Investigating Failures of Automatic Translation in the Case of Unambiguous Gender

Adithya Renduchintala
Facebook AI
adirendu@fb.com

Adina Williams
Facebook AI Research
adinawilliams@fb.com

Abstract

Transformer-based models are the modern work horses for neural machine translation (NMT), reaching state of the art across several benchmarks. Despite their impressive accuracy, we observe a systemic and rudimentary class of errors made by current state-of-the-art NMT models with regards to translating from a language that doesn't mark gender on nouns into others that do. We find that even when the surrounding context provides unambiguous evidence of the appropriate grammatical gender marking, no tested model was able to accurately gender occupation nouns systematically. We release an evaluation scheme and dataset for measuring the ability of NMT models to translate gender morphology correctly in unambiguous contexts across syntactically diverse sentences. Our dataset translates from an English source into 20 languages from several different language families. With the availability of this dataset, our hope is that the NMT community can iterate on solutions for this class of especially egregious errors.

1 Introduction

Neural machine translation models are trained on vast amounts of data and consistently attain strong performance on standard benchmarks (Barrault et al., 2020). Despite this impressive achievement, state-of-the-art MT models are often largely unable to make basic deductions regarding how to correctly inflect nouns with grammatical gender. Previous work measured gender bias by determining how often models translated pronouns coreferent with stereotypical occupation noun stereotypically (e.g., Stanovsky et al. 2019; Prates et al. 2019). Crucially, in this ambiguous setting, the "correct" gender was genuinely under-determined given the context, which allowed for investigating the underlying (often stereotypical) "assumptions" of machine translation models (i.e., that most if not all *nurses* are women). However, gender mistakes in translation go beyond stereotyping: in some cases, assigning the wrong gender to a noun can result in a genuine mistranslation (i.e., a factual error). In this work, we cast the task of measuring gender bias in machine translation as the task of measuring gender *errors* in translation (as opposed to the prevalence of stereotyping in translation). We argue that operationalizing the gender-bias measurement problem with an unambiguous task is much clearer than framing it as an ambiguous task, because, in our setup, morphological gender mistakes are not forgivable.

We introduce a novel unambiguous benchmark dataset that measures whether an MT model can appropriately inflect occupation nouns for gender when translating from an English source into 20 gender-marking target languages. We craft source sentences by manipulating the context of the occupation noun so that the gender of the person referred to (i.e., their gender identity) is clearly specified. For example: *My nurse is a good father* the gender identity of the nurse is unambiguous, because *nurse* is coreferent with *father*. When translating into a target, the occupation noun (*nurse*) requires *masculine* gender marking.

To also enable stereotype measurement within our unambiguous translation task, we vary the gender stereotypicality of occupations (e.g., *nurses* are stereotypically likely to be women while *janitors* are more likely to be men) to determine whether a model's propensity to stereotype contributes to its translation mistakes. Furthermore, we augment our sentences with gender stereotypical adjectives (such as *pretty* and *handsome*, the former being used more frequently in practice to modify nouns referring to women and the latter, to men) to additionally study whether there might be possible interactions between contextual cues, as it is well known that translation systems perform better when provided with more context (i.e., longer sentences; Tiedemann and Scherrer 2017; Miculicich et al. 2018). We expect the incidence of correct inflection

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 3454 - 3469

to rise in cases when a stereotypical contextual cue is also provided. It is our hope that the benchmark will more clearly surface these kinds of errors to the wider NMT community, encouraging us to devise better, targeted mitigation strategies.

Our contributions are as follows: We offer a new unambiguous benchmark to measure MT models' ability to mark gender correctly in 20 target languages (Belarusian, Catalan, Czech, German, Greek, Spanish, French, Hebrew, Croatian, Italian, Latvian, Lithuanian, Polish, Portuguese, Romanian, Russian, Serbian, Ukranian, Urdu) translated from an English source.[1] We find that all tested NMT models reach fairly low accuracy across target languages—at best approximately 70 (Portuguese and German) and at worst below 50 (Urdu). The tested models do better when the trigger refers to a man (e.g., *father*) than when it refers to a woman (e.g., *mother*), and have higher accuracy when the stereotypical gender of the occupation (e.g., *nurse*) matches the gender of the unambiguous trigger (e.g., *mother*), compared to examples for which they don't match (*nurse* and *father*). When we see such blatant translation failures for morphological features as frequent as grammatical gender (which has clear social consequences and strong community buy-in), it becomes very clear that more work is needed to teach our models how to correctly translate morphological information.

2 Methods

Our method crucially relies upon linguistic theory to engineer the context and arrive at unambiguous examples. In most attempts to measure gender bias in NMT, there has been no ground-truth "correct translation"—model "preferences" (Stanovsky et al., 2019; Prates et al., 2019) are reflected by the percentage of examples for which the MT system chooses the gender-stereotypical pronoun as opposed to the anti-gender-stereotypical one. However, since both translations are practically possible in reality (for example, janitors come in all genders), we feel this setting might be overly optimistic about the capabilities of current models.

Our set up has two main components: we have a "trigger" (i.e., a noun or pronoun in the source sentence that unambiguously refers to a person with a particular known gender), and we have an occupation noun which isn't marked for gender in the source language and can be marked with various genders in the target language. We call the former class "triggers" because they are the unambiguous signal which *triggers* a particular grammatical gender marking on the occupation noun. Triggers comprise all "standard" American English pronouns that inflect for gender, and explicitly gendered kinship terms, which were chosen because they are very common concepts cross-linguistically and are gender unambiguous.[2] Occupation nouns were drawn from the U.S. Bureau of Labor Statistics,[3] following Caliskan et al. (2017); Rudinger et al. (2017); Zhao et al. (2018); Prates et al. (2019). We ensure that there is an equal number of triggers and occupation words, so that our benchmark is gender-balanced for binary gender. For a list, see Table 2 and Table 5 in the Appendix.

Source/Target		Label
Src:	My sister is a carpenter$_4$.	Correct
Tgt:	Mi hermana es carpenteria(f)$_4$.	
Src:	That nurse$_1$ is a funny man .	Wrong
Tgt:	Esa enfermera(f)$_1$ es un tipo gracioso .	
Src:	The engineer$_1$ is her emotional mother .	Inconclusive
Tgt:	La ingeniería(?)$_1$ es su madre emocional .	

Table 1: Examples of source-translation pairs. The gender-tags are shown in parenthesis and word-alignments indicated with subscript. *ingeniería* is listed as inconclusive because it is translated as "engineering" and is thus not correctly gendered.

We measure accuracy based on the inflection of the occupation noun, which depends on the syntactic structure of the sentence. To ensure that we have unambiguous sentences, we constructed a short English phrase structure grammar comprising 82 commands to construct our corpus. Previous datasets for measuring gender failures in translation have had a handful unambiguous examples (Stanovsky et al., 2019), but not enough to derive strong conclusions based on unambigous examples alone. Our dataset is unique in having *only* unambiguous examples and having them for a large set of target languages (see also González et al. 2020). We also make use of Binding Theory (Chomsky, 1980, 1981; Büring,

[1] https://github.com/arendu/Unambiguous-gender-bias

[2] Gender identity is not strictly binary. We adopt a binary conception here, because none of our investigated languages grammatically mark genders other than masculine or feminine on occupation nouns. Our gendered trigger words are largely "unambiguous" modulo costume party examples (Ackerman, 2019), where people dress up contra their gender identity: if a man dresses up as his own grandmother, he can be referred to with so-called "unambiguous" triggers such as *grandma* or *she*. We have ensured that our dataset is free from such examples.

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

2005) to ensure that (i) all of our pronoun triggers (both pronominals like *she* and anaphors like *herself*) are strictly coreferring with the occupations and (ii) that no other interpretations are possible.[4]

Having a grammar is useful, since it allows for an increased diversity of source sentences and better control over the context. We will release three grammars which create datasets of three sizes for convenience: extra small (1,536 sentences), small (59, 520 sentences), and extra large (1, 800, 006 sentences). We mainly focus on the extra large dataset (which is a proper superset of the others) for the purposes of the paper. A grammar also allowed us to investigate a couple subsidiary questions about the nature of anaphoric relations: for example, does accuracy depend on whether the occupation precedes or follows the trigger? Moreover, when we include a contextual cue that is predictive of the gender required by the trigger (e.g., *handsome* for *brother*), does accuracy change when we attach it to the occupation (e.g., *that handsome nurse is my brother*) instead of to the trigger (*that nurse is my handsome brother*)? And finally, to what extent do these different syntactic factors interact with each other or vary across languages?

Since we anticipated poor performance on the task, we also devised an easier scenario, where we provide additional contextual cues provided by adjectives about the gender of the relevant entity. Our list of adjectives is the union of single word stereotyped traits drawn from several works in the social psychology literature on gender stereotyping (Bem, 1981; Prentice and Carranza, 2002; Haines et al., 2016; Eagly et al., 2020; Saucier and Iurino, 2020), where they were normed for English.

2.1 Models

We evaluate gendered translation of three pretrained open-source models, (i) *OPUS-MT* is a collection of 1000+ bilingual and multilingual (for certain translation directions) models (Tiedemann and Thottingal, 2020). The architecture of each model was based on a standard transformer (Vaswani et al., 2017) setup with 6 self-attentive layers in both the encoder and decoder network with 8 attention heads in each layer. (ii) *M2M-100* is a large multilingual model which supports "many-to-many" translation directions (Fan et al., 2020). M2M-100 pretrained models are available in three sizes (418 million parameters, 1.2 billion parameters and 15 billion parameters). We employ the small and medium sized models for our experiments which are based on the transformer architecture with 12 encoder and decoder layers and 16 attention heads. (iii) *mBART-50* is another multilingual model (Tang et al., 2020) that is obtained by "many-to-many" direction fine-tuning of a seed mBART denoising auto-encoder model (Liu et al., 2020). The "many–to-many" fine-tuning process is reported to improve multilingual translation by 1 BLEU point, averaged across all translation directions. The mBART-50 models are also based on transformers with 12 encoder and decoder layers with 16 attention heads.

Type	F	M
Trigger	she, her, hers, herself, sister , mother, aunt, grandmother, daughter, niece, wife , girlfriend	he, him, his, himself, brother, father, uncle, grandfather, son, nephew, husband, boyfriend
Occupation	editor, accountant, auditor, attendant, assistant, designer, writer, baker, clerk, cashier, counselor, librarian, teacher, cleaner, housekeeper, nurse, receptionist, hairdresser, secretary	engineer, physician, plumber, carpenter, laborer, driver, sheriff, mover, developer, farmer, guard, chief, janitor, lawyer, CEO, analyst, manager, supervisor, salesperson

Table 2: A sample of words from our dataset. Accuracy is measured on the gender-stereotypical Occupation word. The Trigger provides unambiguous gender information. Co-reference between the two is obligatory.

2.2 Evaluation

To ascertain whether the translation applied the correct morphological marker on the target-side occupation noun, we design a "reference-free" evaluation scheme. Following Stanovsky et al. (2019), we extract token-alignments between the source occupation noun token and its translation in the target side. We also extract morphological features for every token in the target sequence, using a morphological tagger. Thus, we can ascertain the gender associated with the translated occupation noun (as judged by the morphological tagger) and measure the NMT models' accuracy concerning gender translation. We use Dou and

[4]Consider the sentence *Carlotta's dog accompanies her to kindergarten* (Büring, 2005, p.5). In this sentence, we can interpret this sentence as meaning that the dog accompanies either Carlotta or another woman or girl to kindergarten—to strengthen this reading you can append to the front of the sentence the clause something like *whenever Mary's parents have to go to work early, Carlotta's dog accompanies her to kindergarten*. In this way, *her* can refer to either Carlotta or to Mary. We have avoided such ambiguity in our dataset.

Language	M2M (1.2B)			M2M (418M)			mBART-50			OPUS		
	Correct	Wrong	N/A	Correct	Wrong	N/A	Correct	Wrong	N/A	Correct	Wrong	N/A
be	0.47	0.31	0.21	0.39	0.28	0.33						
ca	0.57	0.22	0.22	0.43	0.32	0.25				0.43	0.39	0.19
cs	0.67	0.29	0.04	0.56	0.38	0.06	0.68	0.32	0.01	0.63	0.36	0.01
de	0.73	0.26	0.01	0.54	0.45	0.02	0.61	0.37	0.02	0.61	0.38	0.01
el	0.59	0.35	0.06	0.51	0.37	0.12				0.59	0.39	0.02
es	0.63	0.20	0.17	0.44	0.37	0.18	0.53	0.26	0.22	0.52	0.31	0.17
fr	0.61	0.28	0.11	0.47	0.38	0.15	0.60	0.39	0.01	0.57	0.41	0.02
he	0.57	0.31	0.12	0.51	0.37	0.11	0.57	0.31	0.12	0.55	0.34	0.11
hi	0.51	0.37	0.12	0.49	0.40	0.11	0.49	0.39	0.12			
hr	0.65	0.29	0.05	0.55	0.39	0.07	0.68	0.29	0.03			
it	0.53	0.25	0.22	0.41	0.34	0.24	0.47	0.32	0.21	0.41	0.33	0.26
lt	0.65	0.33	0.02	0.55	0.42	0.03	0.53	0.43	0.04			
lv	0.63	0.35	0.02	0.53	0.44	0.03	0.63	0.33	0.04			
pl	0.65	0.33	0.03	0.54	0.43	0.03	0.59	0.39	0.02			
pt	0.74	0.24	0.02	0.56	0.41	0.03	0.68	0.31	0.02			
ro	0.59	0.33	0.08	0.51	0.41	0.07	0.62	0.32	0.06	0.53	0.40	0.07
ru	0.60	0.38	0.02	0.54	0.42	0.04	0.54	0.36	0.09	0.53	0.47	0.01
sr	0.52	0.43	0.05	0.49	0.44	0.07						
uk	0.59	0.37	0.04	0.51	0.42	0.07	0.67	0.31	0.03	0.56	0.41	0.03
ur	0.44	0.34	0.22	0.44	0.38	0.18	0.42	0.41	0.17			

Table 3: Accuracy for all languages and models. Correct and Wrong indicate the proportion of target sentences wherein the gender marking for the occupation was Correct or Wrong, while N/A represents inconclusive sentences.

Neubig (2021) for word-alignment and Qi et al. (2020) as our morphological tagger. Note that our evaluation scheme only checks if the appropriate gender marking is applied on the occupation noun and does not check if the occupation noun itself has been translated correctly. Thus, we do not prescribe our evaluation scheme as a replacement for traditional MT evaluation using BLEU or chrF++ scores (Papineni et al., 2002; Popović, 2015).

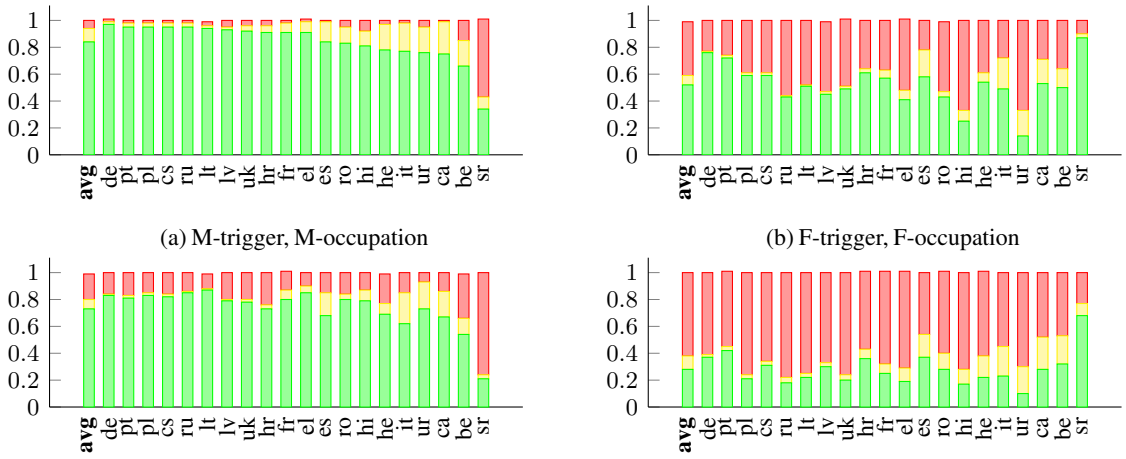
Under our evaluation scheme, there are three possible evaluation outcomes for each sentence. We deem the output (i) *correct* if the gender of the target-side occupation noun is the expected gender (based on the source-side trigger gender). (ii) *wrong* if the gender of the target-side occupation is *explicitly* the wrong gender, and (iii) *inconclusive* if we are unable to make a gender-determination of the target-side occupation noun. A translation can be inconclusive if there are errors in the translation, word-alignments, or morphological tags. In most cases with an inconclusive result, translation errors are the root cause (see Table 1). If errors predominate more for one gender, this itself can be taken as evidence of an imbalance that needs rectification. Note that some of the target languages present for M2M models were not present for mBART and OPUS models—when those models were not trained to translate into a particular target, cells for those languages are left blank in our results tables.

Language	M2M (1.2B)		M2M (418M)		mBART-50		OPUS-MT	
	ΔM	ΔF	ΔM	ΔF	ΔM	ΔF	ΔM	ΔF
be	0.15	**0.23**	0.08	0.08				
ca	0.11	**0.22**	0.06	**0.13**			0.19	**0.27**
cs	0.19	**0.35**	0.03	**0.11**	0.15	**0.23**	0.16	0.19
de	0.13	**0.31**	0.03	**0.15**	0.07	**0.20**	0.15	**0.20**
el	0.07	**0.17**	0.00	**0.08**			0.17	**0.22**
es	0.12	**0.27**	0.08	0.11	0.17	0.19	0.14	0.17
fr	0.12	**0.26**	0.05	**0.14**	0.08	**0.23**	0.14	**0.23**
he	-0.02	**0.28**	-0.11	**0.15**	-0.02	**0.12**	-0.05	**0.33**
hi	-0.02	0.01	0.02	0.04	0.05	**0.14**		
hr	0.19	**0.27**	0.14	**0.28**	0.16	**0.21**		
it	0.12	**0.23**	0.12	**0.20**	0.24	0.22	0.13	**0.25**
lt	0.04	**0.16**	-0.04	**0.09**	0.03	**0.16**		
lv	0.15	**0.18**	0.06	0.10	0.15	**0.27**		
pl	0.18	**0.35**	0.04	**0.15**	0.18	**0.30**		
pt	0.13	**0.21**	0.02	**0.16**	0.11	0.15		
ro	0.10	**0.23**	0.04	**0.15**	0.06	0.09	0.11	**0.25**
ru	0.15	**0.26**	0.06	**0.14**	0.05	**0.18**	0.19	**0.30**
sr	0.18	**0.22**	0.19	0.23				
uk	0.15	**0.29**	0.06	**0.16**	0.13	**0.29**	0.11	**0.24**
ur	-0.02	0.01	0.04	0.03	0.06	0.00		

Table 4: Accuracy drop (Δ) is larger for F-triggers when the occupation is man-stereotypic (M) than for M-triggers when the occupation is woman-stereotypic (F). Bold marks the delta with larger accuracy drop ($>5\%$).

3 Results

Our dataset is very difficult for current models. We observe that accuracy doesn't exceed the low 70s for any language or model (see Table 3). This shows that our dataset is appreciably difficult, and can provide good signal about the failures of our current best models. We additionally find, expectedly, that the larger M2M model outperforms its smaller counterpart (for all languages except Urdu, where performance is comparable). Across the board, M2M with 1.2B parameters slightly outperforms mBART-50, and vastly outperforms the small M2M model with 418M parameters and the OPUS models.



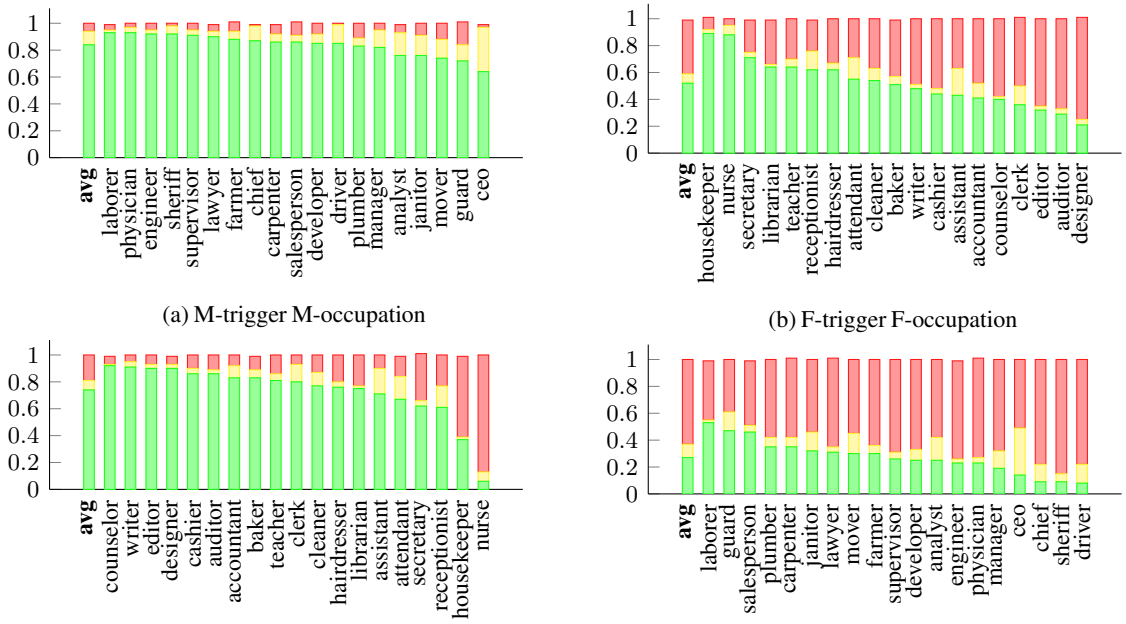
(a) M-trigger, M-occupation

(b) F-trigger, F-occupation

(c) M-trigger, F-occupation

(d) F-trigger, M-occupation

Figure 1: Proportion of correct (green), incorrect (red) and inconclusive (yellow) are provided for each language analyzed. Across the board, for all languages, gender inflection (green) are more correct for masculine triggers, MM (Figure 1a) and MF (Figure 1c) than feminine triggers FF (Figure 1b) and FM (Figure 1d). Accuracy is high for both masculine- and feminine-triggers when the the occupation is indicative of the target gender (Figures 1a and 1b) than when it isn't (Figures 1c and 1d). However, accuracy falls more for F-triggers than for M-triggers when target occupation is indicative of the mismatched gender.



(a) M-trigger M-occupation

(b) F-trigger F-occupation

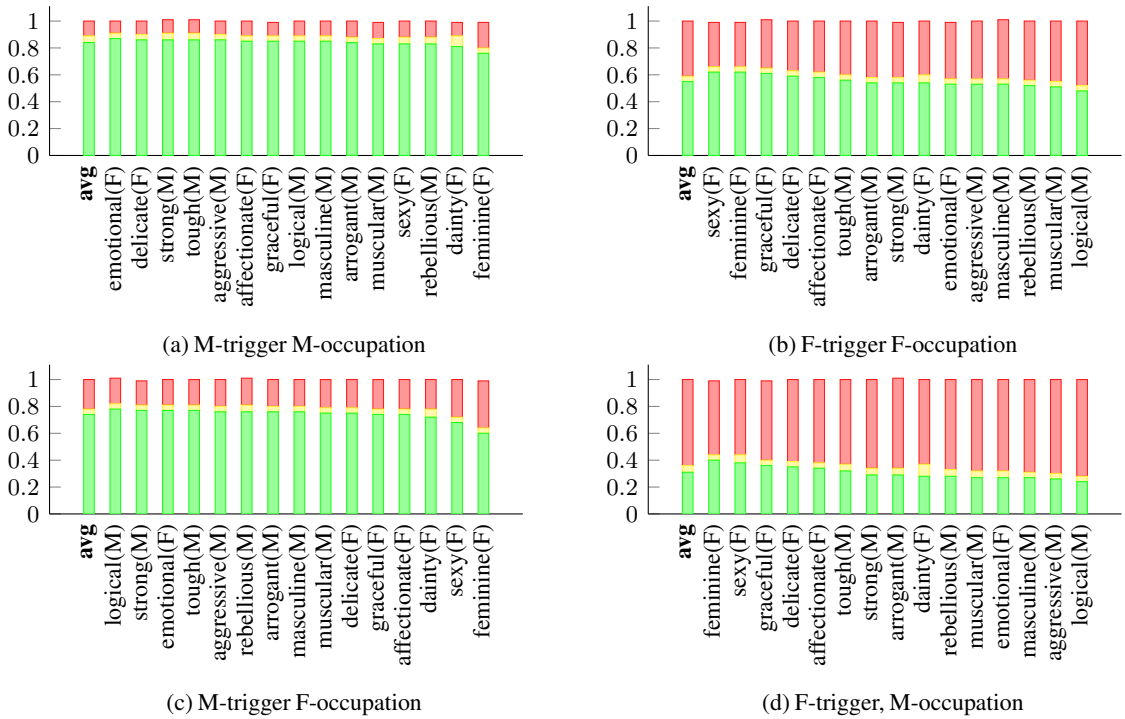
(c) M-trigger F-occupation

(d) F-trigger, M-occupation

Figure 2: Results of accuracy for each occupation noun. For all occupations, accuracy is higher when triggered gender matches the stereotypical gender of the occupation (Figures 2a and 2b), than when it mismatches (Figures 2c and 2d). Accuracy is higher for masculine triggers (Figures 2a and 2c) than for feminine ones (Figures 2b and 2d).

When there is a mismatch between trigger-gender and occupation-gender, accuracy drops. In Table 4, we report ΔM as the difference in accuracy of sentences with (M-Trigger, M-Occupation) and (M-Trigger, F-Occupation) configurations, demonstrating the model's inability to resolve gender mismatches between triggers and occupations (See table 2 for values for the triggers and occupations). We report the same for ΔF where the drop in performance is more pronounced. We take the fact that $\Delta F > \Delta M$ for all languages to be evidence of a more complex type of stereotyping that negatively affects women, namely *androcentrism* (Bem, 1993; Hegarty et al., 2013; Bailey et al., 2019).[5]

[5] Androcentrism is a wide reaching cultural phenomenon that treats the "male experience... as a neutral standard or



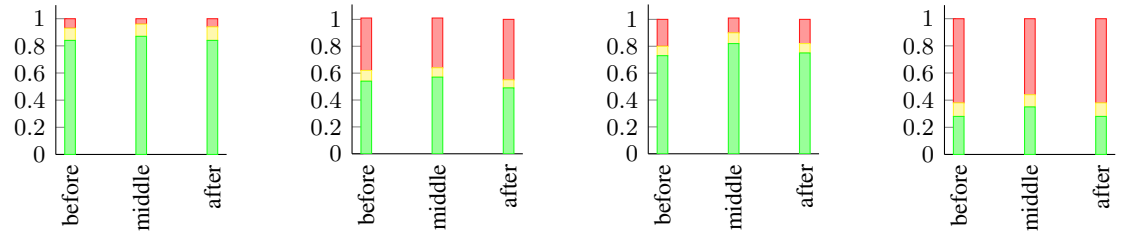
(a) M-trigger M-occupation

(b) F-trigger F-occupation

(c) M-trigger F-occupation

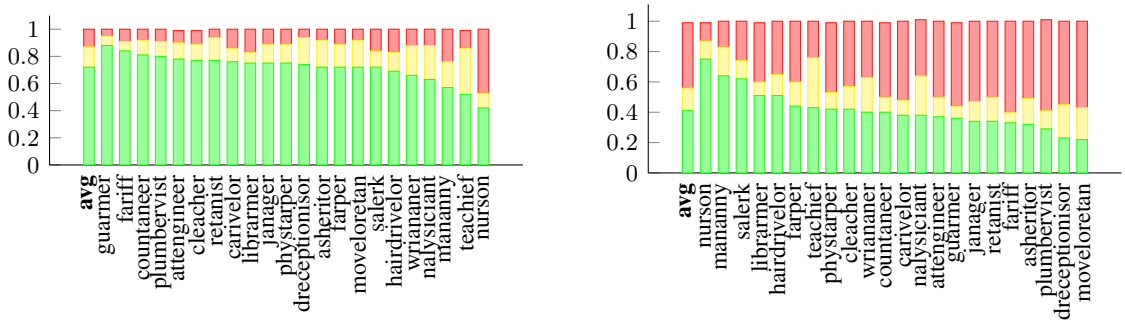
(d) F-trigger, M-occupation

Figure 3: Results for adjective triggers present in any sentential context. Sentences with either a masculine trigger or a masculine-indicative occupation (or both) have higher accuracy regardless of the stereotyped gender of the adjective. Accuracy for sentences where the context adjective matches the trigger is generally higher than for sentences where the context adjective mismatch the gender of the trigger: in Figures 3b and 3d stereotypically feminine adjectives have higher accuracy and in Figure 3c stereotypically masculine adjectives have higher accuracy.



(a) M-trigger, M-occupation (b) F-trigger, F-occupation (c) M-trigger, F-occupation (d) F-trigger, M-occupation

Figure 4: Results for M2M model (1.2B) analysing the relative position of the trigger token and occupation-noun and the trigger token. The "before" category contains source text where the occupation token appears before the trigger token, e.g. *That engineer is my sister*, the "after" category contains source sentences of the form *He works as an engineer* and "middle" category contains a occupation-noun in between two trigger tokens.



(a) M-trigger Nonce-occupation

(b) F-trigger Nonce-occupation

Figure 5: Figures 5a and 5b show the performance for nonce-occupations with M-triggers and F-triggers, respectively.

4 Analysis

In this section, we analyze our results by splitting up languages, occupations, adjective contexts and relative positioning of triggers and occupations using source sentences generated from the small grammar (described in Section 2).

Accuracy is higher when the trigger refers to a man than from when it refers to a woman. As we see in Figure 1, accuracy is lower for the M2M (1.2B) when the trigger requires feminine gender on the occupation, hovering around 40 in most languages. For some languages, such as Urdu, occupation nouns are rarely inflected with the correct gender marking for feminine triggers. The only language for which accuracy on sentences with feminine triggers exceeds 50 is Serbian. In aggregate, these results likely reflect the cultural fact than many languages utilize the masculine form to refer to generic people (Gastil, 1990; Hamilton, 1991).

Accuracy is higher when trigger-gender and occupation-gender match... In Figure 1, the M2M model performs better on inflecting occupations nouns correctly when they are statistically more likely to refer to a person whose gender matches the gender required by the trigger: for example, our models are better at correctly marking *nanny* (stereotypically performed by women) in the context of *mother* than they are at marking *janitor* (stereotypically performed by men). This finding replicates previous work (Stanovsky et al., 2019) that showed that six then-state-of-the-art models were very susceptible to statistical gender biases encoded in occupation words.

...However, gender marking accuracy drops less when the occupation is mismatched with a masculine trigger than when it is mismatched with a feminine one. Although statistical gender biases in how women are presented of the kind presented in Figure 1 are relatively well described in NLP and adjacent fields (Bolukbasi et al., 2016; Hovy and Spruit, 2016; Caliskan et al., 2017; Rudinger et al., 2017; Garg et al., 2018; Garimella et al., 2019; Gonen and Goldberg, 2019; Dinan et al., 2020a,b), we see additional evidence that our NMT systems encode this cultural androcentrism bias in the fact that the drop in accuracy is greater for sentences with feminine triggers (*mother*) and man-stereotypic occupations (*janitor*) than for the converse (compare the magnitude of the drop in Figure 1 and Figure 2 between a and c to the drop between b and d, as well as Table 4).

Models achieve higher accuracy for man-stereotypic than woman-stereotypic occupations (although this varies). To understand particular occupations, we plot the M2M (1.2B) accuracy by occupation averaged across all languages (see Table 5 in the Appendix for the full list of adjectives). Recall that all occupations that are frequent, are either statistically biased towards either men or towards women in the source language, and are balanced in the dataset. We observe that in the case of feminine grammatical gender triggers, only a few woman-stereotypic occupations (e.g. *housekeeper, nurse, secretary* in Figures 2b and 2d) reach the level of accuracy that the model achieves on most man-stereotypic occupations (in Figures 2a and 2c). We also note that variation in accuracy is much higher for woman-stereotypic occupations across both trigger types (compare Figures 2c and 2d), lending support to a cultural androcentrism hypothesis.

Models perform better on sentences when there is a stereotypical adjective that matches the gender of the trigger. We observe an effect of including stereotypical adjectives whereby accuracy is higher when the adjective's stereotypical gender matches the gender that was unambiguously triggered. For example, in Figure 3b shows models translate sentences like *The nanny is my sexy sister* more accurately than *The nanny is my logical sister*, and in Figure 3c sentences like *The sheriff is my logical brother* with higher accuracy than *The sheriff is my feminine brother*. We note that the result holds regardless of whether the adjective precedes the occupation or the trigger (see discussion of Figure 6 and Figure 7 in Appendix A).

Source-side trigger word position does not impact accuracy. We also analyzed if the relative positions of the trigger and occupation tokens (in the source sentence) affect the performance of the model. We split the source sentences into a "before" group wherein all occupation nouns appear before the trigger token, (e.g. *That engineer is my sister*), an "after" group which contained sentences in which the occupation noun appears after the trigger token (e.g. *He works as a engineer*) and a "middle" group where the occupation noun has trigger tokens before and after it (e.g. *He is a nanny who can*

norm for the culture of the species as a whole" (Bem, 1993, p. 41)—one consequence of this cultural phenomenon is that women are restricted to their stereotypical domains (e.g. home, care) more than men are to theirs (e.g. work, science).

inspire himself). Figure 4 shows these findings. We expected the "after" and "middle" category to have better accuracy because the decoding proceeds in a left-to-right manner, which gives allows the model to condition on the target side trigger token when generating the target side occupation token (assuming the target language maintains the same ordering of trigger and occupation tokens). Surprisingly, we do not see a noticeable difference in accuracies between the "before" and "after" categories. We see a small improvement in the "middle" group across evidence that the relative position of the triggers affect the quality of gendered noun translation. Note that the "middle" category has more trigger tokens.

Nonce Word Test. Finally, all of our occupation words genuinely occur in the real world. This means that various idiosyncratic factors, such as word frequency in the training corpora, might have an effect on how well they are translated into other languages. We generate wholly novel nonce occupation words (e.g., *nurson, plumbervist, farper*) which should have no stereotypical gender associations (Appendix C). Therefore, we expect models to do equally well on each word regardless of whether it is in the presence of a masculine or feminine trigger. While Nonce-occupations expectedly have higher levels of inconclusive translations, we do see in Figure 5 that the models are better at resolving a Male-trigger with a Nonce-occupation than a Female-trigger with a Nonce-occupation.

5 Discussion

Recently, several works (Stanovsky et al., 2019; Prates et al., 2019; Gonen and Webster, 2020; González et al., 2020) investigated gender bias in multiple languages with complex morphology, and showed that state-of-the-art MT systems resolve gender-unbalanced occupation nouns (from the US Bureau of Labor Statistics) more often to masculine than feminine pronouns, despite the fact that people of many genders participate in all listed occupations. Our work improves upon these prior approaches by exploring the effects of gender-indicative contexts (e.g., additionally stereotypically masculine and feminine traits and events) in range of syntactic positions (e.g., preceding or following the clue, directly adjacent to the occupation, etc.). While Prates et al. (2019) did investigate some stereotypical traits in their work, they only investigate a few of them, only in the context of the ambiguous paradigm, and were narrowly focused on measuring the translation abilities of one commercial translation product. Recently, Bentivogli et al. (2020) focused on translation quality of occupation-nouns in speech-translation, where they consider the speaker-voice as well as contextual clues. We, on the other hand, explore not only more diverse example traits as well as additional contextual cues, but we do so in unambiguously gendered sentences with a diverse range of sentence structures that allow us to vary the linear precedence of contextual cues as well as their prevalence. Gonen and Webster (2020) also made use of minimally different sentences via an innovative perturbation method that mines examples from real world data and moves away from static word lists; however, their benchmark is also collected for the ambiguous gender setting.

Several works aim to enrich the gender input to an MT system by adding additional gold annotation or context (Stafanovičs et al., 2020; Saunders et al., 2020; Moryossef et al., 2019). This has the additional benefit of making gender tags learnable, but it does not rely on the linguistic signal alone (as we do through leveraging grammatical rules) and instead relies on additional denser annotation. Only two contributions other than our own is known to us to rely only on the particular linguistic structure of the sentence: the first by González et al. (2020) also focused on "unforgivable" grammatical gender-related errors in translation (as well as on other tasks) that come about as a result of syntactic structure and unambiguous coreference. Their approach is somewhat analogous to some of our examples, except that, instead of relying on language-internal properties, we rely on syntactic context to construct unambiguous examples: e.g., particularly those that make use of *own* to make obligatory the local coreference (in this case cataphora) as in *That her **own** child cried, surprised the doctor*. We take our work to be wholly complementary to theirs; Their approach focuses on more source languages, fewer target languages, and a wider range of tasks, we focus on one source language, more target languages, and sentences from a wider range of (source) syntactic structures.

The second work closely related to ours is Renduchintala et al. (2021) which also focuses on unambiguous source sentences. Their work has only a small number of templates for two languages. We propose and create a grammar that encompasses more scenarios where the source sentences contain unambiguous gender indicators for occupation nouns. Our grammar enables us to examine the

effect of adjectives and verbs (which were selected for their association with particular genders) on gendered occupation noun translation accuracy. We also discuss the impact of the relative position of the occupation noun with respect to the gender trigger. Our evaluation scheme allows for more diversity from the NMT model as we do not use a dictionary approach. Our evaluation also focuses on the correctness of morphological markers on the target-side occupation noun and not on the noun itself. Our evaluation scheme also allows us to apply our analysis to more languages.

The present work does not aim to ascertain the cause of models' errors. Our main goal here is to present a novel benchmark for surfacing errors and measuring bias. Since it is relatively well known that generation models, including MT models, often output translations that are less lexically diverse than their training data (Vanmassenhove et al., 2019), several recent works have investigated the effects of gender bias as a function of model training data. Stafanovičs et al. (2020) argues that gender bias in MT models can be lessened if models are trained on denser annotations for identifying the genders of referents.

Concurrently, another approach to pronoun coreference utilized a hand-crafted grammar to generate sentences for measuring fairness (Soremekun et al., 2020), but in the context of NLP tasks other than NMT. Although Soremekun et al. (2020) are interested in measuring performance for unambiguous examples, it does not focus on the NMT use case, and its examples require cross-sentential coreferences, which will likely require a more complex linguistic toolbox than our intrasentential case (Szabolcsi, 2003; Hardmeier and Federico, 2010; Reinhart, 2016). Moreover, the grammar created in that work is much less developed than ours: it does not manipulate the location of the trigger, there is limited syntactic diversity, and there is no incorporation of statistically gender-biased words above and beyond occupation nouns.

At a high level, our work resurfaces problems with morphology in machine translation. While neural machine translation is more fluent than phrase-based machine translation, it has long been observed that even high-resource models can struggle to generate faithful translations that are also syntactically correct (Isabelle et al., 2017) and the problem intensifies for longer sentences with long-distance dependencies (Choshen and Abend, 2019). We highlight yet another morphological failure mode in NMT models in this work. There is also a long history of incorporating morphology and syntax explicitly into NMT models in the hope of reducing the prevalence of such errors (Minkov et al., 2007). For example, Eriguchi et al. (2016) model source-side syntax while Aharoni and Goldberg (2017) proposed models that generate linearized constituency trees. Other works also consider modifications to the attention mechanism in order to improve NMT (Kim et al., 2017).

6 Conclusion

Many of our NLP tasks and datasets are rife with statistical gender biases that reflect, in language, the stereotypical associations we have about gender in our cultures. In this work, we present a new evaluation dataset for measuring gender bias in machine translation for gender unambiguous sentences. Our dataset supports translation from an English source into 20 languages, contains three evaluation datasets of different sizes to accommodate all users, and is designed to answer questions not only about particular occupation words and gender triggering words, but also to further explicate the role of context in how MT systems translate gender morphology. We hope that our dataset will encourage the community to improve on this new setting for measuring gender biases in language.

7 Broader Impact

Our work has proposed a benchmark for measuring morphological gender errors in translation which require adequate representation of the context and may have social repercussions. Our evaluation benchmark measures translation accuracy on an occupation noun on the target side.

In this work, we restrict ourselves to English as a source language. English specifies several kinds of gender, for example, on pronouns, including feminine (*she, her, hers*), masculine (*he, him, his*), nonbinary (*they, them, their, theirs, xe, ze, sie, co, ey*...), and underspecified (*they, them, their, theirs*).[6] We focused solely on binarily gendered contextual clues, although that provides an incomplete picture, for multiple reasons. First, the translation models we evaluated are not yet able to handle underspecified and nonbinary contextual clues consistently, let

[6]Note: Although the sets of morphological forms of underspecified pronouns and nonbinary pronouns overlap, they are not the same phenomena from a linguistic perspective, see Ackerman 2019 i.a.).

alone neopronouns. For example, translating "my parent is a doctor" into German resulted in a translation with a plural verb, and the masculine singular form of the occupation noun (we presume masculine was the majority class in the training data). Second, "they are a doctor"[7] is translated as honorific in German with the pronoun *Sie* (Note that the pronouns for "she" and "they" are homophonous in that language, and are only distinguished by capitalization), but a masculine gender on the occupation, and a plural verb form. If the original translation models that we are aiming to evaluate with our benchmark are unable to translate nonbinary and underspecified examples with any reasonable accuracy at all, this is a much bigger issue requiring its own nuanced investigation. This issue becomes even more complex when you consider what ought to be the appropriate forms of occupation nouns when they refer to nonbinary or individuals we don't know the gender(s) of. Most of the target languages we use do not have a single, standardized way of generating gender-inclusive occupation nouns, because norms regarding complex social/demographic features are currently in flux. Considerations about what ought to be the ideal translation policy will change over time, and will doubtless vary by language and culture. For example, in American English, some prefer *actor* to *actress* as the former is inclusive of all genders. In other languages, specifying more than one gender on the same occupation noun has become preferred (at least in some contexts and among some groups) as another gender-inclusive option. Take, for example, in continental French, the gender on words like "student" can be duplicated as in *étudiante et étudiants*, *étudiant-e-s*, or *étudiant.e.s*, (see Burnett and Pozniak 2021; Pozniak and Burnett 2021; Richy and Burnett 2021 for more information). Even this linguistic innovation however doesn't cover every person's preferences. Some women prefer masculine gender on their occupations because "they have the impression that the masculine forms have a more prestigious connotation than the feminine ones" (Burnett and Pozniak 2021, p.11; Burnett and Bonami 2019).

Acknowledging the range of complexities at play here, for our test benchmark, we fixed the gold translation to obligatorily mark the (binary) gender on the occupation noun in accordance with the explicit gender identity of a person (i.e., it is always preferred for the translation system to explicitly specify a known, binary gender for each occupation noun). Although our approach runs contrary to some preferred ways of referring to people, it is still useful as a tool for uncovering gender biases in current translation systems—it can determine whether the system prefers to translate into the most frequent gender (usually the masculine) while, worryingly, ignoring relevant contextual cues to the contrary. Future iterations of work like this might survey the appropriate ways of specifying nonbinary gender (or purposefully not specifying any gender) in each target language, and develop specific and more fine-grained schemes for measuring statistical gender biases for these situations (Note: considerations like these should be taken into account at the training phase and not just at the evaluation phase).

[7] This sentence unambiguously refers to a single person identifying as non-binary in the English source.

References

Lauren Ackerman. 2019. Syntactic and cognitive issues in investigating gendered coreference. *Glossa: a journal of general linguistics*, 4(1).

Roee Aharoni and Yoav Goldberg. 2017. Towards string-to-tree neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 132–140, Vancouver, Canada. Association for Computational Linguistics.

April H Bailey, Marianne LaFrance, and John F Dovidio. 2019. Is man the measure of all things? a social cognitive account of androcentrism. *Personality and Social Psychology Review*, 23(4):307–331.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Sandra L Bem. 1981. Bem sex role inventory. *Journal of personality and social psychology*.

Sandra L Bem. 1993. *The lenses of gender: Transforming the debate on sexual inequality*. Yale University Press.

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.

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 *Advances in neural information processing systems*, pages 4349–4357.

Daniel Büring. 2005. *Binding theory*. Cambridge University Press.

Heather Burnett and Olivier Bonami. 2019. Linguistic prescription, ideological structure, and the actuation of linguistic changes: Grammatical gender in french parliamentary debates. *Language in Society*, 48(1):65–93.

Heather Burnett and Céline Pozniak. 2021. Political dimensions of gender inclusive writing in parisian universities. *Journal of Sociolinguistics*.

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. 1980. On binding. *Linguistic inquiry*, 11(1):1–46.

Noam Chomsky. 1981. *Lectures on government and binding: The Pisa lectures*. Foris Publications, Holland.

Leshem Choshen and Omri Abend. 2019. Automatically extracting challenge sets for non-local phenomena in neural machine translation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 291–303, Hong Kong, China. Association for Computational Linguistics.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020a. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, Online. Association for Computational Linguistics.

Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020b. 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.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Alice H Eagly, Christa Nater, David I Miller, Michèle Kaufmann, and Sabine Sczesny. 2020. Gender stereotypes have changed: A cross-temporal meta-analysis of us public opinion polls from 1946 to 2018. *American psychologist*, 75(3):301.

Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka. 2016. Tree-to-sequence attentional neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 823–833, Berlin, Germany. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2020. Beyond english-centric multilingual machine translation. *arXiv preprint arXiv:2010.11125*.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

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, Italy. Association for Computational Linguistics.

John Gastil. 1990. Generic pronouns and sexist language: The oxymoronic character of masculine generics. *Sex roles*, 23(11):629–643.

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. 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.

Ana Valeria González, Maria Barrett, Rasmus Hvingelby, Kellie Webster, and Anders Søgaard. 2020. Type B reflexivization as an unambiguous testbed for multilingual multi-task gender bias. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2637–2648, Online. Association for Computational Linguistics.

Elizabeth L Haines, Kay Deaux, and Nicole Lofaro. 2016. The times they are a-changing... or are they not? a comparison of gender stereotypes, 1983–2014. *Psychology of Women Quarterly*, 40(3):353–363.

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

Christian Hardmeier and Marcello Federico. 2010. Modelling pronominal anaphora in statistical machine translation. In *IWSLT (International Workshop on Spoken Language Translation); Paris, France; December 2nd and 3rd, 2010.*, pages 283–289.

Peter Hegarty, Orla Parslow, Y Gávriel Ansara, and Freyja Quick. 2013. Androcentrism: Changing the landscape without leveling the playing field. *The Sage handbook of gender and psychology*, pages 29–44.

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.

Alexander Miserlis Hoyle, Lawrence Wolf-Sonkin, Hanna Wallach, Isabelle Augenstein, and Ryan Cotterell. 2019. Unsupervised discovery of gendered language through latent-variable modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1706–1716, Florence, Italy. Association for Computational Linguistics.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496, Copenhagen, Denmark. Association for Computational Linguistics.

Yoon Kim, Carl Denton, Luong Hoang, and Alexander M Rush. 2017. Structured attention networks. *arXiv preprint arXiv:1702.00887*.

Chunxi Liu, Qiaochu Zhang, Xiaohui Zhang, Kritika Singh, Yatharth Saraf, and Geoffrey Zweig. 2020. Multilingual graphemic hybrid ASR with massive data augmentation. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 46–52, Marseille, France. European Language Resources association.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

Einat Minkov, Kristina Toutanova, and Hisami Suzuki. 2007. Generating complex morphology for machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 128–135, Prague, Czech Republic. Association for Computational Linguistics.

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, Italy. Association for Computational Linguistics.

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, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Céline Pozniak and Heather Burnett. 2021. Failures of gricean reasoning and the role of stereotypes in the production of gender marking in french. *Glossa: a journal of general linguistics*, 6(1).

Marcelo OR Prates, Pedro H Avelar, and Luis C Lamb. 2019. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, pages 1–19.

Deborah A Prentice and Erica Carranza. 2002. What women and men should be, shouldn't be, are allowed to be, and don't have to be: The contents of prescriptive gender stereotypes. *Psychology of women quarterly*, 26(4):269–281.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Tanya Reinhart. 2016. *Anaphora and semantic interpretation*. Routledge.

Adithya Renduchintala, Denise Diaz, Kenneth Heafield, Xian Li, and Mona Diab. 2021. Gender bias amplification during speed-quality optimization in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 99–109, Online. Association for Computational Linguistics.

Célia Richy and Heather Burnett. 2021. Démêler les effets des stéréotypes et le genre grammatical dans le biais masculin: une approche expérimentale. *GLAD!. Revue sur le langage, le genre, les sexualités*, (10).

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, Spain. Association for Computational Linguistics.

Gerard Saucier and Kathryn Iurino. 2020. High-dimensionality personality structure in the natural language: Further analyses of classic sets of english-language trait-adjectives. *Journal of Personality and Social Psychology*, 119(5):1188.

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, Barcelona, Spain (Online). Association for Computational Linguistics.

Allen Schmaltz, Alexander M. Rush, and Stuart Shieber. 2016. Word ordering without syntax. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2319–2324, Austin, Texas. Association for Computational Linguistics.

Ezekiel Soremekun, Sakshi Udeshi, and Sudipta Chattopadhyay. 2020. Astraea: Grammar-based fairness testing. *arXiv preprint arXiv:2010.02542*.

Artūrs Stafanovičs, Toms Bergmanis, and Mārcis Pinnis. 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.

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, Italy. Association for Computational Linguistics.

Anna Szabolcsi. 2003. Binding on the fly: Cross-sentential anaphora in variable-free semantics. In *Resource-sensitivity, binding and anaphora*, pages 215–227. Springer.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

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: Research Track*, pages 222–232, Dublin, Ireland. European Association for Machine Translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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 Extended Adjective Results

When constructing our dataset, we were careful to vary the position of contextual adjectives to get a fuller, more syntactically diverse picture of NMT model performance. We varied whether the stereotypical adjectives (e.g. *logical*, *delicate*) modified the trigger or the occupation. Since English syntax doesn't allow adjectival modification of pronouns, for adjectives modifying triggers, we only considered the subset of sentences with full noun phrases (not pronouns) as triggers. In Figure 3, we observed mainly the cultural androcentrism effect, and wanted to break down those results based on the syntactic position of the adjective. We find that changing the syntactic position of the adjective has little effect on the overall findings Figure 6 and Figure 7. One notable exception is the adjectives *feminine* and *masculine*, which when modifying mismatched occupation nouns attain the lowest performance of our selected subset (see Figures 6a and 6c).

Moreover, we tentatively observed that when the adjectives modify the occupation noun, there is slightly more variance in accuracy than when they modify the trigger; this is more pronounced for feminine triggers than masculine triggers. Despite finding only minor differences for our NMT models for different syntactic positions, we included different and diverse syntactic structures so that our dataset can also be used to evaluate performance on other types of neural architectures, such as LSTMs, which are sometimes found to be more sensitive to word order (Schmaltz et al., 2016), as well as for future models that have yet to exist.

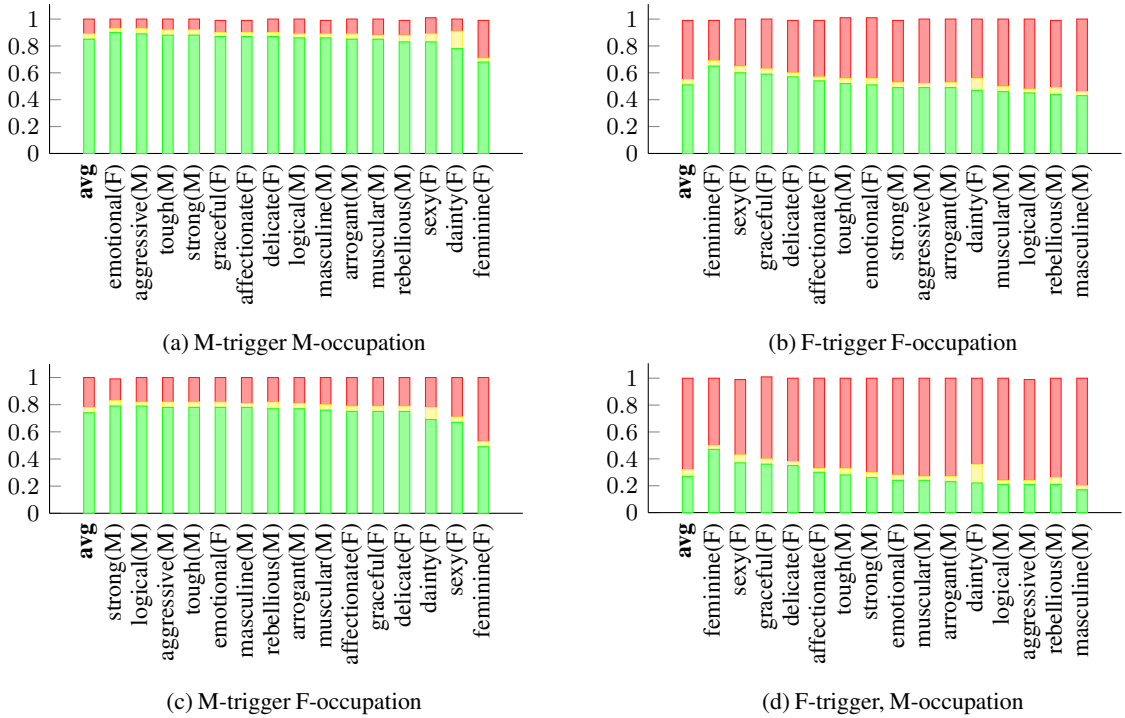
B Context Words List

Our context words (Table 5) were drawn from published literature in social psychology and gender studies (Bem, 1981; Prentice and Carranza, 2002; Haines et al., 2016; Eagly et al., 2020).

C Generating Nonce Words

We trained a simple character level 3-gram language model on the occupations listed in Table 2 and sampled 22 nonce words from this model with the restriction that they be between 4 and 14 characters long to determine whether the higher accuracy for sentences with masculine triggers. We filtered nonce strings using a large list of English words to ensure that none are existing words.[8]

[8]We use a list of 466k English words from https://github.com/dwyl/english-words.



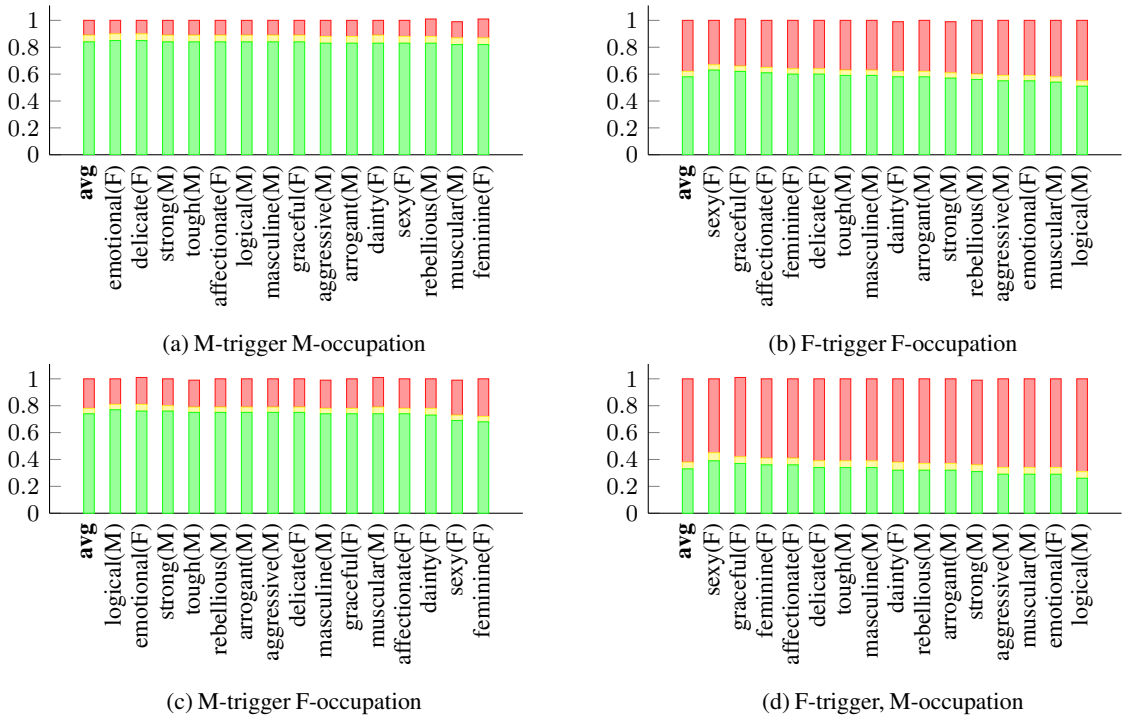
(a) M-trigger M-occupation

(b) F-trigger F-occupation

(c) M-trigger F-occupation

(d) F-trigger, M-occupation

Figure 6: Results for M2M model (1.2B) for sentences where there is a stereotypical adjective modifying the occupation noun. Proportion of correct (green), incorrect (red) and inconclusive (yellow) are provided.



(a) M-trigger M-occupation

(b) F-trigger F-occupation

(c) M-trigger F-occupation

(d) F-trigger, M-occupation

Figure 7: Results for M2M model (1.2B) for sentences where there is a stereotypical adjective modifying a noun phrase trigger. Proportion of correct (green), incorrect (red) and inconclusive (yellow) are provided.

Type	F	M
$\text{Context}_{\text{Adj}}$	affected, affectionate, appreciative, emotional, excitable, imaginative, impressionable, intelligent, organized, outgoing, unambitious, understanding, unintelligent, unselfish, unstable, cautious, changeable, charming, cheerful, childlike, clean, compassionate, complaining, complicated, confused, cooperative, creative, critical, curious, dainty, delicate, dependent, dreamy, family-oriented, fashionable, fault-finding, fearful, feminine, fickle, flatterable, flirtatious, foolish, forgiving, friendly, frivolous, fussy, gentle, graceful, gullible, helpful, honest, kind, loyal, melodramatic, mild, modest, naive, nervous, patient, pleasant, polite, prudish, romantic, self-pitying, sensitive, sentimental, sexy, short, shy, small-boned, smart, soft, soft-hearted, sophisticated, spiritual, submissive, suggestive, superstitious, sympathetic, talkative, tender, timid, touchy, warm, weak, well-dressed, well-mannered, wholesome, worrying, yielding	aggressive, active, adventurous, aggressive, ambitious, analytical, arrogant, assertive, athletic, autocratic, enterprising, independent, indifferent, individualistic, initiative, innovative, intense, inventive, obnoxious, opinionated, opportunistic, unfriendly, unscrupulous, bossy, broad-shouldered, capable, coarse, competitive, conceited, confident, consistent, controlling, courageous, cruel, cynical, decisive, demanding, dependable, determined, disciplined, disorderly, dominant, forceful, greedy, hard-hearted, hardworking, humorous, jealous, lazy, level-headed, logical, loud, masculine, muscular, pleasure-seeking, possessive, precise, progressive, promiscuous, proud, quick, rational, realistic, rebellious, reckless, resourceful, rigid, robust, self-confident, self-reliant, self-righteous, self-sufficient, selfish, serious, sharp-witted, show-off, solemn, solid, steady, stern, stingy, stolid, strong, stubborn, sturdy, tall, tough, well-built, witty
$\text{Context}_{\text{V-OBJ}}$	protect, treat, shame, exploit insult, scare, frighten, distract, escort	reward, glorify, thank, praise, honor, inspire, enrich, appease, congratulate, respect, flatter, destroy, deceive, bore, offend, scold, pay, fight, defeat
$\text{Context}_{\text{V-SUBJ}}$	smile, dance, laugh, play, giggle, weep, faint, scream, gossip, complain, lament, spin, celebrate, clap	succeed, flourish, prosper, win, protest, kill, threaten, rush, speak

Table 5: Gendered context words from our dataset. In the dataset, we also include verbs by argument structure from Hoyle et al. (2019), although we leave their analysis to future work.