Type B Reflexivization as an Unambiguous Testbed
for Multilingual Multi-Task Gender Bias

Ana Valeria González[1], **Maria Barrett**[1,2], **Rasmus Hvingelby**[2], **Kellie Webster**[3], and
Anders Søgaard[1,3]

[1]University of Copenhagen, [2]Alexandra Instituttet, [3]Google Research
{ana,mjb,soegaard}@di.ku.dk,
rasmus.hvingelby@alexandra.dk
websterk@google.com

Abstract

The one-sided focus on English in previous studies of gender bias in NLP misses out on opportunities in other languages: English challenge datasets such as GAP and Wino-Gender highlight model preferences that are "hallucinatory", e.g., disambiguating gender-ambiguous occurrences of 'doctor' as male doctors. We show that for languages with type B reflexivization, e.g., Swedish and Russian, we can construct multi-task challenge datasets for detecting gender bias that lead to *unambiguously wrong* model predictions: In these languages, the direct translation of 'the doctor removed his mask' is not ambiguous between a coreferential reading and a disjoint reading. Instead, the coreferential reading requires a non-gendered pronoun, and the gendered, possessive pronouns are *anti-reflexive*. We present a multilingual, multi-task challenge dataset, which spans four languages and four NLP tasks and focuses only on this phenomenon. We find evidence for gender bias across *all* task-language combinations and correlate model bias with national labor market statistics.

1 Introduction

A reflexive pronoun is an anaphor that requires a *c*-commanding antecedent within its binding domain (Chomsky, 1991).[1] In languages with *Type B* reflexivization (Heine, 2005), the referent of a reflexive possessive pronoun has to be the subject of the clause, while non-reflexive possessive pronouns (so-called *anti-reflexives*) trigger an interpretation where its referent is *not* the subject; see Table 1.

We focus on the subset of those languages in which *anti-reflexive possessive pronouns are gen-*

		TYPE A			TYPE B	
Person	1st	2nd	3rd	1st	2nd	3rd
REFL	✓	✓	✓	✓	✓	

Table 1: In Type B reflexivization (Heine, 2005), 3rd person pronouns cannot be used reflexively. We are interested in Type B languages with gendered pronouns, and where the non-gendered special (3rd person) reflexive marker has a possessive form.

dered, but reflexives are not. This includes Chinese, Russian, Danish, and Swedish, as well as other Scandinavian, Slavic, and Sino-Tibetan languages languages (Bílý, 1981; Battistella, 1989; Kiparsky, 2001).[2] Our motivation for highlighting this particular linguistic phenomenon is that the antecedents of reflexive and anti-reflexive pronouns are grammatically determined; if gender bias leads models (or humans) to predict alternative coreference chains, this violates hard grammatical rules and is thus a clear case of gender bias leading not only to 'hallucinations',[3] but to errors. To see this, consider the following examples:

(1) The surgeon₁ put a book on PRON.POSS.REFL.3RD₁ table. → The book is on the surgeon's₁ table.

(2) The surgeon₁ put a book on PRON.POSS.3RD₂ table. ↛ The book is on the surgeon's₁ table.

Examples (1) and (2) should not be thought of as examples of English, but placeholders for sentences

---

[1]This means that the antecedent should be in the same sentence, be different from the pronoun and not command it, but any ancestor of the antecedent is an ancestor of the pronoun. This is why in *Lea₁'s sister₂ taught herself₁*/₂/₃** the pronoun refers to *sister*, not to *Lea* or a discourse referent.

[2]This rules out languages such as German and French, where the reflexive (e.g., *sich* and *se*) does not have a possessive form (Steinbach, 1998). We focus on the reflexive and anti-reflexive *possessive* forms rather than pure reflexives, since they occur more freely, i.e., not only in the context of reflexive verbs, and they are thus more likely to interact with implicit gender assumptions.

[3]We use the word *hallucination* to refer to gender bias leading models to infer gender without evidence; see Tian et al. (2020) for a similar use of the term in abstractive summarization.

in the languages above since this grammatical distinction is not possible in English: the possessive reflexive (PRON.POSS.REFL.3RD) and the possessive anti-reflexive (PRON.POSS.3RD) in these languages would be translated to the same pronoun in English. In Example (1), the reflexive possessive pronoun is co-referential with the grammatical subject (as indicated by subscripts), which leads to the conclusion that the book is now on a table that is associated with the subject, in other words, the *surgeon's* table. In Example (2), in contrast, when an anti-reflexive possessive pronoun is used, this reading is no longer possible. Instead, Example (2) unambiguously means that the book is on someone else's table. This distinction is not possible in English where the same pronoun (*his/her*) would be used in both Examples (1) and (2): *The surgeon put a book on his table*, which is therefore ambiguous between a disjoint and a coreferent reading.

Language users may be *more likely to prefer the ungrammatical reflexive reading if the gender of the anti-reflexive possessive pronoun matches their (possibly gender-stereotypical) expectations about the referent of the subject*, in this case, *the surgeon*. A masculine possessive pronoun aligns with a prevalent stereotype that surgeons are men; although in the US, in reality, only 62% are.[4] Such a reading is, however, clearly not intended, and this is an example of when gender bias prohibits effective communication. Introducing a new referent in a discourse, usually comes at a cognitive cost when processing the sentence if the referent is not already salient (Grosz et al., 1995). While Example (2) is grammatically unambiguous, language users may occasionally be willing to violate grammatical constraints to avoid the more costly non-coreferential reading, if the meaning of the grammatically correct disjoint reading does not align with their expectations about the world.[5]

The challenge dataset that we present here consists of examples such as the one above and is intended as a diagnostic of implicit gender assumptions in NLP models. It is applicable across four languages (Danish, Russian, Swedish, and Chinese) and four NLP tasks: natural language inference (NLI), machine translation (MT), coreference resolution, and language modeling (LM)). We

will, for example, be interested in whether models are more likely to produce errors when the anti-reflexive pronouns – PRON.POSS.3RD in Example (2) – exhibit the gender that is implicitly associated with the entity in the subject position, i.e., *surgeon*. As should be clear by now, the challenge dataset is fundamentally different from previously introduced challenge datasets in that it focuses on a single linguistic phenomenon that exists across many languages (Lødrup et al., 2011; Honselaar, 1986; Cohen, 1973; Stoykova, 2012) and includes **four languages and four tasks**, and because it focuses on gender bias leading to prediction errors rather than 'hallucinations', i.e., unwarranted disambiguations. To the best of our knowledge, the dataset introduced below is in this way the first of its kind.

Contributions We present a multilingual, multi-task challenge dataset focusing on a specific linguistic phenomenon found in some Scandinavian, Slavic, and Sino-Tibetan languages, namely *gendered possessive anti-reflexive pronouns* in combination with non-gendered possessive reflexive pronouns. We show, by designing multilingual example generation templates by hand, how this phenomenon can interact with gender assumptions in interesting ways. This results in a unique challenge dataset, which we use to detect and quantify gender biases in state-of-the-art and off-the-shelf models across several tasks, including machine translation, natural language inference, coreference resolution, and language modeling. Unlike all other previous challenge datasets focusing on gender bias, our examples quantify *to what extent gender bias in models leads to prediction errors*, rather than unwarranted disambiguation. Data and code is available at https://github.com/anavaleriagonzalez/ABC-dataset

2 The Anti-reflexive Bias Challenge

The ANTI-REFLEXIVE BIAS CHALLENGE (ABC) dataset is designed to force humans and models to align with either widespread gender assumptions or hard grammatical rules. Note, again, that this is in sharp contrast with other gender bias challenge datasets, where gender biases lead to biases in semantic disambiguation, but do not interact with grammatical constraints. Our approach is similar to previous work in other respects:

Similarly to Rudinger et al. (2018) and other recent challenge datasets, ABC relies on hand-

---

[4] http://www.bls.gov/cps/cpsaat11.htm

[5] Note that this is *not* a conflict between syntax and semantics, such as, for example, those studied in Kos et al. (2010), but a conflict between syntax, on the one hand, and belief bias and pragmatics.

written templates, which are used to generate examples in conjunction with lists of occupations. We make use of the 60 occupations listed in Caliskan et al. (2017) containing statistics about gender distributions across professions, taken from the U.S. Bureau of Labor Statistics. Specifically, we generate a base set of 4,560 sentences from 38 templates, two tenses (present and past), and 60 occupations. The 38 templates vary the position of the pronouns, e.g.:

(3) The OCCUPATION lost PRON.POSS.3RD wallet at the house.

(4) The OCCUPATION lost the wallet at PRON.POSS.3RD house.

where PRON.POSS.3RD, in this case, is a place holder for anti-reflexive and reflexive third-person pronouns. Our templates only include transitive verbs.

In our language modelling experiments, we predict the pronoun in question. For NLI and coreference, we introduce three variations of each datapoint (possessive masculine, possessive feminine (anti-reflexive) pronouns and the non-gendered reflexive pronoun). This leads to a total of 13,680 examples for each language. For NLI, we use these as premises and add possible entailments to our templates. See Examples (1) and (2). For machine translation, we use the English versions of Examples (3) and (4) as source sentences, with feminine and masculine third-person pronouns. This leads to 9,120 translation problems. Native speakers manually verified and corrected all templates and sample examples for all tasks. Appendix A shows examples from the four tasks in the four languages. We discuss each task in detail below.

NLI Examples (1) and (2) illustrate the entailment phenomenon that we are interested in. Reflexive possessive pronouns are coreferential with their subjects, which leads to the interpretation that the book is on the surgeon's table. Anti-reflexive pronouns, on the other hand, prevent this reading and leads to an interpretation that a new discourse entity – another person – exists and that the book is located on that person's table.

The general form of our inference examples is as follows:

(5) OCCUPATION.DEF$_1$ [VERB PHRASE] PRON.POSS.REFL.3RD$_1$ OBJECT PREP NOUN.DEF. $\rightarrow$ OCCUPATION.DEF.POSS$_1$ OBJECT [VERB PHRASE.PASSIVE] PREP NOUN.DEF.

(6) OCCUPATION.DEF$_1$ [VERB PHRASE] PRON.POSS.3RD$_2$ OBJECT PREP NOUN.DEF. $\nrightarrow$ OCCUPATION.DEF.POSS$_1$ OBJECT [VERB PHRASE.PASSIVE] PREP NOUN.DEF.

We will primarily be interested in the rate at which state-of-the-art NLI models (wrongly) predict examples of the form in Example (5) to be cases of entailment, and how this depends on whether the possessive pronoun PRON.POSS is masculine or feminine. To generate examples of this form, we translate one prototype example and then identify the variables in the output example. We also make sure to check that there are no morpho-syntactic dependencies, e.g., agreement, between these variables. We then generate all possible examples and have native speakers manually verify the correctness of samples of the generated examples.

Machine Translation For machine translation, we are interested in the way that gender assumptions play a role in the resolution of the gendered possessive pronoun in the source language. As an example, when translating the phrase *The doctor put the book on her table*, an English-Danish translation system would likely generate one of the following two options, a reflexive reading and an anti-reflexive one:

(7) Lægen lagde bogen på *sit* bord

doctor.DEF put book.DEF on PRON.POSS.REFL.3RD table

(8) Lægen lagde bogen på *hendes* bord

doctor.DEF put book.DEF on PRON.POSS.3RD table

While ABC focuses on translating *from English*, it holds that similarly, if we translate the Danish sentence *mekanikeren har brug for sine.*REFL *værktøjer til at arbejde*, which uses a gender-neutral reflexive possessive pronoun *sine*, into English, the model will have to choose between two possible, correct translations:

(9) The mechanic needs *his* tools to work

(10) The mechanic needs *her* tools to work

The machine translation section of the ABC dataset consists of translations from English sentences with gendered possessive pronouns into one of the four target languages (Danish, Russian, Swedish, and Chinese). For a single occupation on the list, this would correspond to two English

sentences (masculine and feminine possessive pronoun) per template. We quantify to what extent models translate English source sentences with possessive masculine or feminine pronouns into target sentences with reflexive pronouns.[6]

Coreference Resolution For coreference resolution, we generate variants of our templates in the four target languages with each of the gendered anti-reflexives and the reflexive pronoun. That is , for a sentence such as:

(11) The firefighter placed *her/his* shoes in the closet

we generate the following examples for Danish:

(12) Brandmanden placerede *hendes* sko i skabet (FEM)

(13) Brandmanden placerede *hans* sko i skabet (MASC)

(14) Brandmanden placerede *sine* sko i skabet (REFL)

In Examples (12) and (13), the use of anti-reflexive pronouns *hans* or the femine anti-reflexive *hendes* means the shoes placed in the closet belong to someone other than the firefighter. In our coreference resolution experiments, we are thus interested in how often models wrongly link the anti-reflexive pronouns (*hans/hendes*) to the occupation. Such predictions violate grammatical constraints and are clear examples of gender assumptions overwriting morpho-syntactic evidence.

Language Modelling For language modeling, we are interested in how likely the models are to predict gendered anti-reflexive possessive pronoun when the original sentence contains a reflexive pronoun. In:

(15) Brandmanden placerede *sine* sko i skabet (REFL)

we compute the sentence perplexity replacing the reflexive pronoun *sine* with a feminine

anti-reflexive (*hendes*) or masculine (*hans*) anti-reflexive pronoun. A difference in perplexity reveals a gender bias, and if the model prefers an anti-reflexive reading, this possibly leads to a grammatically incorrect sentence.[7]

3 Experiments

In this work, we are focused on highlighting a linguistic phenomenon that is useful for diagnosing gender bias, therefore we do not focus on an extensive comparison of model architectures; further work would be required to examine more models. We are interested in the gender associations that *existing* models make. Because of this, we take off-the-shelf translation models and language models. As there were not any state-of-the-art models already pre-trained for coreference in the languages of interest, we train a state-of-the-art architecture for coreference resolution on languages where we could obtain data. To be able to evaluate NLI models on the target languages, we fine-tune a pre-trained model for this task.

As previously found in (Rudinger et al., 2018), gender biases in models tended to correlate with labor statistics of the percentage of female in each occupation according to Bureau of Labor Statistics 2015 [8] released with Caliskan et al. (2017). We correlate our findings with these statistics as well as national statistics.

NLI NLI is originally a three-way classification task. Given two sentences; a premise and a hypothesis, the system classifies the relation between them as `entailment,` `contradiction`, or `neutral`. Since ABC is only intended for diagnosing gender bias in off-the-shelf models, and not for training models, we only consider the `entailment` relation. If the premise contains a reflexive pronoun, the true class is `entailment`, and if the premise contains a masculine or feminine pronoun it is not `entailment`.

XNLI (Conneau et al., 2018) is a manual translation of the English NLI data into 15 languages. Chinese and Russian are among them and we benchmark the model on the XNLI test set. Singh et al. (2019) extend the XNLI train set to a wider set of languages, including Danish and Swedish but there is not test set for benchmarking. We use

---

[6]In the context of examples such as Example (9) and (10), using an anti-reflexive pronoun in the target translation may seem more like a hallucination than violating grammatical constraints, and we acknowledge that in machine translation, as well as in language modeling, the difference concerning existing gender bias challenge datasets is less pronounced than with NLI and coreference resolution. Nevertheless, note that the model not only hallucinates a gender attribution, but also co-referentiality, making it relatively simple to construct semantically impossible examples, e.g., *The mechanic needs his tools, but not his own tools*. Furthermore, introducing a new referent without evidence also violates pragmatic economy principles (Grosz et al., 1995; Gardent and Webber, 2001). Google Translate incorrectly translates into a sentence with two reflexive pronouns (violating the semantic principle of bivalence).

[7]See also the footnote above on whether our machine translation examples diagnose model 'hallucinations' or unambiguous prediction errors.

[8]http://www.bls.gov/cps/cpsaat11.htm

cross-lingual language model pre-training (XLI) (Conneau and Lample, 2019), i.e., we fine-tune on English NLI training data. For Chinese and Russian, we use a publicly available implementation[9] of the XLM-15 model (Conneau and Lample, 2019) and fine-tune it using a batch size of 4 and a learning rate of 0.000005 for 35 epochs, which led to the best performance on the XNLI development set. For Danish and Swedish, we use the XLM-100 model, which we fine-tune for 28 epochs.

Machine Translation For machine translation, we evaluate models for English→ {Danish, Russian, Swedish, Chinese} *to assess how often they predict the non-gendered reflexive possessive pronouns when the source possessive pronoun is masculine versus feminine*. For all languages, we report the performance of Google Translate. Additionally, for the languages where an off-the-shelf, near-state-of-the-art system was publicly available, we also report performance. For Chinese, we use the pre-trained models provided by Sennrich et al. (2017)[10] (E-WMT). For Russian, we use the winner system of WMT19 (Ng et al., 2019), which is provided as part of the Fairseq toolkit (F-WMT).[11]

Coreference Resolution For coreference resolution, we are interested in whether the system violates grammatical rules by placing an anti-reflexive possessive pronoun in a cluster. We train coreference resolution models for Chinese and Russian using the model and code of Joshi et al. (2019). For Chinese, we use the Chinese version of Ontonotes as our training data, which is made up of about 1800 documents for training. For Russian, we use the RuCor corpus (Ju et al., 2014), which is small, containing only 181 documents total, but has been used to train coreference models for Russian before (Ju et al., 2014; Sysoev et al., 2017). The task consists of predicting the spans that make up a coreference cluster. We train the model using the hyperparameters specified in the source code [12]. We use a maximum segment length of 128. See Appendix §B for statistics of the coreference resolution datasets used for training. While we do not have coreference resolution systems we can evalu-

ate for Danish and Swedish, we include challenge examples for these languages that can be used to detect bias in future systems for these languages.

Language Modeling For our language modeling experiments, we use the pre-trained BERT masked language modeling architecture (Devlin et al., 2019). We turn pronoun prediction into a Cloze task (Taylor, 1953) where the pronoun is masked and then the probabilities of each possible alternative taken to compute the sentence-level perplexity. We use Chinese BERT (for Chinese) and multilingual BERT for Russian, Danish, and Swedish.[13] The overall perplexities of these models on our challenge examples, are low; again, this is because of the simple vocabulary and constructions used in the examples. We nevertheless see strong gender bias in the language models, especially for Danish and Chinese.

4 Results

Our evaluation results are found in Table 2, with results on Danish (da), Russian (ru), Swedish (sv), and Chinese (zh), and for machine translation (MT), natural language inference (NLI), coreference resolution (COREF), and language modeling (LM).

NLI For NLI, the XLM models generally over-predict entailment for anti-reflexive pronouns. The models perform well on benchmark data, e.g., 0.742 on the Chinese XNLI test set, but much worse (0.330) on our challenge examples. For Chinese and Danish, the models perform slightly better on sentences with masculine anti-reflexive pronouns, whereas they perform slightly better on sentences with feminine anti-reflexives in Russian and Swedish. For all four languages, we see significant negative correlations between relative error increase on sentences with feminine pronouns and the ratio of women in corresponding occupations; see §5 for how a discussion of the statistics. This suggests that the very poor performance numbers on sentences with anti-reflexive pronouns is, in part, the result of gender bias.

Machine translation For machine translation, we also observe strong negative correlations, suggesting gender bias. In the manual analysis of the output translations, we see a very clear pattern that English masculine possessive pronouns

---

[9]https://github.com/facebookresearch/XLM

[10]https://github.com/EdinburghNLP/nematus

[11]https://github.com/pytorch/fairseq/tree/master/examples/wmt19

[12]https://github.com/mandarjoshi90/coref/blob/master/experiments.conf

[13]https://github.com/google-research/bert/blob/master/multilingual.md

Task	Lang	System	Benchmark	ABC	Significance
NLI	da	XLM-100	–	0.380	✓
	ru	XLM-15	0.736	0.370	✓
	sv	XLM-100	–	0.362	✓
	zh	XLM-15	0.742	0.330	✓
MT	da	Google Translate	0.204	0.395	✓
	ru	Google Translate	0.260	0.406	✓
		F-WMT	0.268	0.421	✓
	sv	Google Translate	0.211	0.422	✓
	zh	Google Translate	0.460	0.594	†
		E-WMT	0.360	0.194	✓
COREF	ru	E2ECOREF-BERT	0.602	0.090	†
	zh		0.630	0.600	✓
LM	da	BERT	2.4	11.4	✓
	ru		3.9	13.4	✓
	sv		1.2	11.2	✓
	zh		6.7	22.1	✓

Table 2: GENDER BIAS RESULTS. Performance on benchmarks and ABC. We look at differences in performance between genders for each task and correlate these with labor market statistics. ✓: Pearson's ρ of error Δ on sentences with feminine pronouns and % of women in corresponding occupations significant ($p < 0.01$); see §5 for a discussion of the labor market statistics. †: Systems insensitive to variation in pronouns.

are more likely to translate into reflexive pronouns in the target languages, than feminine possessive pronouns. For Danish, 93.7% of masculine pronouns were translated into reflexives, whereas only 72.9% of feminine pronouns were. For Russian, the two systems were consistent in this respect and both translated 69.3% of masculine pronouns and 18.1% of feminine pronouns into reflexive pronouns. For Swedish, the numbers were 90.0% and 73.1%, respectively. For Chinese, where the reflexive pronoun is used less frequently,[14] the machine translation models only produced a few translations with reflexive pronouns (for masculine source pronouns).

These differences are not reflected in BLEU scores, and in our correlations we correlate the increase in pronoun translation errors for source sentences with feminine pronouns and the ratio of women in the corresponding occupations. In general, our models achieve high BLEU scores on our challenge examples, which are all syntactically simple and use simple, in-vocabulary words.

Coreference Resolution For coreference resolution, we observe clear performance differences between our Chinese and Russian models. This possibly reflects the fact that the Russian model was trained on a very small dataset and is less likely to generalize. For both models, we observe a clear bias towards clustering masculine anti-reflexive pronouns with their grammatical subjects, despite how this violates grammar. The Chinese model, which exhibits a strong gender bias, errs on 17% of sentences with masculine anti-reflexive pronouns, and on 14.6% of sentences with feminines anti-reflexives. For Russian, the differences are small, but note the model is trained on limited data, e.g., 140 documents. Out of around 13,000 examples, the model only predicts clusters for 475 pronouns, and 400 of those are in reflexive case. The remaining 75 are masculine (0 feminine). In other words: we see a similar tendency to Chinese, but since the overall performance is poor, and the model is in general rather insensitive to differences in pronouns, we do not include correlation results.

Language Modeling Also for language modeling, we observe consistent bias when predicting a masculine pronoun in place of a reflexive for all

---

[14]The systems are trained on a combination of traditional and simplified Chinese; the latter variant does not include the reflexive pronoun.

languages. These differences are higher for Chinese and Russian. We are not interested in the model's ability to generate a particular pronoun, the more interesting observation is whether the perplexities for sentences containing masculine possessives are lower than for predicting feminine possessives when forcing the model to predict these in place of a reflexive. Our results show that perplexities are lower for masculine possessives in all languages with the biggest differences of 3.7 sentence perplexity for Russian.

5 Analysis: Biased statistics?

We used occupations from Caliskan et al. (2017) in creating our template data; this database also includes U.S. occupation statistics. In our results in Table 2, however, we rely on national statistics instead, but how much of a bias would it be to rely on the original American statistics? In this section, we explain how we collected the national statistics and show how they strongly correlate with the American statistics, but also that national statistics are slightly better at detecting gender bias:

Our Danish labor market statistics come from Larsen et al. (2016), as well as Statistics Denmark[15] and Bevægelsesregisteret,[16] which is a national database over authorised health staff. Some numbers (paramedic, scientist and receptionist) are based on graduation statistics. The Russian labor market statistics were mostly obtained from the Federal State Statistic Service.[17] For occupations not contained on this website we obtained the numbers from separate sources such as the Center of Fire Statistics (CFS) of International Association of Fire and Rescue Services (CTIF)[18] and the Organisation for Economic Cooperation and Development's statistics website[19]. We obtain most of our Swedish labor market statistics from Statistics Sweden (SCB).[20] We use the most recent statistic from 2017, which considers people aged 16-64 (Eriksson and Nguyen, 2019). For clerk and worker, we found labor market statistics in SCB (2018). For medical jobs, we used member statistics by Swedish Medical Association (SLF) from 2016.[21] Finally, we obtain statistics for China from Na-
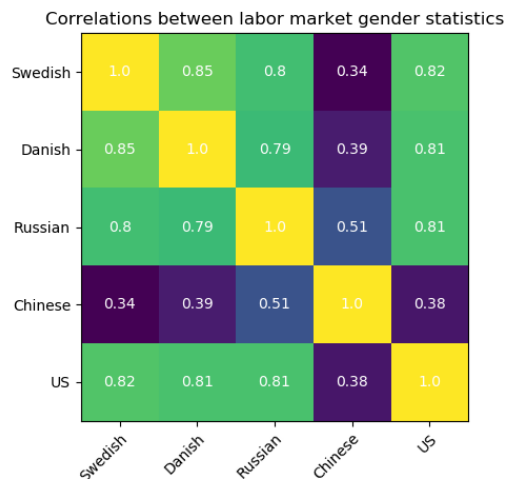

Figure 1: Correlations between collected labor statistics. Numbers > 0.7 are significant ($p < 0.01$).

tional Bureau of Statistics (2004), which is based on census data from 2000.[22]

While labor statistics correlate strongly across countries (Table 1), U.S. statistics are not universal; e.g., almost all pathologist in the U.S. are women (97.5%), whereas the percentage for Denmark is 60%. In the U.S. and Sweden, the painter profession is very male-dominated, like mechanic and electrician (5.70% and 8% women, respectively), whereas in Russia, 57.0% of painters are women.

Correlation Results To assess the potential bias of using U.S. labor market statistics in multilingual experiments, we correlate the gender bias of models for language l with labor statistics from the U.S. and the country in which l is a national language, i.e., we correlate performance differences on Swedish ABC examples with both U.S. and Swedish labor statistics, Danish ABC examples with U.S. and Danish labor statistics, etc. We do so for the subset of occupations, where national gender statistics are available: **NLI** Correlations were stronger with national rather than U.S. statistics for Danish and Swedish (-0.35 vs. -0.28; -0.36 vs. -0.34). **Machine Translation** Correlations were stronger with national rather than U.S. statistics for Russian and Swedish (-0.31 vs. -0.20; -0.31 vs. -0.14). **Coreference Resolution** For coreference,

[22]We did not find reliable gender statistics for all occupations for all countries, but for 51 (Denmark), 50 (Sweden), 38 (Russia), and 10 (China) occupations. One reason was a mismatch between how gender statistics are reported in official reports, including how jobs are grouped. We release the numbers we were able to collect and will continuously work on obtaining more statistics.

we were able to correlate only the results for Chinese due to the fact that the coreference model for Russian only predicted clusters for sentences with male pronouns. The correlations with U.S. and Chinese labor market statistics were not significantly different because we only had statistics for 10 occupations. **Language Modeling** Correlations were stronger with national rather than U.S. statistics on average, but not significantly so.

6   Related Work

The ABC dataset is not first to focus on pronouns and gender bias. The UD English-Pronouns[23] (Munro, 2020), a manually constructed, gender-balanced benchmark of English sentences with pronouns, was motivated by the observation that the genitive pronoun *hers* only occurs three times in the English Universal Dependencies (Nivre et al., 2016). The gendered, ambiguous pronoun (GAP) dataset (Webster et al., 2018) is a coreference resolution dataset of human-annotated ambiguous pronoun-name examples from English Wikipedia. Prates et al. (2018) constructed a translation challenge dataset of simple sentences in gender-neutral languages such as Hungarian and Yoruba and English target sentences such as *he/she is an engineer* to estimate gender biases in machine translation. Both these challenge datasets focus on gender hallucinations, not unambiguous errors induced by gender bias. Some of our examples share similarities with the English WinoGender schema (Rudinger et al., 2018). Consider the following minimal pair of Winograd schema taken from their paper:

(16)   The paramedic$_1$ performed CPR on the passenger$_2$ even though PRON$_1$ knew it was too late.

(17)   The paramedic performed CPR on the passenger$_2$ even though PRON$_2$ was already dead.

In the Winograd schema, the context, i.e., the second clause, is supposed to disambiguate the pronoun on semantic grounds. In Example (16), the pronoun refers to the paramedic, because the patient is unlikely to know whether CPR is too late. In Example (17), the pronoun refers to the patient, because it is impossible to perform CPR if you are dead. Our examples, in contrast, do not disambiguate pronouns on semantic grounds and this is why we are interested in reflexive possessive pronouns: they always refer to the subject, and their anti-reflexive counterparts never do, so there

---

[23]universaldependencies.org/

is no grammatical ambiguity. The disadvantage with semantic disambiguation, we argue, is that it ultimately becomes a subjective competition of belief biases. It is generally impossible to perform CPR if you are dead, but special cases exist:

(18)   Dr Jones$_1$ has turned into a zombie! He$_1$ performed CPR on the passenger even though he$_1$ was already dead.

The ABC dataset evaluates to what extent gender bias leads to unambiguous NLP errors not based on semantic grounds. Finally, Zhao et al. (2018) also include English examples with reflexive pronouns that can be resolved on syntactic grounds, such as:

(19)   The secretary called the physician and told *him* about a new patient.

This construction, however, is less interesting than the reflexive possessive pronominal construction, since in this case, pronouns are always co-referential with the object position, regardless of the pronoun. In sum, the ABC challenge dataset is, to the best of our knowledge, the first dataset to focus on cases where gender bias leads to unambiguous errors; it is also the first multilingual, multi-task gender bias challenge dataset, and the first to focus on anti-reflexive pronouns.

7   Conclusion

In this work we have introduced the Anti-reflexive Bias Challenge (ABC) dataset for multilingual, multi-task gender bias detection, the first of its kind, including four languages and four tasks: machine translation, natural language inference, coreference resolution and masked language modeling. The ABC dataset focuses on a specific linguistic phenomenon that does not occur in English but is found in languages with *Type B reflexivization*: namely, anti-reflexive gendered pronouns. This phenomenon is shown to be useful for exposing unambiguous gender bias, because it quantifies to what extent gender bias leads to prediction *errors*, in contrast to just *unwarranted disambiguations* ('hallucinations'). The problem of anti-reflexive gendered pronouns has, to the best of our knowledge, not received attention before in the NLP literature, which tends to focus heavily on English (Bender and Friedman, 2018). Our evaluations of state-of-the-art models across the four tasks generally reveal significant gender biases leading to false predictions. Additionally, we find that for some

tasks, these associations are more in line with national labor market gender statistics than with U.S. statistics, revealing another way that anglocentric biases can prohibit the detection of gender biases in our models.

Acknowledgements

References

Edwin Battistella. 1989. Chinese reflexivization: a movement to infl approach. *Linguistics*, 27:987–1012.

Emily Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *TACL*, 6:587–604.

Milan Bílý. 1981. *Intrasentential pronominalization and functional sentence perspective*. Lund Slavonic Monographs.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Noam Chomsky. 1991. *Lectures on Government and Binding*. Mouton de Gruyter.

David Cohen. 1973. Hindi'apnaa': A problem in reference assignment. *Foundations of Language*, 10(3):399–408.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7057–7067.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Linn Eriksson and Nicole Nguyen. 2019. The swedish occupational register 2017. *Yrkesstrukturen i Sverige*.

Claire Gardent and Bonnie Webber. 2001. Towards the use of automated reasoning in discourse disambiguation. *Journal of Logic, Language and Information*, 10.

Barbara Grosz, Aravind Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21.

Bernd Heine. 2005. On reflexive forms in creoles. *Lingua*, 115:201–257.

Wim Honselaar. 1986. Reflections on the russian reflexive possessive pronoun . *Studies in Slavic and General Linguistics*, 8:235–248.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2019. SpanBERT: Improving pre-training by representing and predicting spans. *arXiv preprint arXiv:1907.10529*.

Toldova S Ju, A Roytberg, AA Ladygina, MD Vasilyeva, IL Azerkovich, M Kurzukov, G Sim, DV Gorshkov, A Ivanova, A Nedoluzhko, et al. 2014. Ru-eval-2014: Evaluating anaphora and coreference resolution for russian. In *Computational linguistics and intellectual technologies: Proceedings of the international conference "Dialogue*, pages 681–694.

Paul Kiparsky. 2001. Disjoint reference and the typology of pronouns. In *More than Words. A Festschrift for Dieter Wunderlich*.

Miriam Kos, Theo Vosse, Daniëlle van den Brink, and Peter Hagoort. 2010. About edible restaurants: Conflicts between syntax and semantics as revealed by erps. *Frontiers in Psychology*, 1(222).

M Larsen, H Holt, and MR Larsen. 2016. 15:16 et kønsopdelt arbejdsmarked: Udviklingstræk, konsekvenser og forklaringer. *SFI: Det Nationale Forskningscenter for Velfærd*, 170.

Helge Lødrup, Miriam Butt, and Tracy Holloway King. 2011. Norwegian possessive pronouns: Phrases, words or suffixes. In *Proceedings of the LFG11 Conference*, pages 339–359. Oslo: CSLI Publications.

Robert Munro. 2020. *Human-in-the-Loop Machine Learning*. Manning.

National Bureau of Statistics. 2004. *Women and Men in China: Facts and figures*.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proc. of WMT*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Marcelo OR Prates, Pedro H 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.

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.

Statistics Sweden SCB. 2018. *Women and Men in Sweden: Facts and figures*.

Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. 2017. The University of Edinburgh's neural MT systems for WMT17. In *Proceedings of the Second Conference on Machine Translation*, pages 389–399, Copenhagen, Denmark. Association for Computational Linguistics.

Jasdeep Singh, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Bert is not an interlingua and the bias of tokenization. In *EMNLP*.

Markus Steinbach. 1998. *Middles in German*. Ph.D. thesis, Humboldt University.

Velislava Stoykova. 2012. The inflectional morphology of bulgarian possessive and reflexive-possessive pronouns in universal networking language. *Procedia Technology*, 1:400–406.

A Sysoev, I Andrianov, and A Khadzhiiskaia. 2017. Coreference resolution in russian: State-of-the-art approaches application and evolvement. In *Proceedings of International Conference Dialogue-2017*, pages 327–338.

Wilson L Taylor. 1953. "Cloze procedure": A new tool for measuring readability. *Journalism Bulletin*, 30(4):415–433.

Ran Tian, Shashi Narayan, Thibault Sellam, and Ankur Parikh. 2020. Sticking to the facts: Confident decoding for faithful data-to-text generation. In *Arxiv*, page 1910.08684.

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

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

A   Example Data

We present example data for each task.

A.1   NLI

For each occupation, template and language in our NLI benchmark, we have three premise-hypothesis pairs. The hypotheses are identical. For the premises, we have three versions, one for each pronoun (the non-gendered reflexive possessive pronoun and and the two gendered, anti-reflexive possessive pronouns). Examples with reflexives are labeled `entailment`; examples with anti-reflexives are labeled `neutral`. An example datapoint for this task, with the label `entailment`, is shown in Table 3.

NATURAL LANGUAGE INFERENCE	
A) PREMISE AND B) HYPOTHESIS	
da	a) Ingeniøren tager *sine* briller fra skrivebordet b) Ingeniørens briller tages fra skrivebordet
sv	a) Ingenjören tar *sina* glasögon från skrivbordet b) Ingenjörens glasögon tas bort från skrivbordet.
ru	a) Inzhener snimayet *svoi* ochki so stola. b) Ochki inzhener snyaty so stola.
zh	a) Gōngchéngshī cóng zhuōzi shàng zhāi xià *zìjǐ* de yǎnjìng b) Gōngchéngshī de yǎnjìng yǐ bèi cóng zhuōzi shàng yí kāi.
ENGLISH TRANSLATION	
en	a) The engineer takes *his/her* glasses from the desk b) The engineer's glasses are taken from the desk

Table 3: Example data for NLI. For NLI, we only generate entailments and neutral statements. The English translation is shown for reference only.

A.2   Machine Translation

For machine translation, we have 4560 pairs of source sentences with masculine (*his*) and feminine possessive pronouns (*her*), respectively. We translate these into the target languages using off-the-shelf models and assess the tendency of these models to predict reflexive possessive pronouns in the target languages, instead of anti-reflexive possessive pronouns. An example of the data can be found in Table 4.

MACHINE TRANSLATION	
SOURCE SENTENCE	
en	The engineer takes *his/her* glasses from the desk
TRANSLATIONS	
da	Ingeniøren tager *sine* briller fra skrivebordet
sv	Ingenjören tar *sina* glasögon från skrivbordet
ru	Inzhener snimayet *svoi* ochki so stola.
zh	Gōngchéngshī cóng zhuōzi shàng zhāi xià *zìjǐ* de yǎnjìng

Table 4: Example data for machine translation.

A.3   Coreference Resolution

For coreference resolution, we are interested in whether the model is more likely to cluster a masculine possessive pronoun with the subject of the sentence than a feminine pronoun, even when this reading violates grammatical constraints. In Table 5, we list examples of how the task data would look. In brackets, we have mentions of an entity that can be clustered together by the system as belonging to the same coreference chain.

A.4   Language Modeling

For language modeling, we take a sentence containing a reflexive pronoun and swap the reflexive for the possessive masculine and feminine anti-reflexives; we then compute the perplexities of the original and perturbed sentences. Example of how this is framed can be found in Table 6.

B   Coreference Dataset Statistics

In table 7 we show the number of documents used to train each system. For Chinese, the data is available with predefined train, development and test sets. For Russian, however, this is not specified, therefore we split the data 80-20-20.

COREFERENCE RESOLUTION	
da	[Ingeniøren] tager *[sine/hans/hendes]* briller fra skrivebordet
sv	[Ingenjören] tar *[sina/hans/hennes]* glasögon från skrivbordet
ru	[Inzhener] snimayet *[svoi/yego/yeye]* ochki so stola.
zh	[Gōngchéngshī] cóng zhuōzi shàng zhāi xià *[zìjǐ/tā/tā]* de yǎnjìng
ENGLISH TRANSLATION	
en	[The engineer] takes *[his/her]* glasses from the desk

Table 5: Example data for coreference resolution. In brackets, we have the mentions that the system could cluster as coreferent. We include an English translation only for reference.

LANGUAGE MODELING	
da	**Truth**: Ingeniøren tager *sine* briller fra skrivebordet **Prediction(Fem)**: Ingeniøren tager *hendes* briller fra skrivebordet **Prediction(Masc)**: Ingeniøren tager *hans* briller fra skrivebordet
sv	**Truth**: ingenjören tar *sina* glasögon från skrivbordet **Prediction(Fem)**: ingenjören tar *hennes* glasögon från skrivbordet **Prediction(Masc)**: Ingenjören tar *hans* glasögon från skrivbordet
ru	**Truth**: Inzhener snimayet *svoi* ochki so stola. **Prediction(Fem)**: Inzhener snimayet *yeye* ochki so stola. **Prediction(Masc)**: Inzhener snimayet *yego* ochki so stola.
zh	**Truth**: Gōngchéngshī cóng zhuōzi shàng zhāi xià *zìjǐ* de yǎnjìng **Prediction(Fem)**: Gōngchéngshī cóng zhuōzi shàng zhāi xià *tā* de yǎnjìng **Prediction(Masc)**: Gōngchéngshī cóng zhuōzi shàng zhāi xià *tā* de yǎnjìng

Table 6: Example data for the language modeling task

Lang	Training	Dev	Test
zh	1810	252	218
ru	144	18	18

Table 7: Statistics for the coreference data used for training.