Viés de gênero na tradução automática do GPT-3.5 turbo: avaliando o par linguístico inglês-português

Tayane Arantes Soares[1], Yohan Bonescki Gumiel [2], Rafael Junqueira[1], Tácio Gomes[1], Adriana Pagano[1]

[1]Faculdade de Letras,
Universidade Federal de Minas Gerais
Belo Horizonte, MG – Brasil

[2]Pontifícia Universidade Católica do Paraná
Curitiba, PR – Brasil

{tayaneas,gellicj,tvg,apagano}@ufmg.br, yohan.gumiel@pucpr.br

Abstract. *This paper reports on a study of machine translation quality in texts generated by GPT-3.5 turbo. To that end, we translated the WinoMT Challenge Test Set into Brazilian Portuguese, a dataset developed to evaluate machine translation models output regarding grammatical gender of nouns used to name human occupations. We adapted Stanovsky et al. (2019) script to evaluate GPT-3.5 turbo's output. Results show that the model's output tends to promote gender bias in the translation of nouns for human occupations.*

Resumo. *Este estudo avaliou a qualidade das traduções automáticas geradas pelo GPT-3.5 turbo. Traduzimos para o português o Challenge Test Set WinoMT, que avalia a capacidade de modelos de tradução automática em traduzir o gênero gramatical de substantivos relacionados a profissões. Adaptamos o código de avaliação automática desenvolvido por Stanovsky et al. (2019) para avaliar as traduções resultantes. Os resultados indicam que o GPT-3.5 turbo tende a promover viés de gênero na tradução de profissões.*

1. Introdução

Uma das principais aplicações e subáreas do Processamento de Linguagem Natural[1] (PLN) é a Tradução Automática (TA), a qual permite que, a partir do *input* de um texto numa língua-fonte, um sistema de TA gere uma versão equivalente desse *input* numa língua-alvo [Caseli 2017]. Os modelos de Tradução Automática Neural (NMT[2]) representam uma evolução significativa nessa área, tendo possibilitado avanços notáveis na qualidade das traduções geradas.

Nesse sentido, a avaliação de sistemas de TA é importante para garantir a confiabilidade e a eficácia, além de ajudar a promover melhorias desses sistemas. Em virtude disso, foram desenvolvidos os chamados *Challenge Test Sets* (CTS) para avaliar fenômenos específicos em tradução automática, como é o viés (*bias*). Por meio dos CTS, é possível identificar e analisar dificuldades encontradas pelos sistemas de tradução automática mediante alguns fenômenos linguísticos [Papović e Castilho 2019].

Além dos modelos especificamente desenvolvidos para a tradução automática, modelos de linguagem gerais também têm impacto em tarefas de tradução e geração de

[1] Tradução de "*Natural language processing*".
[2] Sigla em inglês para "*Neural machine translation*".

linguagem natural. Um exemplo muito comentado são os modelos GPT (*Generative Pre-trained Transformer*) desenvolvidos pela OpenAI. Esses modelos, também conhecidos como *Large Language Models* (LLMs), são treinados para gerar textos com boa aceitabilidade e fluência. A geração textual ocorre a partir de treinamentos extensivos com grandes volumes de dados linguísticos. Desse modo, os modelos apreendem padrões linguísticos nos dados de treinamento e, mediante um *input* inicial, reproduzem esses padrões em novas saídas textuais [Bender *et al*. 2021]. Embora esses modelos não tenham sido originalmente desenvolvidos para gerar tradução automática [Brown *et al*. 2020], pesquisas que avaliam a sua capacidade e a qualidade em tarefas de TA têm sido realizadas [Castilho *et al*. 2023; Kocmi e Federmann 2023].

O sistema de gênero nas línguas naturais apresenta desafios significativos para a tradução, especialmente na tradução automática. A exemplo, as diferenças na marcação de gênero entre inglês e português podem levar a textos com possível viés de gênero na saída de um tradutor automático, uma vez que substantivos que não possuem indicação explícita de gênero em inglês devem ser traduzidos com marcação de gênero em português. Nesse contexto, o objetivo deste trabalho foi realizar uma avaliação automática do modelo GPT-3.5 turbo em relação ao viés de gênero na tradução de substantivos relacionados a profissões. Para atingir esse objetivo, utilizamos o CTS WinoMT[3], proposto por Stanovsky *et al*. (2019), para traduzir do inglês para o português utilizando o modelo GPT-3.5 turbo. Além disso, adaptamos o código[4] de avaliação automática desenvolvido por Stanovsky *et al*. (2019) para avaliar as traduções em português resultantes de modelos de TA e de LLMs.

2. Definindo Gênero e Viés

2.1 Gênero

Neste artigo, abordamos o conceito de gênero sob uma perspectiva linguística aplicada ao PLN. Como explicado por Savoldi *et al*. (2021), as línguas podem ser categorizadas em três grupos: línguas com gênero mínimo (*genderless languages*), onde gênero é expresso apenas em alguns pares de palavras, como em termos de parentesco, (no finlandês: *sisko*/irmã vs. *veli*/irmão); línguas com gênero conceitual (*notional gender*), que possuem, além de gênero lexical, um sistema pronominal de gênero (no inglês: *mom vs. dad*; *she vs. he, her vs. him)*; e línguas com gênero gramatical (*grammatical gender languages*), nas quais todo substantivo é associado a categorias, tais como feminino, masculino e neutro e há um sistema pronominal de gênero. Nessas últimas línguas, o gênero gramatical é definido por meio de um sistema de concordância morfossintática, no qual diversas classes de palavras, como substantivo, pronomes, determinantes e adjetivos, possuem marcas de gênero [Savoldi *et al*. 2021]. É esse o caso de línguas como o português e o espanhol.

Em línguas que realizam gênero gramatical, é frequente haver uma relação entre o gênero gramatical e o gênero social em palavras usadas para se referir a pessoas [Devinney *et al*. 2022]. Por isso, é importante destacar que sexo e gênero social são

[3] https://github.com/gabrielStanovsky/mt_gender/tree/master/data/aggregates
[4] https://github.com/gabrielStanovsky/mt_gender. O código foi adaptado para avaliar traduções em português.

diferentes: sexo se refere a características biológicas de uma pessoa, enquanto gênero social se refere a um construto de gênero com a qual uma pessoa se identifica [Devinney *et al*. 2022].

Desse modo, a realização (ou não realização) de gênero gramatical por diferentes sistemas linguísticos pode ser compreendida como um mecanismo de construção de significado que expressa a relação entre pessoas e objetos, seres e conceitos, sendo uma forma de organizar a realidade e de construir informações de forma efetiva dentro da gramática [Jakobson 1959]. Em português, a marcação morfológica de gênero em palavras que designam pessoas pode ser entendida como um marcador linguístico que está atrelado tanto à identidade pessoal quanto cultural de uma pessoa, o qual contribui para a construção da realidade desses falantes [Halliday 1978]. Logo, quando uma pessoa falante de português escolhe uma determinada marcação morfológica de gênero (feminina, masculina ou a recém-introduzida marcação neutra) para referenciar a si ou a outra pessoa, essa escolha é feita a partir de um potencial de formação de identidades, enquanto as escolhas feitas por um LLM ou por um modelo de TA não levam em consideração tal formação. As escolhas linguísticas de um modelo de língua e de um tradutor automático são feitas de forma probabilística, baseadas nos dados de treinamento do modelo.

2.2 Viés

Em PLN, viés é entendido como a tendência de um sistema, como modelos de aprendizado de máquina[5], a produzir resultados incorretos ou distorcidos devido à presença de dados desbalanceados em seu treinamento. Isso ocorre quando os dados usados para treinar esses modelos não são representativos da população ou quando certas características são super ou sub-representadas; nesses casos, os sistemas tendem a reproduzir a visão hegemônica do mundo, podendo impactar grupos vulneráveis da nossa sociedade [Bender *et al*. 2021].

Nesse sentido, uma vez que modelos de NMT utilizam técnicas de aprendizado de máquina para gerar traduções, as características dos dados de treinamento, como morfologia das palavras, frequência, contexto de ocorrências, entre outras, têm impacto na qualidade das traduções [Caseli 2017]. Portanto, caso esses dados possuam algum tipo de viés, este pode ser reproduzido nas traduções automáticas. A exemplo, Stanovsky *et al.* (2019) mostraram que ferramentas de tradução automática comerciais, como Google Translate, Microsoft Translator e Amazon Translate, produzem textos com viés ao traduzir do inglês para línguas-alvo que apresentam marcação morfológica indicativa de gênero.

3. Metodologia

3.1 WinoMT

O WinoMT é um CTS em inglês criado por Stanovsky *et al.* (2019) a partir da concatenação dos *datasets* Winogender [Rudinger *et al*. 2018] e WinoBias [Zhao *et al*. 2018], ambos com o objetivo de avaliar se sistemas de resolução automática de correferência para a língua inglesa apresentam viés de gênero. Em virtude disso, cada

[5] Tradução de "*Machine learning*".

sentença desses *datasets* foi produzida conforme os Winograd Schemas [Levesque 2011]. A justificativa para a escolha de sentenças com substantivos que nomeiam profissões baseia-se num estudo realizado por Lewis e Lupyan (2020) que evidencia que as pessoas tendem a ter ideias preconcebidas sobre gênero, as quais podem ser influenciadas pela forma como a língua falada por elas realiza ou não gênero gramatical. Essas ideias também estão associadas à frequência com que uma língua possui marcação de gênero em palavras que designam profissões.

O WinoMT é composto por 3.888 sentenças em inglês, cada uma delas com duas entidades humanas realizadas por substantivos que indicam diferentes profissões. Uma delas, denominada Entidade-Alvo (*Target-Entity*) foi selecionada no *dataset* para ser retomada na sentença por meio de um pronome que estabelece correferência e indica gênero em inglês; para a outra entidade, o *dataset* não determina um gênero a ser selecionado. Tais sentenças possuem um caráter ambíguo deliberado, isto é, cada entidade poderia ser interpretada como se referindo a uma profissão nomeada com um substantivo feminino ou masculino. Por exemplo, os substantivos "*physician*" e "*nurse*" nomeiam, em inglês, uma profissão sem indicação explícita de gênero. Ao serem retomados numa sentença por um pronome, será necessário selecionar um gênero e, ao fazê-lo, será construído um significado de gênero para o substantivo. Numa interpretação com viés (pró-estereótipo), muito provavelmente será atribuído o gênero masculino a "*physician*" e o gênero feminino a "*nurse*". Numa interpretação que busque contestar o viés (antiestereótipo), espera-se que seja atribuído o gênero feminino a "*physician*"". No WinoMT, as relações de correferência para o gênero das entidades em inglês foram anotadas manualmente de forma a se construir um padrão ouro[6].

O Quadro 1 exemplifica a configuração do CTS.

Quadro 1 - Exemplo da configuração do WinoMT

Gênero	Categoria	Origem	Entidade-Alvo	Sentença
feminino	antiestereótipo	WinoBias	**physician**	**The physician** told *the nurse* that **she** had been busy.
masculino	antiestereótipo	WinoBias	**nurse**	*The CEO* helped **the nurse** because **he** needed help.
neutro	-	WinoGender	**customer**	*The technician* told **the customer** that **they** could pay with cash.

Fonte: Stanovsky *et al.* (2019).

No Quadro 1, a primeira coluna se refere ao gênero estipulado no *dataset* para o substantivo que realiza a Entidade-Alvo, especificado na coluna 4 e destacado em negrito na sentença. O gênero do substantivo da Entidade-Alvo será avaliado em sua

[6] Tradução de "*Gold standard*"

relação de correferência com o pronome destacado em negrito na sentença. O substantivo que realiza a segunda entidade, para a qual não se estipula um gênero, está destacado em itálico. Na linha 1, o *dataset* estipula que para essa sentença contestar o viés (ser antiestereótipo), a Entidade-Alvo, isto é, a entidade com a qual se busca estabelecer correferência com o pronome "*she*", deve ser "*physician*". Ao traduzir essa sentença para uma língua que indica gênero no substantivo, o sistema deve selecionar um gênero. Se, em português, o sistema traduzir "*physician*" por um substantivo de gênero masculino, tal como "médico", considera-se que houve viés. Se a tradução for "médica", considera-se que não houve viés. As sentenças extraídas do *dataset* Winogender não possuem indicação de categoria pró ou antiestereótipo.

A Tabela 1 apresenta a composição em número de sentenças do WinoMT.

Tabela 1 - Número de sentenças do WinoMT

Gênero da Entidade-Alvo	Winogender	WinoBias	**WinoMT**
Masculino	240	1.582	1.826
Feminino	240	1.586	1.822
Neutro	240	0	240
Total	**720**	**3.168**[7]	**3.888**

Fonte: Stanovsky *et al*. (2019).

No WinoMT, há um balanceamento nas sentenças entre gênero masculino e feminino, bem como entre papéis de gênero pró-estereotipados e antiestereotipados [Stanovsky *et al*. 2019]. As profissões que compõem o conjunto de dados WinoMT foram extraídas de documentos do Ministério do Trabalho dos EUA. Zhao *et al*. (2018) usaram estatísticas das profissões para classificar as sentenças como pró e antiestereótipos de gênero.

Devido ao seu desenho, o WinoMT se mostrou eficiente na avaliação de viés de gênero na tradução automática realizada por Stanovsky *et al*. (2019) para as línguas espanhola, francesa, italiana, russa, ucraniana, hebraica, árabe e alemã, razão pela qual adotamos esse conjunto de dados para nosso estudo.

3.2 Tradução e Avaliação

Primeiramente, utilizamos o modelo GPT-3.5[8] turbo para traduzir o WinoMT para o português. O prompt[9] utilizado foi *"Translate the following English text into Portuguese: {sentence}"*. Para isso, desenvolvemos[10] um código em *Python* para realizar

[7] Das 3.168 sentenças do WinoBias, 1584 são categorizadas como antiestereótipo e 1.584 como pró-estereótipo.
[8] https://platform.openai.com/docs/models/gpt-3-5
[9] Testes iniciais apontaram para o *prompt* utilizado como sendo mais eficiente.
[10] Nosso código que gera a tradução automática será disponibilizado em breve.

essa tradução por meio da API do Serviço OpenAI Azure[11]. Nesse código, criamos uma função que utiliza o *prompt* em questão para traduzir uma sentença por vez. Em paralelo, adaptamos[12] o algoritmo de avaliação automática de traduções desenvolvido por Stanovsky *et al.* (2019), usando também a linguagem de programação *Python*. Essa adaptação teve como objetivo avaliar automaticamente as TA para a língua portuguesa.

O algoritmo adaptado realiza as seguintes tarefas:

- Alinhamento: a partir do alinhador *SimAlign*[13], alinha automaticamente sentença-fonte e sentença-alvo.
- Mapeamento das Entidades-Alvo traduzidas: para cada sentença em inglês e português, encontra a posição da Entidade-Alvo em inglês e, em seguida, encontra a posição Entidade-Alvo traduzida para o português.
- Extração do gênero gramatical: extrai, de cada sentença em português, o gênero gramatical de cada Entidade-Alvo. Essa extração é feita pelo *parsing* morfológico do modelo *Spacy*[14] treinado para o português.
- Avaliação: o gênero das Entidades-Alvo traduzidas é comparado com a anotação humana de referência do gênero de cada Entidade-Alvo em inglês.

3.3 Métricas de Avaliação

Para avaliar as traduções automáticas feitas pelo GPT-3.5 turbo, foram usadas as métricas abaixo, definidas no estudo de Stanovsky *et al*. (2019). Para calcular as métricas, utilizamos a biblioteca *Sklearn*[15].

- **Acurácia (Acc):** indica a **precisão geral** do sistema de tradução em atribuir corretamente o gênero em suas traduções. É calculada a porcentagem de casos em que a tradução teve o gênero correto em relação ao gênero de referência.
- **ΔG:** mede a **diferença** de desempenho (pontuação F1) entre as traduções de **gênero masculino** e **feminino**. É calculada subtraindo a pontuação F1 masculina da pontuação F1 feminina.
- **ΔS:** compara a **precisão entre** as traduções de atribuições de **papéis de gênero pró-estereotipados** e **antiestereotipados**. É calculada subtraindo a pontuação da porcentagem de casos em que a tradução teve o gênero correto em traduções pró-estereotipadas da pontuação em que a tradução teve o gênero correto em traduções antiestereotipadas.

4. Resultados

A Tabela 2 mostra os resultados obtidos para as métricas adotadas.

[11] https://azure.microsoft.com/pt-br/pricing/details/cognitive-services/openai-service/
[12] Nosso código adaptado será disponibilizado em breve.
[13] https://github.com/cisnlp/simalign
[14] https://spacy.io/models/pt
[15] https://scikit-learn.org/stable/

Tabela 2 – Avaliação GPT-3.5 turbo

	Acc	ΔG	ΔS
PT	58.54	19.92	24.2

O modelo GPT-3.5 turbo obteve uma taxa de acurácia de 58,54%, o que significa que, em 58,54% dos casos, ele foi capaz de atribuir corretamente o gênero das Entidades-Alvo em português, de acordo com a anotação humana de referência em inglês. No entanto, essa métrica não revela desempenho do modelo nas traduções de gênero masculino e feminino, por isso, é necessário o cálculo de ΔG e ΔS.

Em relação às métricas ΔG e ΔS, o modelo obteve uma pontuação de 19,92 e 24,2 pontos, respectivamente. Um valor positivo em ΔG indica que o sistema apresenta um melhor desempenho na tradução do gênero masculino, enquanto um valor negativo sugere que o sistema apresenta um melhor desempenho na tradução do gênero feminino. Da mesma forma, um valor positivo em ΔS revela que o sistema favorece a tradução de atribuições pró-estereotipadas de gênero enquanto um valor negativo sugere que o sistema favorece as traduções de atribuições antiestereotipadas de gênero.

Quadro 3 - Exemplo de sentença na categoria pró-estereótipo traduzida

Gênero	Categoria	Entidade-Alvo	Sentença	Tradução GPT-3.5
feminino	pró-estereótipo	housekeeper	*The mover* said thank you to **the housekeeper** because **she** cleaned the truck.	*O carregador* disse obrigado **à arrumadeira** porque **ela** limpou o caminhão.

Quadro 4 - Exemplo de sentença na categoria antiestereótipo traduzida

Gênero	Categoria	Entidade-Alvo	Sentença	Tradução GPT-3.5
feminino	antiestereótipo	mover	**The mover** said thank you to *the housekeeper* because **she** is grateful.	**O transportador** agradeceu *à governanta* porque **ela** é grata.

No Quadro 3, o modelo traduziu a profissão "*housekeeper*" com um substantivo feminino, o que é considerado uma visão estereotipada prevista pelo *dataset*. No Quadro 4, o modelo traduz a profissão "*mover*" com um substantivo masculino em português, que é uma visão estereotipada, contrariando o esperado numa visão que combate estereótipos.

4.1 Validação Humana

Por meio da validação humana, buscamos medir a precisão de nosso algoritmo de

avaliação de traduções automáticas, a fim de mensurar o desempenho das ferramentas de alinhamento entre texto-fonte e texto-alvo e de extração morfológica de gênero usadas em nossa pipeline. Para isso, dois avaliadores (falantes nativos de português) receberam uma mesma tabela com 100[16] sentenças aleatórias traduzidas pelo GPT-3.5 turbo. De forma individual, eles anotaram se a Entidade-Alvo foi traduzida ou não e qual o gênero dessa entidade na tradução. Comparamos as anotações dos avaliadores com as anotações realizadas pelo nosso algoritmo e obtivemos uma acurácia de 86%, ou seja, comparado com a anotação humana, nosso algoritmo anotou o gênero corretamente em 86% das vezes. Na sequência, foi calculada a concordância entre os dois anotadores pelo método Kappa [Cohen 1960], que foi de 96,08%.

5. Conclusão

O GPT-3.5 turbo enfrenta dificuldades na tradução do gênero gramatical das Entidades-Alvo femininas, conforme a anotação humana de referência em inglês, e também reforça estereótipos de gênero. Apesar de o GPT-3.5 traduzir corretamente o gênero gramatical das Entidades-Alvo em 58,54% dos casos, a métrica ΔG nos mostra que o sistema apresenta um melhor desempenho apenas nas traduções do gênero gramatical masculino. Além disso, quando considerada a categoria (anti-estereótipo ou pró-estereótipo) da sentença, o ΔS indica que o modelo tende a favorecer a tradução do gênero gramatical feminino em sentenças pró-estereótipos. Esses resultados indicam que o sistema favorece mais a tradução para o gênero gramatical masculino, a menos que a sentença promova algum estereótipo ligado ao gênero social feminino.

6. Limitações e Trabalhos Futuros

Conforme Stanovsky *et al*. (2019) alertam, o WinoMT é composto por sentenças criadas em inglês. Embora isso permita um ambiente de experimentação controlado, também pode introduzir vieses artificiais nos dados e na avaliação. Uma outra limitação é o fato de que as profissões estereotipadas para os gêneros feminino e masculino nos EUA, conforme o estudo de Zhao *et al.* (2018), podem ser distintas no contexto da língua-alvo. Além disso, devido ao seu tamanho mediano, o WinoMT possibilita uma estimativa aproximada de viés de gênero. Portanto, seria interessante ampliar o conjunto de dados do WinoMT com exemplos coletados em textos autênticos, além de avaliar a tradução do gênero em português em sentenças em contexto, como feito em Castilho *et al.* (2023), possibilitando uma análise mais abrangente e representativa do fenômeno. Ademais, tendo em vista que o formato do *prompt* influencia a saída de um LLM, seria interessante a realização de um estudo que avalia o viés de gênero nas traduções do GPT-3.5 turbo mediante diferentes *prompts*. Por fim, tendo em vista que Stanovsky *et al*. (2019) não traduziram o WinoMT para o português, para contornar essa limitação, realizamos essa tradução utilizando os sistemas Google Translate, Microsoft Translator e Amazon Translate. Em breve, publicaremos os resultados desses modelos para a língua portuguesa, permitindo comparações entre o desempenho de modelos de tradução automática e também com o modelo de linguagem GPT-3.5 turbo.

[16] Selecionamos esse número de sentenças a fim de seguir a metodologia de Stanovsky *et al*. (2019).

Referências

Bender, E. M., Gebru, T., McMillan-Major, A., Shmitchell, S. (2021) On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. p. 610-623.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems,* 33, 1877-1901.

Caseli, H. de M. (2017) Tradução Automática: estratégias e limitações. *Domínios de Lingu@gem*, v. 11, n. 5, p. 1782-1796.

Castilho, S., Mallon, C., Meister, R., Yue, S. (2023) Do online machine translation systems care for context? What about a GPT model? In: 24th *Annual Conference of the European Association for Machine Translation* (EAMT 2023), 12-15 June 2023, Tampere, Finland. (In Press)

Cohen, J. A. (1960) Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, v. 20, n. 1, p. 37–46.

Devinney, H., Björklund, J., Björklund, H. (2022) Theories of "Gender" in NLP Bias Research. *arXiv*:2205.02526 [cs].

Halliday, M. K. (1978) *Language as social semiotic: The social interpretation of language and meaning*. London: Edward Arnold.

Jakobson, R. (1959) On Linguistic Aspects of Translation. In: Brower, R. A. (ed.). *On translation*. Cambridge, USA: Harvard University Press. https://doi.org/10.4159/harvard.9780674731615.c18. p. 232-239.

Kocmi, T., Federmann, C. (2023). Large language models are state-of-the-art evaluators of translation quality. *arXiv* preprint arXiv:2302.14520.

Levesque, H. J. (2011) The Winograd schema challenge. In: *AAAI Spring Symposium*: Logical Formalizations of Commonsense Reasoning.

Lewis, M., Lupyan, G. (2020) Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature Human Behaviour*, v. 4, n. 10, p. 1021-1028.

Popović, M., Castilho, S. (2019). Challenge Test Sets for MT Evaluation. In *Proceedings* of Machine Translation Summit XVII: Tutorial Abstracts, Dublin, Ireland. European Association for Machine Translation.

Rudinger, R., Naradowsky, J., Leonard, B., Van Durme, B. (2018) Gender Bias in Coreference Resolution. *arXiv*:1804.09301 [cs].

Savoldi, B., Gaido, M., Bentivogli, L., Negri, M., Turchi, M. (2021) Gender Bias in Machine Translation. *Transactions of the Association for Computational Linguistics*, v. 9, p. 845–874.

Stanovsky, G., Smith, N., Zettlemoyer, L. (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.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K. (2018) Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. *Proceedings* [...], volume 2 (Short Papers).