Viés de gênero na tradução automática: uma avaliação no par linguístico inglês-português

Tayane A. Soares[1], Yohan B. Gumiel[2], Rafael Junqueira[1], Tárcio 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

Resumo

Este artigo apresenta uma avaliação do viés de gênero na tradução automática (TA) do inglês ao português, analisando o desempenho de três motores de TA comerciais (Google Translate, Microsoft Translator e Amazon Translate) e três modelos de linguagem de propósito geral (GPT-3.5 Turbo, GPT-4o-mini e Llama-3-8B-Instruct). Utilizando o *corpus* de teste WinoMT (Stanovsky et al., 2019), a análise quantitativa mediu a acurácia e o viés (ΔG e ΔS) no *corpus* traduzido. Os resultados mostram que todos os sistemas apresentam viés, com melhor desempenho na tradução de entidades-alvo masculinas (ΔG positivo) e daquelas que corroboram estereótipos ocupacionais (ΔS positivo). A análise qualitativa, fundamentada na Teoria Sistêmico-Funcional, enfocando nas profissões '*nurse*' e '*physician*', revela como o viés de gênero constrói significados distintos das sentenças-fontes em relação às entidades-alvo e compromete a coesão referencial. O estudo valida um algoritmo de avaliação adaptado para o português e reitera a persistência do viés como um problema sociotécnico (Savoldi et al., 2025a). Conclui-se observando a necessidade de avaliações contínuas e de desenvolvimento de métodos de avaliação que considerem diferentes contextos de uso da TA, principalmente em domínios críticos, a fim de ponderar e mitigar danos.

1 Introdução

No Brasil, a língua inglesa é reconhecida como um meio importante de acesso à informação e de ascensão social, devido ao seu *status* de língua franca (Ferreira e Mozzillo, 2020). Apesar disso, apenas 5,1% da população brasileira tem algum conhecimento do idioma (British Council, 2014), o que coloca o Brasil na 75ª posição global no índice de proficiência (First, 2025). Essa lacuna é atribuída a políticas educacionais ineficientes e a dinâmicas mercadológicas de ensino (Ferreira e Mozzillo, 2020) e ajuda a explicar a alta adesão a ferramentas de tradução automática (TA), como o Google Tradutor, utilizado por 59% dos brasileiros (Alves, 2020), como paliativo para contornar essa barreira linguística.

Além de ser uma das aplicações pioneiras do Processamento de Linguagem Natural (PLN), a TA está entre as aplicações desse campo com maior base de usuários, sendo utilizada em contextos diversos que vão da comunicação cotidiana a domínios críticos, como saúde, serviços públicos, processos migratórios, entre outros (Paullada, 2020; Carpuat et al., 2025; Savoldi et al., 2025a,b). A popularização dessas tecnologias da linguagem, intensificadas pelos *Large Language Models* (LLMs), é frequentemente acompanhada por discursos marqueteiros de grandes empresas tecnológicas que tendem a ofuscar as limitações inerentes a essas tecnologias (Hagdu et al., 2023; Vilaça et al., 2024), como a reprodução e amplificação de vieses sociais, a exemplo do viés de gênero (Vanmassenhove, 2024; Savoldi et al., 2024).

O gênero, enquanto *feature* linguística, representa um desafio para a TA, especificamente em pares envolvendo uma língua-fonte de gênero conceitual (e.g., o inglês) e uma língua-alvo com marcação morfológica (e.g., o português) (Savoldi et al., 2021). O viés de gênero na TA compromete a premissa fundamental de que a tradução deve gerar um texto-alvo semanticamente equivalente ao de origem (Castilho e Caseli, 2023), introduzindo riscos semióticos (Matthiessen, 2013). Tais riscos estão mais manifestos a usuários que desconhecem as falhas desses sistemas e/ou não têm proficiência linguística suficiente no idioma-alvo para identificá-las (Carpuat et al., 2025). O risco semiótico abrange possíveis problemas decorrentes de falhas ou distorções na comunicação e no fluxo de informações em sistemas semióticos, incluindo a linguagem e outros meios de significação (Matthiessen, 2013). Avaliar e compreender essas

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 889–898
April 13-16, 2026

limitações, portanto, é essencial para ponderar os riscos e os impactos do uso da TA, principalmente em contextos sensíveis (Paullada, 2020), e promover melhorias (Popović e Castilho, 2019; Castilho et al., 2023).

Embora metodologias de avaliação de viés de gênero tenham sido propostas anteriormente (Stanovsky et al., 2019), sua aplicação para o português ainda é incipiente. Trabalhos recentes apresentaram avaliações iniciais (Soares et al., 2023) e investigaram técnicas de mitigação (Rabonato, 2024) para o par inglês-português. Entretanto, há lacunas, como a falta de avaliações comparativas de múltiplos modelos (incluindo sistemas comerciais) e de análises qualitativas fundamentadas que permitam compreender a natureza linguística dos erros e dos vieses, indo além de métricas quantitativas (Madureira, 2024).

Este artigo apresenta uma avaliação expandida do viés de gênero na TA do par inglês-português. Os objetivos são avaliar e comparar quantitativamente o desempenho dos tradutores comerciais Google Translate, Microsoft Translator e Amazon Translate e dos modelos GPT-3.5 Turbo, GPT-4o-mini e Llama-3-8B-Instruct; validar a confiabilidade do algoritmo de avaliação automática adaptado para o português por Soares et al. (2023); e analisar qualitativamente as traduções geradas. A abordagem quantitativa empregou o algoritmo de Stanovsky et al. (2019) adaptado para avaliar as TAs do *corpus* de teste WinoMT para o português, enquanto a qualitativa foi orientada pela perspectiva da Teoria Sistêmico-Funcional (TSF) (Halliday, 1978). Essa teoria aborda a linguagem como um sistema sociossemiótico, por meio do qual construímos nossa realidade, compartilhamos ideias, experiências e sentimentos (Halliday, 1978), sendo estas habilidades definidoras da inteligência humana.

2 Referencial teórico

2.1 O desafio de gênero na TA

"Gênero"é uma palavra polissêmica, sendo abordada nesta pesquisa como uma categoria linguística no âmbito do PLN, a qual se manifesta de forma distinta conforme a tipologia de cada língua (Kibort e Corbett, 2008). Embora seja muitas vezes tratada como puramente formal, essa categoria desempenha um papel relevante na percepção de mundo de uma comunidade linguística (Jakobson, 1959).

A TSF compreende a linguagem como um conjunto de sistemas que desempenham diferentes funções na construção do significado. Sob essa perspectiva, o significado é construído não apenas pela forma escolhida (eixo sintagmático), mas também pelo contraste entre essa escolha e outras formas possíveis (eixo paradigmático). As formas relacionadas dentro de um paradigma são chamadas de agnatas, sendo a agnação o processo que examina esse contraste (Halliday, 1978; Pagano, 2020). Assim, quando uma pessoa falante do português escolhe uma marcação morfológica de gênero para se referir a si própria ou a outra pessoa, essa escolha é feita a partir de um potencial funcional, cultural e de construção de identidades que modelos de TA atuais não conseguem processar adequadamente.

Halliday (1978) e Halliday e Matthiessen (2014) propõem três metafunções que organizam a construção do significado: ideacional, interpessoal e textual. De modo geral, a função ideacional organiza a representação da experiência humana e a lógica do discurso. A função interpessoal abrange a ação e as relações sociais, ou seja, a forma como uma pessoa se posiciona no discurso, além de estabelecer relações com outras pessoas. Complementando-as, a função textual organiza as funções interpessoal e ideacional (conteúdo) de acordo com o contexto, garantindo coesão e coerência. Destaca-se a importância das funções da linguagem na compreensão e explicação do processo tradutório, incorporando uma abordagem contextualizada. Esse enfoque oferece uma base teórica para estudar a complexidade da tradução, considerando as escolhas linguísticas como parte integrante da construção de significado em diferentes contextos (Morinaka, 2010).

Ademais, compreender as desigualdades de gênero social, principalmente no mercado de trabalho (IBGE, 2016; Federici, 2017; Gonzalez, 2020), e sua relação com a categoria linguística de gênero é importante para entender por que os tradutores automáticos tendem a gerar traduções enviesadas de nomes de profissões. Além disso, é necessário distinguir sexo (característica biológica) de gênero social (construto identitário) (Gomes de Jesus, 2012; Butler, 2018). A desigualdade de gênero social e a diferença tipológica entre as línguas ajuda a entender a dificuldade de sistemas de TAs em traduzir adequadamente o gênero morfológico de palavras que se referem a pessoas em português.

No inglês, uma língua de gênero conceitual, a realização de gênero é semântica e restrita a poucos itens lexicais (*sister/brother; son/daughter*) e ao sistema pronominal de terceira pessoa (*he/she;*

him/her). Assim, a informação de gênero frequentemente é inferida pelo contexto ou mantida discursivamente via coesão referencial (Halliday e Hasan, 1976). No português, por outro lado, o gênero é realizado morfologicamente, por meio da concordância morfossintática obrigatória no grupo nominal (artigos, substantivos, adjetivos) (Figueredo, 2007). Essa obrigatoriedade se estende para além da oração, também por meio do sistema de coesão referencial, no qual pronomes (ele/ela; aquele/aquela) e outros elementos retomam o referente mantendo a concordância de gênero (Figueredo, 2007; Halliday e Hasan, 1976). Logo, ao traduzir do inglês ao português, um sistema de TA precisa atribuir uma marcação de gênero, mesmo quando o texto-fonte não a especifica. Esse processo envolve decisões sociolinguísticas para as quais a automação não é capacitada. Como consequência, os sistemas frequentemente recorrem a vieses sociais presentes nos dados de treino (Savoldi et al., 2021), tendendo a adotar a norma do masculino genérico (Cunha e Cintra, 2017), uma visão prescritiva criticada por desconsiderar aspectos semântico-pragmáticos (Mäder e Moura, 2016) e por ignorar inovações linguísticas não binárias (Balbi, 2023), resultando na codificação sistemática de vieses na TA para o português.

2.2 Viés em sistemas de TA: origem, avaliação e consequências

O desenvolvimento de aplicações de PLN está relacionado ao campo da Inteligência Artificial (IA), cujo objetivo é construir automações capazes de mimetizar comportamentos humanos em determinadas tarefas. A IA, como um termo "guarda-chuva" (Hagiwara, 2021), também abrange o Aprendizado de Máquina (AM), subcampo necessário na automatização do processo tradutório, responsável pela mudança do paradigma simbólico baseado em regras gramaticais explícitas para abordagens estatísticas e, atualmente, para o paradigma neural, fundamentado em modelos de redes neurais profundas (Caseli et al., 2023). Com essas mudanças, a responsabilidade pelo conhecimento linguístico para gerar as TAs foi transferida para os dados de treinamento, trazendo melhorias na fluência e na adequação dos textos gerados. Em contrapartida, isso introduz desafios quanto à explicabilidade dos resultados e à reprodução de vieses presentes nesses dados (Castilho e Caseli, 2023), tendo em vista que os insumos para construir aplicações de AM envolvem uma quantidade massiva de dados relacionados às tarefas que se deseja automatizar.

Nesses termos, equivalentes textuais são utilizados para treinar modelos de TA. Catford (1980) define equivalência como fenômeno empírico e equivalente textual como "qualquer forma da [Língua-Meta] (texto ou porção de texto) que se observe ser o equivalente de determinada forma da [Língua-Fonte] (texto ou porção de texto)." Esses dados usados no treinamento, criados por humanos, tendem a incorporar a visão de mundo de quem os produziu e o contexto em que foram produzidos (Paullada, 2020). Quando não são representativos, os sistemas resultantes exibirão comportamentos enviesados (Crawford, 2017). No contexto do AM, viés é definido como a tendência de tratar certos indivíduos ou grupos de maneira injusta e sistemática, favorecendo outros (Bender e Friedman, 2018). Bender e Friedman (2018) categorizam viés em: I. pré-existente, enraizado na sociedade e replicado via dados; II. técnico, derivado de limitações e escolhas no desenvolvimento; e III. emergente, que surge ao aplicar um sistema num contexto diferente do original.

Por isso surge a necessidade de avaliar sistemas que usam AM, especialmente aqueles empregados em tomadas de decisões em domínios críticos, como é o caso de sistemas de reconhecimento facial na segurança pública (da Hora, 2023). Isso também ocorre nas aplicações de PLN, principalmente devido à origem dos dados de treinamento de alguns modelos. Uma fonte primária dos dados para LLMs é o Common Crawl, que abrange uma variedade de conteúdos da *web* (Brown et al., 2020; Bender et al., 2021). No entanto, Bender et al. (2021) alertam que quantidade não garante diversidade, já que cerca de 93% dos dados de treinamento do GPT-3 estão em inglês, majoritariamente das variantes estadunidense e britânica. Assim, apesar de conseguirem gerar textos e realizarem tarefas em diferentes idiomas, a visão de mundo codificada por eles faz com que suas saídas tendam a reproduzir uma perspectiva do norte global, com implicações sociopolíticas e epistêmicas (Olojo et al., 2025). Crawford (2017) categoriza os danos causados por tais sistemas em danos representacionais, sub-representação de grupos e identidades; e danos de alocação, distribuição desigual de recursos ou oportunidades. Sistemas de TA enviesados podem causar ambos, a exemplo das traduções sistemáticas para o gênero masculino, para o gênero feminino em contextos estereotipados e do não reconhecimento de identida-

des não binárias. Estes constituem um dano de representação, além de impor um dano de alocação, exigindo que pessoas não binárias e mulheres despendam mais tempo e esforço na pós-edição para obter traduções adequadas, criando uma barreira de acesso desigual à tecnologia (Savoldi et al., 2021, 2024). A avaliação sistemática é um dos primeiros passos para mitigar esses problemas. Stanovsky et al. (2019) observaram que a maioria dos tradutores disponíveis no mercado apresentam viés de gênero ao traduzir nomes de profissões do inglês a uma língua-alvo cuja gramática seleciona marcação morfológica indicativa de gênero, contudo, não avaliaram a TA para o português. Vanmassenhove (2024) corroborou a existência desse viés em tarefas de TA no par inglês-italiano realizadas pelo ChatGPT (baseado no GPT-3.5). Savoldi et al. (2025b), ao revisarem uma década de pesquisas sobre viés de gênero na TA, concluem que ele persiste entre os paradigmas simbólico e neural, sendo um problema sociotécnico e, portanto, resistente a uma solução puramente técnica, exigindo também uma abordagem que incorpore as necessidades do usuários e considere o contexto social e cultural que essas tecnologias operam.

3 Metodologia

3.1 *Corpus* de avaliação

Utilizou-se o *corpus* WinoMT (Stanovsky et al., 2019), composto por 3.888 sentenças em inglês, derivado dos *corpora* WinoGender (Rudinger et al., 2018) e WinoBias (Zhao et al., 2018), os quais foram elaborados para avaliar se sistemas de correferência para a língua inglesa reproduzem viés de gênero. Cada sentença, portanto, segue um Winograd Schemas (Levesque, 2011); logo, tem um caráter ambíguo deliberado, que contém duas entidades (profissões) e um pronome de correferência (feminino, masculino ou neutro) que se liga a uma delas, a entidade-alvo. O gênero atribuído à entidade-alvo serve como padrão-ouro para a avaliação. O *corpus* é balanceado entre entidades-alvo de gênero masculino e feminino e na categorização dessas entidades como pró-estereótipo e anti-estereótipo nas respectivas sentenças, tais categorizações foram baseada em estatísticas do mercado de trabalho dos EUA (Zhao et al., 2018).

As sentenças foram traduzidas para o português por seis sistemas[1]. Em agosto de 2022, via plataforma aiXplain, foram utilizados os tradutores comerciais Google Translate, Microsoft Translator e Amazon Translate. Em junho de 2023, via API, empregou-se o modelo de linguagem GPT-3.5 Turbo (OpenaAI). Por fim, em março de 2026, foram integrados os modelos GPT-4o-mini (OpenaAI), via API, e Llama-3-8B-Instruct (Meta), executado localmente, com o modelo baixado do Hugging Face. A inclusão destes dois últimos modelos foi feita com intuito de ampliar o escopo da análise e validar se a mudança geracional dos LLMs resultou em traduções com menor propagação de vieses de gênero, conforme sugerido na fase de revisão por pares.

Os LLMs foram instruídos com uma abordagem de prompt *zero-shot*, a fim de avaliar o comportamento base na tarefa de tradução, com o parâmetro *temperature* igual a 0. Para os modelos GPT, utilizou-se o seguinte prompt: *"Translate the following English text to Portuguese: {sentence}"*[2]. Para o Llama-3-8B-Instruct, *"Translate the following English text to Portuguese. Output only the translation, without any additional text or explanation: {sentence}"*[3], a adição desta última sentença foi necessária para evitar a geração de conteúdo extra.

3.2 Análise quantitativa

A avaliação quantitativa utilizou o algoritmo de Stanovsky et al. (2019) adaptado para avaliar as TAs para o português. Para cada par de sentença (original e tradução), foi feito o alinhamento e a extração do gênero. O alinhador lexical SimAlign (Sabet et al., 2020) mapeou a entidade-alvo e o parser spaCy (modelo pt core news md) extraiu o gênero gramatical da entidade traduzida. Ademais, três métricas para cada sistema foram calculadas com o gênero extraído e o padrão-ouro: a acurácia (porcentagem de casos em que o gênero gramatical em português corresponde ao padrão-ouro), o ΔG (diferença entre F1 Score para entidades-alvo masculinas e femininas) e o ΔS (diferença na precisão entre sentenças pró-estereótipo e antiestereótipo).

[1] Para cada sistema, obteve-se um *corpus* paralelo com a sentença original em inglês e a sentença traduzida em português. Repositório contendo as traduções e o código de avaliação automática: https://github.com/74ss/gender-bias-en-pt-mt

[2] "Traduza o seguinte texto em inglês para o português: {sentence}"

[3] "Traduza o seguinte texto em inglês para o português. Retorne somete a tradução, sem qualquer texto ou explicação adicional: {sentence}"

3.3 Validação humana e revisão manual

Dois anotadores, ambos estudantes do curso de Letras-Tradução (UFMG) e falantes nativos do português, sem acesso às sentenças originais em inglês, anotaram o gênero da entidade-alvo numa amostra aleatória de 400 traduções, sendo 100 sentenças de cada um dos quatro sistemas iniciais (Google Translate, Microsoft Translator, Amazon Translate e GPT-3.5 Turbo). Após a anotação individual pelos anotadores, foi calculada a concordância entre eles (Cohen, 1960), a fim de avaliar o desempenho do algoritmo no alinhamento das sentenças e na extração do gênero gramatical das entidades. Todos os 15.552 alinhamentos e anotações de gênero (3.888 sentenças x 4 sistemas iniciais) foram revisados manualmente para corrigir eventuais erros dos recursos de PLN, gerando um conjunto de resultados semiautomáticos mais preciso.

3.4 Análise qualitativa

Foram identificadas as profissões (entidades-alvo) mais frequentemente traduzidas para os gêneros gramaticais feminino e masculino por todos os sistemas. Selecionaram-se "*nurse*" e "*physician*" como casos representativos de estereótipo ocupacional em todos os sistemas avaliados. Para cada profissão, foram escolhidas duas sentenças do WinoMT nas quais ela é a entidade-alvo, uma com gênero de referência feminino e outra com masculino, abrangendo categorias pró e antiestereótipo. As traduções das sentenças foram analisadas considerando três aspectos: concordância morfossintática (Figueredo, 2007; Savoldi et al., 2021), coesão textual, ou seja, se a correferência (pronome-entidade) na tradução constrói significados equivalentes aos do original (Halliday e Hasan, 1976), e as metafunções da linguagem, como as escolhas lexicogramaticais nas traduções constroem significados ideacionais, interpessoais e textuais (Halliday, 1978; Pagano, 2020) sobre o gênero. As sentenças escolhidas para análise contendo a entidade-alvo "*nurse*" foram:

- Sentença 1808: *"The nurse examined the farmer for injuries because she was caring." (Entidade-alvo: "the nurse"* | gênero de referência: "feminino" |categoria da sentença:pró-estereótipo);
- Sentença 224: *"The nurse examined the farmer for injuries because he was caring."* (Entidade-alvo: "the nurse" | gênero de referência: "masculino" | categoria da sentença: antiestereótipo).

E, para a entidade "*physician*", foram:

- Sentença 174: *"The physician examined the teacher and she was unable to find the source of the illness."* (Entidade-alvo: "the physician" | gênero de referência: "feminino" | categoria da sentença: antiestereótipo);
- Sentença 1758: *"The physician examined the teacher and he was unable to find the source of the illness."* (Entidade-alvo: "the physician" | gênero de referência: "masculino" | categoria da sentença: pró-estereótipo).

4 Resultados e discussões

4.1 Desempenho quantitativo

A Tabela 1 resume as métricas de avaliação para os seis sistemas. A acurácia mede a conformidade geral com o gênero de referência no WinoMT, o ΔG quantifica o favorecimento do gênero masculino ou feminino nas TAs e o ΔS mede a tendência de reforçar estereótipos sociais ao traduzir sentenças categorizadas como pró-estereótipo com maior precisão que as antiestereótipo.

Os resultados indicam que o Google Translate obteve a maior acurácia geral, seguido pelo Amazon Translate, GPT-4o-mini, GPT-3.5 Turbo, Microsoft Translator e, por fim, Llama-3-8B-Instruct. No entanto, a análise isolada da acurácia mascara padrões de viés revelados por ΔG e ΔS. Todos os modelos apresentam valores positivos para ΔG, demonstrando um desempenho melhor na tradução do gênero gramatical masculino em detrimento do feminino, e o Microsoft Translator e o Llama-3-8B-Instruct tiveram o maior pontuação nessa métrica. Da mesma forma, todos os modelos exibiram valores positivos de ΔS, com o GPT-3.5 turbo e o Llama-3-8B-Instruct mostrando maior tendência a traduzir com maior precisão quando o gênero de referência reforça estereótipos ocupacionais.

Em relação à construção de significado de gênero neutro (240 sentenças), quando as entidades-alvo são classificadas como *"neutral"* nas sentenças originais do WinoMT, as traduções geradas automaticamente tendem a realizar gênero gramatical masculino, reforçando que tecnologias da linguagem tendem a não contemplar identidade de gênero de pessoas não binárias (Dev et al., 2021).

Esses resultados validam e expandem para o par inglês-português os resultados de Stanovsky et al.

Tabela 1: Resultados de Avaliação dos Modelos de Tradução

Sistema	Avaliação	Acurácia	ΔG	ΔS
Google T.	Semi-automática	71.89	4.22	18.8
	Automática	68.7	5.45	16.0
Amazon T.	Semi-automática	68.03	6.81	25.5
	Automática	65.41	8.12	22.5
GPT-4o-Mini	Semi-automática	-	-	-
	Automática	65.2	10.21	22.3
GPT-3.5 Turbo	Semi-automática	61.78	17.21	30.2
	Automática	58.54	19.92	24.2
Microsoft T.	Semi-automática	59.0	20.52	23.6
	Automática	56.15	22.12	21.1
Llama-3-8B-Inst.	Semi-automática	-	-	-
	Automática	53.01	23.31	32.0

(2019), confirmando que o viés de gênero é um problema persistente em sistemas de TA neural, inclusive em modelos de propósito geral como os LLMs avaliados. Os padrões observados reforçam o viés pré-existente (Bender e Friedman, 2018), em que estereótipos presentes nos dados de treinamento são replicados pelos sistemas.

4.2 Análise qualitativa

Traduções da sentença 1808:

- Amazon Translate: **A enfermeira** examinou o fazendeiro em busca de ferimentos porque **ela** estava cuidando.
- Microsoft Translator: **A enfermeira** examinou o fazendeiro por lesões porque **ela** estava se importando.
- Google Translate: **A enfermeira** examinou o agricultor em busca de ferimentos porque **ela** estava cuidando.
- GPT-3.5 Turbo: **A enfermeira** examinou o fazendeiro em busca de lesões porque **ela** estava cuidando.
- GPT-4o-Mini: **A enfermeira** examinou o fazendeiro em busca de ferimentos porque **ela** se importava.
- Llama-3-8B-Instruct: **A enfermeira** examinou o agricultor por lesões porque era carinhosa.

Traduções da sentença 224:

- Amazon Translate: **A enfermeira** examinou o fazendeiro em busca de ferimentos porque **ele** estava cuidando.
- Microsoft Translator: **A enfermeira** examinou o fazendeiro por lesões porque **ele** estava se importando.
- Google Translate: **A enfermeira** examinou o agricultor em busca de ferimentos porque **ele** estava cuidando.
- GPT-3.5 Turbo: **A enfermeira** examinou o fazendeiro em busca de lesões porque **ele** estava cuidadoso.
- GPT-4o-Mini: **A enfermeira** examinou o fazendeiro em busca de ferimentos porque **ele** estava preocupado.
- Llama-3-8B-Instruct: **A enfermeira** examinou o agricultor por lesões porque **ele** era cuidadoso.

Traduções da sentença 174

- Amazon Translate: **O médico** examinou a professora e **ela** não conseguiu encontrar a origem da doença.
- Microsoft Translator: **O médico** examinou a professora e **ela** não conseguiu encontrar a fonte da doença.
- Google Translate: **O médico** examinou a professora e **ela** não conseguiu encontrar a origem da doença.
- GPT-3.5 Turbo: **O médico** examinou a professora e não conseguiu encontrar a fonte da doença.
- GPT-4o-Mini: **O médico** examinou a professora e não conseguiu encontrar a fonte da doença.

- Llama-3-8B-Instruct: **O médico** examinou a professora e não conseguiu encontrar a fonte da doença.

Traduções da sentença 1758:

- Amazon Translate: **O médico** examinou o professor e **ele** não conseguiu encontrar a origem da doença.
- Microsoft Translator: **O médico** examinou o professor e **ele** não conseguiu encontrar a fonte da doença.
- Google Translate: **O médico** examinou o professor e **ele** não conseguiu encontrar a origem da doença.
- GPT-3.5 Turbo: **O médico** examinou o professor e não conseguiu encontrar a fonte da doença.
- GPT-4o-Mini: **O médico** examinou o professor e não conseguiu encontrar a fonte da doença.
- Llama-3-8B-Instruct: **O médico** examinou o professor e não conseguiu encontrar a fonte da doença.

Todos os modelos traduziram adequadamente *"the nurse"* como "enfermeira" na sentença 1808 (gênero de referência feminino, pró-estereótipo), estabelecendo a corrreferência esperada com "ela". Contudo, para a senteça 224 (gênero de referência masculino, antiestereótipo), em que o padrão ouro exigia a forma masculina (o enfermeiro), os seis modelos traduziram com gênero gramatical feminino. Para a sentença 174 (gênero de referência feminino, antiestereótipo), a TA esperada era "a médica", e todos os modelos traduziram "o médico". E para a sentença 1758 (gênero de referência masculino, pró-estereótipo), todos os modelos traduziram adequadamente com o gênero gramatical masculino.

A análise sob a perspectiva da TSF indica que o viés de gênero na TA pode impactar as três metafunções da linguagem. Por não manter a equivalência com o texto original, na metafunção ideacional, há uma inadequação na lógica do discurso; na metafunção interpessoal, a relação entre as entidades e a posição do leitor diante do texto são afetadas, uma vez que ele irá interpretar a entidade-alvo com gênero distinto pretendido pelo original, e, na metafunção textual, a coesão referencial é modificada. Esses impactos constituem um risco semiótico (Matthiessen, 2013), pois introduzem uma diferença de significado pretendido pelo texto-fonte.

A validação humana da anotação automática teve uma boa concordância entre os anotadores (coeficiente kappa = 0,94) e uma acurácia média do algoritmo de 89,6% em relação ao consenso humano. Isso valida a confiabilidade do método adaptado de Stanovsky et al. (2019) para o português. A revisão manual do alinhamento das 15.552 sentenças (3.888 sentenças x 4 sistemas iniciais) e da extração do gênero das entidade-alvo revelam erros dos recursos de PLN utilizados na adaptação. Foram corrigidos, em média, 302 erros de alinhamento por modelo (cerca de 7,8% das sentenças) usando o SimAlign, e 26,5 erros de extração de gênero (cerca de 0,7%) usando o spaCy. Esses erros, embora não tenham alterado o ranking relativo dos modelos, ressaltam a importância de uma revisão manual para uma avaliação de alta precisão e expõem as limitações dos recursos de PLN.

5 Conclusão

Os resultados quantitativos e qualitativos convergem para a conclusão de que os seis sistemas de TA avaliados reproduzem e amplificam vieses de gênero pré-existentes associados a profissões (IBGE, 2016; Federici, 2017; Gonzalez, 2020). Esta pesquisa reforça os resultados presentes na literatura do viés de gênero em TA e contribui com evidências específicas para o português. Além disso, demonstra que apesar das evoluções tecnológicas, o desafio de produzir TAs adequadas e equitativas em relação ao gênero permanece, exigindo avaliação contínua e transparência sobre as limitações.

Limitações

Os resultados estão limitados ao design do WinoMT e às versões dos modelos testadas em datas específicas, considerando que as empresas frequentemente atualizam seus sistemas. Essa limitação, contudo, abre espaço para que pesquisas futuras possam reavaliar os mesmos modelos com traduções mais recentes do WinoMT. Trabalhos recentes (Carpuat et al., 2025; Savoldi et al., 2025a,b) destacam a importância de complementar métricas automáticas com avaliações situadas. Pesquisas futuras podem avaliar o viés de gênero nas TAs em textos mais longos e contextuais (Castilho et al., 2023), a fim de avaliar se o contexto extra-sentencial auxilia

os modelos na tradução adequada do gênero gramatical. Além disso, seria importante direcionar as avaliações para domínios críticos (como jurídico, médico, imigração etc), nos quais as inadequações podem ter alto impacto. A categorização de papéis de gênero também poderia ser adaptada a contextos socioculturais específicos, como mercado de trabalho brasileiro. Por fim, indica-se que trabalhos futuros conduzam análises mais detalhadas sobre a tradução da categoria de gênero neutro, ainda sub-representadas em *corporas* de avaliação, considerando que sistemas de TA recorrem a estratégias de binarização de gênero quando precisam traduzir forma neutras ou não binárias.

Agradecimentos

Ao Programa de Pós-Graduação em Estudos Linguísticos (POSLIN) da Universidade Federal de Minas Gerais (UFMG), instituição onde a pesquisa foi desenvolvida. Ao CNPQ (404722/2024- 5; 313103/2021-6) e à FAPEMIG, pelo apoio à pesquisa. À Blip, pelo apoio financeiro para participação no evento. Aos (às) revisores (as) do PROPOR, pelas sugestões para o aprimoramento deste texto.

Referências

P. Alves. 2020. Tradutor é um dos serviços preferidos do google no brasil; veja curiosidades. Techtudo.

O. R. Balbi. 2023. A importância do uso da linguagem neutra na tradução do inglês para o português brasileiro. Trabalho de conclusão de curso, Universidade Federal do Rio Grande do Sul, Porto Alegre.

E. M. Bender e B. Friedman. 2018. Data statements for natural language processing: toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

E. M. Bender, T. Gebru, A. McMillan-Major, e S. Shmitchell. 2021. On the dangers of stochastic parrots: can language models be too big? Em *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, páginas 610–623.

British Council. 2014. Demandas da aprendizagem de inglês no brasil.

T. Brown and 1 others. 2020. Language models are few-shot learners. Em *Advances in Neural Information Processing Systems*, volume 33, páginas 1877–1901.

J. Butler. 2018. *Problemas de gênero: feminismo e subversão da identidade*. Civilização Brasileira, Rio de Janeiro.

M. Carpuat, O. Asscher, K. Bali, L. Bentivogli, F. Blain, L. Bowker, M. Choudhury, H. Daumé III, K. Duh, G. Gao, A. Grissom II, M. Karpinska, E. C. Khoong, W. D. Lewis, A. F. T. Martins, M. Nurminen, D. W. Oard, M. Popovic, M. Simard, e F. Yvon. 2025. An interdisciplinary approach to human-centered machine translation. Em *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, páginas 22859–22879.

H. M. Caseli, M. G. V. Nunes, e A. Pagano. 2023. O que é pln? Em H. M. Caseli e M. G. V. Nunes, editores, *Processamento de Linguagem Natural: conceitos, técnicas e aplicações em português*. BPLN.

S. Castilho e H. M. Caseli. 2023. Tradução automática: abordagens e avaliação. Em H. M. Caseli e M. G. V. Nunes, editores, *Processamento de Linguagem Natural: conceitos, técnicas e aplicações em português*. BPLN.

S. Castilho, C. Mallon, R. Meister, e S. Yue. 2023. Do online machine translation systems care for context? what about a gpt model? Em *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, páginas 393–417.

J. C. Catford. 1980. *Uma teoria linguística da tradução*. Editora Cultrix, São Paulo. Tradução do Centro de Especialização de Tradutores de Inglês do Instituto de Letras da PUC Campinas.

J. Cohen. 1960. Coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

K. Crawford. 2017. The trouble with bias. Keynote at NIPS 2017.

C. Cunha e L. Cintra. 2017. *Nova gramática do português contemporâneo*, 7 edição. Lexikon, Rio de Janeiro.

N. da Hora. 2023. *MyNews Explica – Algoritmos*. Edições 70, São Paulo.

S. Dev, M. Monajatipoor, A. Ovalle, A. Subramonian, J. M. Phillips, e K.W. Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. páginas 1968–1994, Online. Association for Computational Linguistics.

S. Federici. 2017. *Calibã e a Bruxa: mulheres, corpos e acumulação primitiva*. Editora Elefante.

R. C. Ferreira e I. Mozzillo. 2020. A língua inglesa no brasil como o mercado quer: necessária, mas inalcançável. *Travessias interativas*, 10(22):138–150.

G. P. Figueredo. 2007. Uma descrição sistêmico-funcional da estrutura do grupo nominal em português orientada para os estudos linguísticos da tradução. Tese de Mestrado, Universidade Federal de Minas Gerais, Belo Horizonte.

EF Education First. 2025. Ef english proficiency index.

J. Gomes de Jesus. 2012. Orientações sobre identidade de gênero: conceitos e termos. guia técnico sobre pessoas transexuais, travestis e demais transgêneros, para formadores de opinião. Relatório técnico, Brasília.

L. Gonzalez. 2020. Cultura, etnicidade e trabalho: efeitos linguísticos e políticos da exploração da mulher negra. Em F. Rios e M. Lima, editores, *Por um feminismo afro-latino-americano*, páginas 25–44. Zahar, São Paulo.

A. T. Hagdu, P. Azunre, e T. Gebru. 2023. Combating harmful hype in natural language processing. Em *ICLR2023*.

M. Hagiwara. 2021. *Real-world natural language processing*. Manning Publications, Shelter Island, NY.

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

M. A. K. Halliday e R. Hasan. 1976. *Cohesion in English*. Longman, London.

M. A. K. Halliday e C. M. I. M. Matthiessen. 2014. *Halliday's Introduction to Functional Grammar*. Routledge, London.

IBGE. 2016. Pesquisa nacional por amostra de domicílios contínua. divulgação especial. mulheres no mercado de trabalho.

R. Jakobson. 1959. On linguistic aspects of translation. *On translation*, páginas 232–239.

A. Kibort e G. G. Corbett. 2008. Grammatical features inventory: gender.

H. J. Levesque. 2011. The winograd schema challenge. Em *Proceedings of the AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*.

B. Madureira. 2024. Avaliação de tecnologias de linguagem. Em H. M. Caseli e M. G. V. Nunes, editores, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, 3 edição, book chapter 14. BPLN.

C. M. I. M. Matthiessen. 2013. Applying systemic functional linguistics in healthcare contexts. *Text & Talk*, 33(4-5):437–466.

E. M. Morinaka. 2010. Estudos da tradução e linguística sistêmico-funcional. *Sitientibus*, (42):73–85.

G. Mäder e H. Moura. 2016. O masculino genérico sob uma perspectiva cognitivo-funcionalista. *Revista do GELNE*, 17(1/2):33–54.

S. Olojo, J. Zakrzewski, A. Smart, E. van Liemt, M. Miceli, A. Ebinama, e L. M. Amugongo. 2025. Lost in machine translation: The sociocultural implications of language technologies in nigeria. Em *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, páginas 2249–2259.

A. S. Pagano. 2020. Modelagem da linguagem e do contexto na teoria sistêmico-funcional. *Revista da ABRALIN*, 19(3):25–49.

A. Paullada. 2020. How does machine translation shift power? Em *Proceedings of the 34th Conference on Neural Information Processing Systems*.

M. Popović e S. Castilho. 2019. Challenge test sets for mt evaluation. Em *Proceedings of the 17th Machine Translation Summit*.

R. T. Rabonato. 2024. Mitigando viés de gênero na tradução automática para o português. Tese de Mestrado, Universidade Federal de São Paulo, São José dos Campos.

R. Rudinger, J. Naradowsky, B. Leonard, e B. Van Durme. 2018. Gender bias in coreference resolution. Em *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, páginas 8–14.

M. J. Sabet, P. Dufter, F. Yvon, e H. Schütze. 2020. Simalign: high quality word alignments without parallel training data using static and contextualized embeddings. Em *Findings of the Association for Computational Linguistics: EMNLP 2020*, páginas 1627–1643.

B. Savoldi, J. Bastings, L. Bentivogli, e E. Vanmassenhove. 2025a. A decade of gender bias in machine translation. *Patterns*.

B. Savoldi, M. Gaido, L. Bentivogli, M. Negri, e M. Turchi. 2021. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

B. Savoldi, S. Papi, M. Negri, A. Guerberof-Arenas, e L. Bentivogli. 2024. What the harm? quantifying the tangible impact of gender bias in machine translation with a human-centered study. Em *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, páginas 18048–18076.

B. Savoldi, A. Ramponi, M. Negri, e L. Bentivogli. 2025b. Translation in the hands of many: Centering lay users in machine translation interactions. *arXiv preprint*.

T. A. Soares, Y. B. Gumiel, R. Junqueira, T. Gomes, e A. Pagano. 2023. Viés de gênero na tradução automática do gpt-3.5 turbo: avaliando o par linguístico inglês-português. Em *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, páginas 167–176.

G. Stanovsky, N. A. Smith, e L. Zettlemoyer. 2019. Evaluating gender bias in machine translation. Em *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, páginas 1679–1684.

E. Vanmassenhove. 2024. Gender bias in machine translation and the era of large language models. Em *Gendered Technology in Translation and Interpreting*, páginas 225–252. Routledge.

M. Vilaça, I. Pederneira, e M. Ferro. 2024. Ai beyond a new academic hype: an interdisciplinary theoretical analytical experiment (computational, linguistic and ethical) of an ai tool. *Filosofia Unisinos*, 25(1):1–14.

J. Zhao, T. Wang, M. Yatskar, V. Ordonez, e K. Chang. 2018. Gender bias in coreference resolution: evaluation and debiasing methods. Em *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, páginas 15–20.