://doi.org/10.4324/9781315158938-33

Hector J. Levesque. 2014. On our best behaviour. *Artificial Intelligence*, 212(1):27–35. https://doi.org/10.1016/j.artint.2014.03.007

Roger J. R. Levesque. 2011. *Sex Roles and Gender Roles*. Springer, New York, USA. https://doi.org/10.1007/978-1-4419-1695-2_602

Molly Lewis and Gary Lupyan. 2020. Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature Human Behaviour*, 4(10):1021–1028.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. Towards robust and privacy-preserving text representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 25–30, Melbourne, AU. Association for Computational Linguistics.

Daniel J. Liebling, Michal Lahav, Abigail Evans, Aaron Donsbach, Jess Holbrook, Boris Smus, and Lindsey Boran. 2020. Unmet needs and opportunities for mobile translation AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–13, New York, USA. Association for Computing Machinery. https://doi.org/10.1145/3313831.3376261

Anna Lindqvist, Emma A. Renström, and Marie Gustafsson Sendén. 2019. Reducing a male bias in language? Establishing the efficiency of three different gender-fair language strategies. *Sex Roles*, 81(1–2):109–117. https://doi.org/10.1007/s11199-018-0974-9

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV Subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, SI. European Language Resources Association (ELRA).

Katherine A. Liu and Natalie A. Dipietro Mager. 2016. Women's involvement in clinical trials: Historical perspective and future implications. *Pharmacy Practice*, 14(1):708. https://doi.org/10.18549/PharmPract.2016.01.708, PubMed: 27011778

Ártemis López. 2019. Tú, yo, elle y el lenguaje no binario. http://www.lalinternadeltraductor.org/n19/traducir-lenguaje-no-binario.html. Accessed: 2021-02-25.

Ártemis López, Susana Rodríguez Barcia, and María del Carmen Cabeza Pereiro. 2020. Visibilizar o interpretar: Respuesta al Informe de la Real Academia Española sobre el lenguaje inclusivo y cuestiones conexas. http://www.ngenespanol.com/el-mundo/la-rae-rechaza-nuevamente-el-lenguaje-inclusivo/. Accessed: 2021-02-25.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. In *Logic, Language, and Security*, volume 12300 of *Lecture Notes in Computer Science*, pages 189–202, Springer. https://doi.org/10.1007/978-3-030-62077-6_14

John Lyons. 1977. *Semantics*, volume 2, Cambridge University Press, Cambrdige, UK.

Thomas Manzini, Lim Yao Chong, Alan W. Black, and Yulia Tsvetkov. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 615–621, Minneapolis, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1062

Marianna Martindale and Marine Carpuat. 2018. Fluency over adequacy: A pilot study in measuring user trust in imperfect MT. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 13–25, Boston, USA. Association for Machine Translation in the Americas.

Chandler May. 2019. Deconstructing gender prediction in NLP. In *Conference on Neural Information Processing Systems (NIPS) – Keynote*. Vancouver, CA.

Sally McConnell-Ginet. 2013. Gender and its relation to sex: The myth of 'natural' gender. In Greville G. Corbett, editor, *The Expression of Gender*, pages 3–38. De Gruyter Mouton, Berlin, DE.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, IT. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1334

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning.

Shachar Mirkin, Scott Nowson, Caroline Brun, and Julien Perez. 2015. Motivating personality-aware machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1102–1108, Lisbon, PT. Association for Computational Linguistics. https://doi.org/10.18653/v1/D15-1130

Margaret Mitchell, Dylan Baker, Nyalleng Moorosi, Emily Denton, Ben Hutchinson, Alex Hanna, Timnit Gebru, and Jamie Morgenstern. 2020. Diversity and inclusion metrics in subset selection. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, pages 117–123, New York, USA. Association for Computing Machinery.

Britta Mondorf. 2002. Gender differences in English syntax. *Journal of English Linguistics*, 30:158–180. https://doi.org/10.1177/007242030002005

Johanna Monti. 2017. Questioni di genere in traduzione automatica. *Al femminile. Scritti linguistici in onore di Cristina Vallini*, 139:411–431.

Johanna Monti. 2020. Gender issues in machine translation: An unsolved problem? In Luise von Flotow and Hala Kamal, editors, *The Routledge Handbook of Translation, Feminism and Gender*, pages 457–468, Routledge. https://doi.org/10.4324/9781315158938-39

Amit Moryossef, Roee Aharoni, and Yoav Goldberg. 2019. Filling gender & number gaps in neural machine translation with black-box context injection. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 49–54, Florence, IT. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3807

Heiko Motschenbacher. 2014. Grammatical gender as a challenge for language policy: The (im)possibility of non-heteronormative language use in German versus English. *Language policy*, 13(3):243–261. https://doi.org/10.1007/s10993-013-9300-0

Anthony Mulac, James J. Bradac, and Pamela Gibbons. 2001. Empirical support for the gender-as-culture hypothesis. *Human Communication Research*, 27:121–152. https://

doi.org/10.1111/j.1468-2958.2001.tb00778.x

El Mundo. 2018. La RAE rechaza nuevamente el lenguaje inclusivo. https://www.ngenespanol.com/el-mundo/la-rae-rechaza-nuevamente-el-lenguaje-inclusivo/. Accessed: 2021-02-25.

David A. B. Murray. 2003. Who is Takatāpui? Māori language, sexuality and identity in Aotearoa/New Zealand. *Anthropologica*, pages 233–244. https://doi.org/10.2307/25606143

Terttu Nevalainen and Helena Raumolin-Brunberg. 1993. Its strength and the beauty of it: The standardization of the third person neuter possessive in early modern English. In Dieter Stein and Ingrid Tieken-Boon van Ostade, editors, *Towards a Standard English*, pages 171–216. De Gruyter, Berlin, DE.

Matthew L. Newman, Carla J. Groom, Lori D. Handelman, and James W. Pennebaker. 2008. Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, 45(3):211–236. https://doi.org/10.1080/01638530802073712

Dong Nguyen, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. 2016. Computational sociolinguistics: A survey. *Computational linguistics*, 42(3):537–593. https://doi.org/10.1162/COLI_a_00258

Jan Niehues, Eunah Cho, Thanh-Le Ha, and Alex Waibel. 2016. Pre-translation for neural machine translation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1828–1836, Osaka, JP. The COLING 2016 Organizing Committee.

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

Malvina Nissim and Rob van der Goot. 2020. Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, 46(2):487–497. https://doi.org/10.1162/coli_a_00379

Parmy Olson. 2018. The algorithm that helped google translate become sexist. https://www.forbes.com/sites/parmyolson/2018/02/15/the-algorithm-that-helped-google-translate-become-sexist/?sh=d675b9c7daa2. Accessed: 2021-02-25.

Dimitrios Papadimoulis. 2018. *Gender-Neutral Language in the European Parliament*. European Parliament 2018.

Benjamin Papadopoulos. 2019. *Morphological Gender Innovations in Spanish of Genderqueer Speakers*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, USA. Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

Amandalynne Paullada, Inioluwa D. Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2020. Data and its (dis)contents: A survey of dataset development and use in machine learning research. In *NeurIPS 2020 Workshop: ML Retrospectives, Surveys & Meta-analyses (ML-RSA)*. Vitual.

James Pennebaker and Lori Stone. 2003. Words of wisdom: Language use over the life span. *Journal of Personality and Social Psychology*, 85:291–301. https://doi.org/10.1037/0022-3514.85.2.291, PubMed: 12916571

Marcelo O. R. Prates, Pedro H. C. Avelar, and Luís C. Lamb. 2018. Assessing gender bias in machine translation: A case study with Google Translate. *Neural Computing and Applications*, pages 1–19.

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

Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W. Crandall, Nicholas A. Christakis, Iain D. Couzin, Matthew O. Jackson, et al. 2019. Machine behaviour. *Nature*, 568(7753):477–486. https://doi.org/10.1038/s41586-019-1138-y, PubMed: 31019318

Isabelle Régner, Catherine Thinus-Blanc, Agnès Netter, Toni Schmader, and Pascal Huguet. 2019. Committees with implicit biases promote fewer women when they do not believe gender bias exists. *Nature Human Behaviour*, 3(11):1171–1179. https://doi.org/10.1038/s41562-019-0686-3, PubMed: 31451735

Argentina A. Rescigno, Johanna Monti, Andy Way, and Eva Vanmassenhove. 2020. A case study of natural gender phenomena in translation: A Comparison of Google Translate, Bing Microsoft Translator and DeepL for English to Italian, French and Spanish. In *Proceedings of the Workshop on the Impact of Machine Translation (iMpacT 2020)*, pages 62–90, Online. Association for Machine Translation in the Americas.

Alexander S. Rich and Todd M. Gureckis. 2019. Lessons for artificial intelligence from the study of natural stupidity. *Nature Machine Intelligence*, 1(4):174–180. https://doi.org/10.1038/s42256-019-0038-z

Christina Richards, Walter P. Bouman, Leighton Seal, Meg J. Barker, Timo O. Nieder, and Guy TSjoen. 2016. Non-binary or genderqueer genders. *International Review of Psychiatry*, 28(1):95–102. https://doi.org/10.3109/09540261.2015.1106446, PubMed: 26753630

Barbara J. Risman. 2018. Gender as a Social Structure. In Barbara Risman, Carissa Froyum, and William J. Scarborough, editors, *Handbook of the Sociology of Gender*, pages 19–43, Springer. https://doi.org/10.1007/978-3-319-76333-0

Nicholas Roberts, Davis Liang, Graham Neubig, and Zachary C. Lipton. 2020. Decoding and diversity in machine translation. In *Proceedings of the Resistance AI Workshop at 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*. Vancouver, CA.

Suzanne Romaine. 1999. *Communicating Gender*. Lawrence Erlbaum, Mahwah, USA. https://doi.org/10.4324/9781410603852

Suzanne Romaine. 2001. A corpus-based view of gender in British and American English. *Gender across Languages*, 1:153–175. https://doi.org/10.1075/impact.9.12rom

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Ram Samar. 2020. Machines Are Indifferent, We Are Not: Yann LeCuns Tweet Sparks ML bias debate. https://analyticsindiamag.com/yann-lecun-machine-learning-bias-debate/. Accessed: 2021-02-25.

Kalinowsky Santiago. 2018. Todos/Todas/Todes. Interview with Megan Figueroa, host; Carrie Gillon, host. In *The Vocal Fries [Podcast]*. Vancouver, CA.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Danielle Saunders and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.690

Danielle Saunders, Rosie Sallis, and Bill Byrne. 2020. Neural machine translation doesn't

translate gender coreference right unless you make it. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 35–43, Online. Association for Computational Linguistics.

Londa Schiebinger. 2014. Scientific research must take gender into account. *Nature*, 507(9). https://doi.org/10.1038/507009a, PubMed: 24598604

Ari Schlesinger, W. Keith Edwards, and Rebecca E. Grinter. 2017. Intersectional HCI: Engaging identity through gender, race, and class. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 5412–5427, New York, USA. Association for Computing Machinery. https://doi.org/10.1145/3025453.3025766

Natalie Schluter. 2018. The Glass Ceiling in NLP. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2793–2798, Brussels, BE. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1301

Muriel R. Schulz. 1975. The semantic derogation of woman. In Barrie Thorne and Nancy Henley, editors, *Sex and Language. Difference and Dominance*, pages 64–75, Newbury House, Rowley, USA.

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3419–3425, Hong Kong, CN. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1341

Sabine Sczesny, Christa Nater, and Alice H. Eagly. 2018. Agency and communion: Their implications for gender stereotypes and gender identities, *Agency and Communion in Social Psychology*, pages 103–116. Taylor and Francis. https://doi.org/10.4324/9780203703663-9

Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 59–68, New York, USA. Association for Computing Machinery. https://doi.org/10.1145/3287560.3287598

Rico Sennrich. 2017. How grammatical is character-level neural machine translation? Assessing MT quality with contrastive translation pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382, Valencia, ES. Association for Computational Linguistics. https://doi.org/10.18653/v1/E17-2060

Deven S. Shah, Hansen A. Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.468

Alyx J. Shroy. 2016. Innovations in gender-neutral French: Language practices of nonbinary French speakers on Twitter. *Ms., University of California, Davis*.

Jeanette Silveira. 1980. Generic masculine words and thinking. *Women's Studies International Quarterly*, 3(2–3):165–178. https://doi.org/10.1016/S0148-0685(80)92113-2

Fabian H. Sinz, Xaq Pitkow, Jacob Reimer, Matthias Bethge, and Andreas S. Tolias. 2019. Engineering a less artificial intelligence. *Neuron*, 103(6):967–979. https://doi.org/10.1016/j.neuron.2019.08.034, PubMed: 31557461

Janet Smith. 2003. Gendered structures in Japanese. *Gender across Languages*, 3:201–227. https://doi.org/10.1075/impact.11.12shi

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge, USA. The Association for Machine Translation in the Americas.

Artūrs Stafanovičs, Mārcis Pinnis, and Toms Bergmanis. 2020. Mitigating gender bias in machine translation with target gender annotations. In *Proceedings of the Fifth Conference on Machine Translation*, pages 629–638, Online. Association for Computational Linguistics.

Dagmar Stahlberg, Friederike Braun, Lisa Irmen, and Sabine Sczesny. 2007. Representation of the sexes in language. *Social Communication*, pages 163–187.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, IT. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1164

Simone Stumpf, Anicia Peters, Shaowen Bardzell, Margaret Burnett, Daniela Busse, Jessica Cauchard, and Elizabeth Churchill. 2020. Gender-inclusive HCI research and design: A conceptual review. *Foundations and Trends in Human–Computer Interaction*, 13(1):1–69. https://doi.org/10.1561/1100000056

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature Review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, IT. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1159

Tony Sun, Kellie Webster, Apu Shah, William Y. Wang, and Melvin Johnson. 2021. They, them, theirs: Rewriting with gender-neutral English. *arXiv preprint arXiv:2102.06788*.

Harini Suresh and John V. Guttag. 2019. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*.

Masashi Takeshita, Yuki Katsumata, Rafal Rzepka, and Kenji Araki. 2020. Can existing methods debias languages other than english? First attempt to analyze and mitigate Japanese Word Embeddings. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 44–55, Online. Association for Computational Linguistics.

Tina Tallon. 2019. A century of ''shrill'': How bias in technology has hurt women's voices. *The New Yorker*.

Rachael Tatman. 2017. Gender and dialect bias in youtube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, ES. Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1606

Peter Trudgill. 2000. *Sociolinguistics: An Introduction to Language and Society*. Penguin Books, London, UK.

Anne M. Turner, Megumu K. Brownstein, Kate Cole, Hilary Karasz, and Katrin Kirchhoff. 2015. Modeling workflow to design machine translation applications for public health practice. *Journal of Biomedical Informatics*, 53:136–146. https://doi.org/10.1016/j.jbi.2014.10.005, PubMed: 25445922

Amos Tversky and Daniel Kahneman. 1973. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5(2): 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131. https://doi.org/10.1126/science.185.4157.1124, PubMed: 17835457

Emiel van Miltenburg. 2019. *Pragmatic Factors in (Automatic) Image Description*. Ph.D. thesis, Vrije Universiteit, Amsterdam, NL.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting Gender Right in Neural Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, BE. Association for Computational Linguistics.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 222–232, Dublin, IE. European Association for Machine Translation.

Mihaela Vorvoreanu, Lingyi Zhang, Yun-Han Huang, Claudia Hilderbrand, Zoe Steine-Hanson, and Margaret Burnett. 2019. From gender biases to gender-inclusive design: An empirical investigation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–14, New York, USA. Association for Computing Machinery. https://doi.org/10.1145/3290605.3300283

Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2015. It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 9.

Mario Wandruszka. 1969. *Sprachen: Vergleichbar und Vnvergleichlich*, R. Piper & Co. Verlag, Munich, DE.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? Annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-5618

Zeerak Waseem, Smarika Lulz, Joachim Bingel, and Isabelle Augenstein. 2020. Disembodied machine learning: On the illusion of objectivity in NLP. OpenReview Preprint.

Kellie Webster, Marta R. Costa-jussà, Christian Hardmeier, and Will Radford. 2019. Gendered ambiguous pronoun (GAP) shared task at the gender bias in NLP workshop 2019. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 1–7, Florence, IT. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3801

Ilka B. Wolter and Bettina Hannover. 2016. Gender role self-concept at school start and its impact on academic self-concept and performance in mathematics and reading. *European Journal of Developmental Psychology*, 13(6):681–703. https://doi.org/10.1080/17405629.2016.1175343

Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender bias in multilingual embeddings and cross-lingual transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2896–2907, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.260

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, DK. Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1323

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

for Computational Linguistics. https://doi.org/10.18653/v1/N18-2003

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, BE. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1521

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. Examining gender bias in languages with grammatical gender. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5276–5284, Hong Kong, CN. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1531.

Lal Zimman. 2020. Transgender language, transgender moment: Toward a trans linguistics. In Kira Hall and Rusty Barrett, editors, *The Oxford Handbook of Language and Sexuality*, Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190212926.013.45

Lal Zimman, Evan Hazenberg, and Miriam Meyerhoff. 2017. Trans peoples linguistic self-determination and the dialogic nature of identity. *Representing trans: Linguistic, legal and everyday perspectives*, pages 226–248.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, IT. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1161