Examining Covert Gender Bias: A Case Study in Turkish and English Machine Translation Models

Chloe Ciora*, Nur Iren*, Malihe Alikhani
University of Pittsburgh
{chloeciora,nei3,malihe}@pitt.edu

Abstract

As Machine Translation (MT) has become increasingly more powerful, accessible, and widespread, the potential for the perpetuation of bias has grown alongside its advances. While overt indicators of bias have been studied in machine translation, we argue that covert biases expose a problem that is further entrenched. Through the use of the gender-neutral language Turkish and the gendered language English, we examine cases of both overt and covert gender bias in MT models. Specifically, we introduce a method to investigate asymmetrical gender markings. We also assess bias in the attribution of personhood and examine occupational and personality stereotypes through overt bias indicators in MT models. Our work explores a deeper layer of bias in MT models and demonstrates the continued need for language-specific, interdisciplinary methodology in MT model development.

1 Introduction

Various forms of biases are encoded in the way that people use language (Rudinger et al., 2018; Butler, 1990). Similar to other Natural Language Processing (NLP) tasks, learned models used in MT systems include social biases as they learn correlations from their training data that have encoded stereotypes. Specifically, several studies (Prates et al., 2020; Cho et al., 2019; Baeza-Yates, 2019) have shown that translations from a gender-neutral language to a language with gendered pronouns are biased in the selection of pronouns in the target language.

However, this is not the only way bias can manifest in MT. For example, Figure 1 demonstrates marked gender in the female case of the same sentence while remaining neutral in the male case. Since the translations are both accurate, unless the two sentences are presented together, the asymmetry in gender reference is not immediately obvious. The example demonstrates the use of optional referential gender in Turkish, highlighting the need to frame gender bias in MT around language-specific social and cultural knowledge.

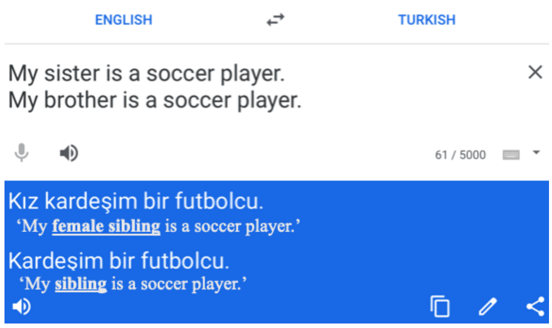

Figure 1: Using Google Translate, "My sister is a soccer player" accurately translates to "My **female sibling** is a soccer player" while "My brother is a soccer player" is translated to "My **sibling** is a soccer player". Gender is overtly marked only when the subject is female.

While previous mitigation efforts have focused on debiasing training data (Elaraby et al., 2018; Costa-jussà and de Jorge, 2020; Stafanovičs et al., 2020; Saunders and Byrne, 2020), the issue of covert bias has not been adequately addressed, and goes far beyond the perpetuation of outdated stereotypes. In order to ensure that the true meaning of the source is accurately represented during the translation process, understanding the linguistic and social context of the utterance is necessary.

In this paper, we examine both overt and covert gender biases in commercially-used MT models through the use of a gender-neutral language, Turkish, and a gendered language, English. Our study investigates explicit stereotype bias through the assignment of pronouns according to stereotypes regarding occupation and personality. We also investigate how additional qualifiers to job descriptions affect results: for example, are "good doctors"

* Equal contribution.

Proceedings of the 14th International Conference on Natural Language Generation (INLG), pages 55–63,
Aberdeen, Scotland, UK, 20-24 September 2021.

more likely to be men than "bad" ones? Similarly, we measure how a reference to personhood changes pronoun results. Lastly, we shed light on the presence of asymmetrical gender in MT models by analyzing explicit gender markings in Turkish translations of gender-specific English sentences. We not only ask if gender markings occur more for female subjects, but also if gender markings are more likely when the stereotype of the predicate does not align with the gender of the subject.

To this end, we created a parallel corpus of 1,617 Turkish and English job titles. We also compiled a list of descriptive adjectives based on Turkish stereotypes and formed appropriate Turkish sentences with and without a reference to personhood. Lastly, we formed a dataset of English sentences by pairing a gendered English subject word (that has no gendered translation in Turkish) with a gender-stereotyped action or description. Our code and data can be found in our GitHub repository.[1]

2 Related Work

Previous works on bias in embeddings and models (Bolukbasi et al., 2016; Zhao et al., 2019; Stanovsky et al., 2019), as well as corpora (Babaeianjelodar et al., 2020), have demonstrated that gender bias exists in the core of MT models. Additionally, Stanovsky et al. (2019) introduced a challenge set in measuring bias from English to languages with morphological gender.

One common approach in bias evaluation is to translate from a gender-neutral language to a gendered language and examine the pronouns selected for occupations and adjectives (Prates et al., 2020; Farkas and Németh, 2020; Cho et al., 2019). We used a modified version of these methods by ensuring that the occupation exists in the target language as well as the source language and that the adjectives used are actual stereotypes in Turkey (Sakallı et al., 2018)[2]. Our remaining experiments are inspired by socio-linguistics research in Turkish. First, Braun (2001) discusses how neutral Turkish words describing people, such as *insan* ("human"), tend to be biased towards male interpretations. In NLP, Mehrabi et al. (2020) examines a related bias in English named-entity recognition where fewer female names are recognized as "person" entities than male ones. Our work will similarly examine gender and personhood bias but in MT models. Second, Braun (2001) describes asymmetrical gender markings in the Turkish language, concluding that male gender remains unmarked regardless of context, whereas female gender tends to be overly expressed. For example, female children are more likely to be referred to with marked gender (*kız çocuğu* "girl child" instead of *çocuk* "child") than male children. The exception to this pattern is when the subject is exceptionally stereotyped as feminine (e.g. *hizmetçi* "househelper"). We will extend the study of this phenomenon to MT.

3 Experiments

We used four commercially available MT models in our experiments: Google Translate, Amazon Translate, Microsoft Translator, and SYSTRAN. For reproducibility purposes, all translations were executed in April of 2021. All datasets can be found on our GitHub[1].

3.1 He is a Doctor, She is a Nurse? Gender Bias in Job Occupation

We examined the distributions of the pronouns selected in English when Turkish sentences were translated following the template[3]: "He/She is a(n) <occupation>", and compared them to the 2020 Turkish (Türkiye Istatistik Kurumu, 2021) and US (U.S. Bureau of Labor, 2020) workforce statistics. Inspired by Farkas and Németh (2020), a second template "He/She is a <adjective> <occupation>" was also formed using the words *çok kötü* ("very bad"), *kötü* ("bad"), *iyi* ("good"), and *çok iyi* ("very good") as attributive adjectives to determine their influence.

We retrieved occupation lists from Turkish and US government agencies[4] and matched occupations that exist in both countries [5]. Some occupation titles were modified for clarity, and some were removed due to gender requirements or a lack of census data, as described in detail in Appendix A. Through our matching process, we were able to match 1,617 occupations.

[1] https://github.com/NurIren/Gender-Bias-in-TR-to-EN-MT-Models

[2] Turkish is also a commonly used gender-neutral language in previous works (Prates et al., 2020; Lauscher and Glavaš, 2019; Zhao et al., 2020), but these works use an intermediary translator to form their Turkish datasets.

[3] The same template was also used by Prates et al. (2020).

[4] Turkish Employment Agency (İŞKUR) and the United States Department of Labor Bureau of Labor Statistics

[5] Using both the major and minor occupational titles of International Standard Classification of Occupations (ISCO-08)

3.2 He is Smart, She is Beautiful? Bias in Adjective Use

We pulled stereotypes from a study where Turkish undergraduate students were asked to provide adjectives that describe men and women (Sakallı et al., 2018). We compiled the list of adjectives presented by this work and removed any that were lexically gendered, leaving 97 total adjectives. Each adjective was then labeled as either masculine-coded (e.g. *agresif* "agressive") or feminine-coded (e.g. *güçsüz* "weak") if more than 60% of the time that word was used to describe a certain gender. All others were considered to be neutral.

The adjectives were first placed into the template "O <adjective>" (He/She is <adjective>)[6] to assess the adjective stereotypes and then into the template "O <adjective> birisidir" ("He/She is someone who is <adjective>)[7] in order to assess if the introduction of the "personhood factor" changed the assumed gender in the translations.

[6]Since Turkish is an agglutinative language, the proper suffixes were also appended to each adjective in order to fit the first template.

[7]Note that although the translation may seem unnatural in English, this is a common utterance in Turkish.

3.3 Bias Through Asymmetrical Gender Markings

English sentences were formed with grammatically gendered subjects, followed by a predicate including a stereotypical occupation, description, or activity. For example, "My sister is an engineer" contains a female subject and a stereotypically masculine predicate. These sentences were then translated to Turkish to measure if the subject was gender-marked. We aim to answer several questions. First, are sentences with male subjects less likely to mark gender than sentences with female subjects? Second, is gender more likely to be marked when the stereotype of the predicate does not align with the gender of the subject?

We selected four subject words that are gendered in English but are grammatically neutral in Turkish. For example, there are no commonly used words for "brother" and "sister"; the only options are "sibling" (*kardeş*), "male sibling" (*erkek kardeş*), or "female sibling" (*kız kardeş*). For each of the predicate categories (occupation, description, and activity), we selected five that were stereotypically masculine and five stereotypically feminine according to Turkish gender stereotypes (Sakallı et al.,

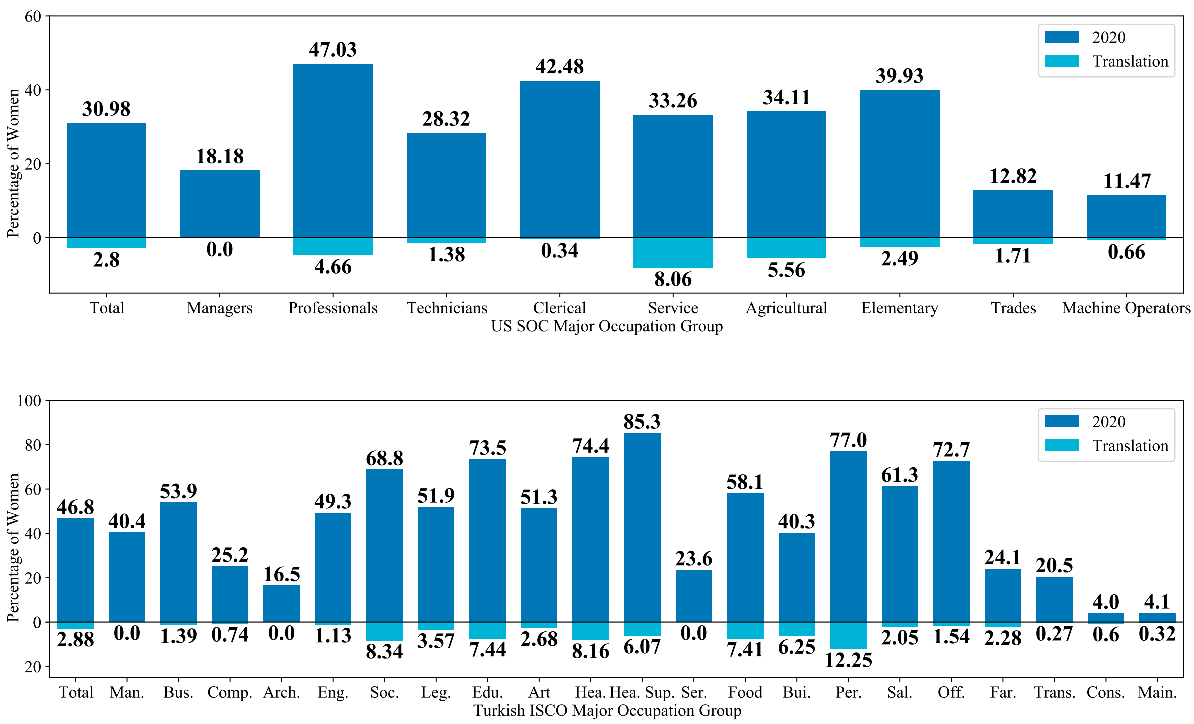

Figure 2: Comparison of the percent of women in the Turkish (bottom) and US (top) labor force in 2020 with the average of the MT models broken down by ISCO-08 and SOC major groups. In both breakdowns, the translation results clearly do not match the labor force. Additionally, the percent of female translations tends to increase in ISCO groups (bottom) with higher female participation. Full group names can be found in Appendix A.

2018; Vatandaş, 2011).

By checking the translations for overt gender markings, translators can be evaluated for asymmetry. We compared the results across the gender of each original English subject word as well as the stereotypical gender of each predicate. With 10 sentence templates in each category for the four gendered subject words, we constructed 120 sentences for each gender in total.

4 Results

In this section, we evaluate[8] aggregate results across all experiments.

4.1 Gender Bias in Occupations

Overall, the percent of female pronouns selected by the MT models were: 1.11% with Google, 1.18% with Amazon, 3.83% with Microsoft, and 5.07% with Systran. Figure 2 demonstrates that this is drastically low compared to female participation in the 2020 workplace in Turkey (31.78%) and the US (47%).

The SOC 2018 group breakdown reveals that, for occupation groups where female participation is either approximately equivalent to or greater than male participation, the models tended to translate the occasional occupation with a female pronoun. Occupations where women are the minority tended to have none or nearly no female translations. Additionally, stereotypical occupations like nurses, fashion designers, and beauticians[9] were consistently translated with female pronouns. Overall, assuming the translation results in each job category should match the corresponding labour statistic, our results were statistically significant ($p < 0.01$).

4.2 Impact of an Attributive Adjective Preceded by Occupation

As shown in Table 1, when an adjective was introduced, sentences originally assigned a female pronoun were more likely to be assigned a male pronoun instead. For each attributive adjective, this was statistically significant ($p < 0.01$). Furthermore, as the adjective changed from *çok iyi* "very good" to *çok kötü* "very bad", the amount of female pronouns that changed to male increased, but the reverse occurred for male pronouns. For example, using Google, Amazon, and SYSTRAN, the Turkish sentence "*O bir Yoğun Bakım Hemşiresi*" yielded the translation "She is an intensive care unit nurse", but the sentence "*O çok kötü bir Yoğun Bakım Hemşiresi*" yielded "He is a very bad intensive care unit nurse".

Adjective	"She"→"He"	"He"→"She"
Very Good	**0.1272**	0.0044
Good	**0.1503**	0.0039
Bad	**0.3353**	0.0005
Very Bad	**0.3815**	0.0010

Table 1: The proportion of pronouns that changed (female to male or male to female) due to the addition of an attributive adjective, cumulative across all translators.

4.3 Turkish Gender Stereotypes in Person Descriptors

For the first sentence template ("He/She is <adjective>"), the first outstanding result is that only 6.74% percent of the pronouns assigned were female (SYSTRAN: 24.5%, Google: 2%, Microsoft: 3.1%, Amazon: 2%) which indicates a strong bias towards male pronouns overall. Secondly, the sentences that were translated to a female pronoun were much more likely to have contained a female-coded adjective. This was highly significant ($p < .01$) in comparison to the amount of female pronouns generated by sentences with male-coded adjectives and significant ($p < .05$) in comparison to neutral ones. The reverse did not hold true for male pronouns; while 83.34% of all sentences that were assigned a female pronoun contained female-coded adjective, only 46.70% of translations with male pronouns were male-coded.

4.4 Analyzing Gendered Personhood

Following from the previous section, we analyze if adding a personhood modifier to the adjective sentences affects pronoun use. Of the sentences that were assigned female gender in the first template, 74.07% changed to male pronouns in the second template when personhood was introduced. The opposite is not the case; only 2.76% of adjectives with male pronouns in Template 1 were female in Template 2. Overall, each translator was significantly more likely to assign a male pronoun when the original sentence contained a personhood modifier ($p < 0.01$).

4.5 Asymmetrical Gender Analysis

As shown in Figure 3, for male subject words, 47.7% of the translations did not mark gender

[8] One sided t-tests performed with equal variance and $p < 0.01$ unless specified otherwise.

[9] A full list of occupations assigned female pronouns can be found in the appendix.

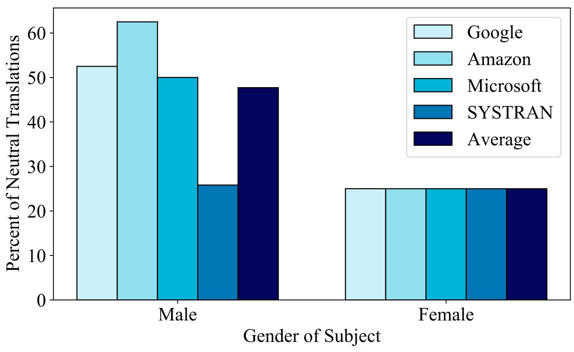

Figure 3: The percentages of translations that used the neutral case according to the gender of the subject per translator as well as the average. While the translations with male subject words had an almost even split, female subject words left gender unmarked only 25% of the time.

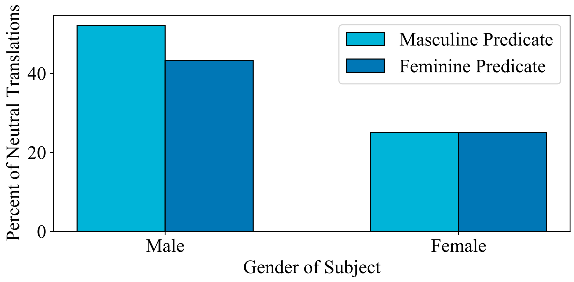

Figure 4: The percent of translations that used the neutral case and didn't preserve gender, across male and female stereotyped predicates, as well as masculine or feminine subjects. For male subject words, gender is significantly more likely to be overtly expressed if the predicate is stereotypically feminine ($p < 0.05$).

and used the neutral form. However, only 25% of the female subject words used the neutral case. This was due to one word, *yeğen* ("niece/nephew"), that remained neutral 100% of the sentences for both male and female subject words. We theorize that this derives from spoken Turkish as *yeğen* ("niece/nephew") is not frequently gender-marked.

Figure 4 demonstrates that when the predicate was stereotypically masculine and the subject word was male, the MT models assumed that the gender of the subject did not need to be overtly expressed, and gender was not preserved 52.1% of the time. For example, "The young men are soccer players" (masculine predicate) did not preserve gender in the translation while "The young men are secretaries" (feminine predicate) did. However, gender was overtly expressed in 56.6% of translations when a stereotypically female predicate was paired with a male subject. Female subject words did not follow this pattern—in fact, for all subject words other than niece/nephew, gender was overtly marked in 75% of the translations. In summary, although male gender was only marked when the content of the sentence deviated from the masculine social norm, female gender was marked in the overwhelming majority of cases, and was consistently treated as aberrational regardless of context.

5 Conclusion

We have examined gender bias exhibited by commercially used MT models in the case of Turkish and English translations. We have shown evidence of overt gender bias through occupation and adjective stereotypes, and covert gender bias through asymmetrical gender and personhood bias. Furthermore, our experiments show consistent evidence of male bias in a neutral context. Male gender was assumed in reference to gender-equal occupations and stereotype-neutral adjectives, and the same phenomenon extends to the manifestation of overt gender markings where male subjects were more likely to be assigned the neutral case. However, when the context was not neutral, stereotype bias routinely affected results across all experiments.

Previous bias mitigation discussions have focused on fair pronoun assignments (Prates et al., 2020; Cho et al., 2019; Baeza-Yates, 2019). Additionally, Google Translate has recently implemented a gender-specific translation feature (Kuczmarski, 2018; Johnson, 2020). While pronoun assignment is a salient and ongoing concern, our study demonstrates how the problem of gender bias can be far more complex. Our experiments show that domain and cultural knowledge are required and these techniques are not necessarily transferable across languages. We advocate for the inclusion of language-specific differences and the design of mitigation models that are linguistically aware and socially grounded. We hope that our work will bring more attention to such interdisciplinary work, prompt continued research in how gender bias is expressed in NLP, and assist with mitigation efforts.

Acknowledgements

We would like to thank Sami Iren for his assistance and insights in verifying our translation data sets for accuracy.

References

Marzieh Babaeianjelodar, Stephen Lorenz, Josh Gordon, Jeanna Matthews, and Evan Freitag. 2020. Quantifying gender bias in different corpora. In *Companion Proceedings of the Web Conference 2020*, WWW '20, page 752–759. Association for Computing Machinery.

Ricardo Baeza-Yates. 2019. Bias on the web. *Symposium on Fairness and Transparency*, page 9.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364. Curran Associates Inc.

Friederike Braun. 2001. The communication of gender in turkish. In *Gender Across Languages*, pages 283–310. John Benjamins.

Judith Butler. 1990. *Gender trouble: feminism and the subversion of identity*. Routledge, New York, NY.

Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. On measuring gender bias in translation of gender-neutral pronouns. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 173–181. Association for Computational Linguistics.

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

Mostafa Elaraby, Ahmed Y Tawfik, Mahmoud Khaled, Hany Hassan, and Aly Osama. 2018. Gender aware spoken language translation applied to english-arabic. In *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, pages 1–6. IEEE.

Anna Farkas and Renáta Németh. 2020. How to measure gender bias in machine translation: Optimal translators, multiple reference points. *arXiv preprint arXiv:2011.06445*.

Melvin Johnson. 2020. A scalable approach to reducing gender bias in google translate. *Google AI Blog*.

James Kuczmarski. 2018. Reducing gender bias in google translate. *The Keyword*.

Anne Lauscher and Goran Glavaš. 2019. Are we consistently biased? multidimensional analysis of biases in distributional word vectors. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 85–91. Association for Computational Linguistics.

Ninareh Mehrabi, Thamme Gowda, Fred Morstatter, Nanyun Peng, and Aram Galstyan. 2020. Man is to person as woman is to location: Measuring gender bias in named entity recognition. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, HT '20, page 231–232. Association for Computing Machinery.

Marcelo Prates, Pedro Avelar, and Luís Lamb. 2020. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, 32.

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. Association for Computational Linguistics.

Nuray Sakallı, Beril Türkoğlu, and Abdulkadir Kuzlak. 2018. How are women and men perceived? structure of gender stereotypes in contemporary turkey. *Nesne Psikoloji Dergisi*, 6:309–336.

Danielle Saunders and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736. Association for Computational Linguistics.

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. 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. Association for Computational Linguistics.

Türkiye Istatistik Kurumu. 2021. İşgücü İstatistikleri, 2020.

U.S. Bureau of Labor. 2020. Labor force statistics from the current population survey.

Celalettin Vatandaş. 2011. Toplumsal cinsiyet ve cinsiyet rollerinin algılanışı. *Istanbul Journal of Sociological Studies*, 0:29 – 56.

Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender bias in multilingual embeddings and cross-lingual transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2896–2907. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. pages 629–634.

A Occupation Data Set Details

This section provides additional details on how the occupation data set was created. The final data set includes matches that are exact matches and matches that are similar. Similar matches fall into one of the following categories:

1. One occupation is a more specific or broad version of it's matching occupation.
2. One occupation uses a slightly different title but describes a similar job.
3. Specifically for educational occupations, the matching occupation describes a different educational level. This helps include occupations that generally exist, but due to different education system setups, are offered at different levels.

Some of the occupation titles have been slightly modified in order to better describe the occupation it matches. These modifications fall into one of the following categories:

1. The occupation title includes punctuation like hyphens or parentheses that describe the occupation. These titles were modified to include the details provided by that occupation.
2. The occupation is split into multiple occupations because it is two separate occupations in the matching country.
3. Specific job details not included in the matching occupation were removed.

Although there were matches, certain occupations were not included for the following reasons:

1. Any religious occupation, due to gender requirements of the majority of those occupations, were not included.
2. Gender specific Turkish occupations. This includes occupations that are either culturally gendered or lexically have gender.
3. Due to different governmental regulations and requirements surrounding gender, military occupations were not included.

Lastly, we list all occupation group names and their abbreviations in Tables 2 and 3.

Abbreviation	SOC Major Group Title
Man.	Management
Bus.	Business and Financial Operations
Comp.	Computer and Mathematical
Arch.	Architecture and Engineering
Eng.	Life and Physical Engineering
Soc.	Community and Social Service
Leg.	Legal
Edu.	Education Training and Library
Art.	Arts, Design, Entertainment, Sports and Media
Hea.	Healthcare Practitioners and Technical
Hea. Sup.	Health Practitioner Support Technologists and Technicians
Ser.	Service
Food	Food Preparation
Bui.	Building and Grounds Cleaning and Management
Per.	Personal Care and Service
Sal.	Sales and Office
Off.	Office Administration Support
Far.	Farming, Fishing and Forestry
Trans.	Transportation and Material Moving
Cons.	Construction and Extraction
Main.	Installation, Maintenance, and Repair

Table 2: Full US SOC occupation titles.

Abbreviation	ISCO Major Group Title
Technicians	Technicians and Associate Professionals
Clerical	Clerical Support Workers
Service	Service and Sales Workers
Agricultural	Skilled Agricultural, Forestry, and Fishery Workers
Trades	Craft and Related Workers
Machine Operators	Plant Machine Operators and Assemblers
Elementary	Elementary Operators

Table 3: Turkish ISCO group names.

B Occupations with Female Generated Pronouns

Table 4 lists all occupations that were assigned female pronouns by at least 3 out of 4 translators.

Occupational Health Spec.	Skin Care Instructor
Barbering Instructor	Emergency Room RN
Registered Nurse	Housekeeping Aide
Surgical Nurse Practitioner	Interior Design Professor
Fashion Designer	CCU Nurse
Certified Diabetes Educator	Bridal Gown Fitter
Cosmetology Instructor	Clinical Nurse Specialist
Makeup Artist	Beautician
Pediatric Registered Nurse	

Table 4: Occupations assigned mostly female pronouns.

The matching Turkish occupation titles can be found in the GitHub[1].

C Sentence Templates in Turkish

Table 5 lists original sentence templates in Turkish.

Original Turkish Template	English Translation
O bir <occupation>	He/she is a <occupation>
O bir <adjective>	He/she is <adjective>
O bir <adjective> <occupation>	He/she is a <adjective> <occupation>
O <adjective> birisidir	He/she is someone who is <adjective>

Table 5: Turkish sentence templates. In the third template, the adjective was one of: "çok iyi" (*very good*), "iyi" (*good*), "kötü" (*bad*), or "çok kötü" (*very bad*).