Conectando Personagens, Revelando Gêneros: Uma Análise Computacional da Literatura em Português

Bárbara Duarte[1], Gabriella L. Araujo[1], Marcele Araponga[1], Mariana O. Silva[1], Michele A. Brandão[1], Mirella M. Moro[1]

[1]Department of Computer Science
Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG – Brazil

{barbaraduarte,gabriellalima}@dcc.ufmg.br,
louise20marcele03@ufmg.com.br,
{mariana.santos,michele.brandao,mirella}@dcc.ufmg.br

Abstract. *Gender representation in literary pieces is fundamental for understanding how narratives reflect social norms. This study investigates gender dynamics in Portuguese-language literature through two analyses. By using 551 works from the 18th to 20th centuries, we model character relationships as co-occurrence networks, revealing structural gender imbalance: male-male interactions dominate, and male characters tend to exhibit higher agency. Through an author-gender-balanced corpus of 58 works, we show that the author's gender significantly shapes character representation: male authors construct male-dominated networks, while female authors create more balanced ones.*

Resumo. *A representação de gênero em obras literárias é fundamental para compreender como narrativas refletem normas sociais. Este trabalho investiga a dinâmica de gênero na literatura de língua portuguesa por meio de duas análises. Utilizando 551 obras dos séculos XVIII ao XX, modelamos relações entre personagens como redes de coocorrência, revelando um desequilíbrio estrutural: interações entre homens dominam, e personagens masculinos apresentam maior agência. Utilizando um corpus balanceado por gênero autoral de 58 obras, mostramos que o gênero do autor influencia significativamente a representação dos personagens: autores homens constroem redes dominadas por personagens masculinos, enquanto autoras criam redes mais equilibradas.*

1. Introdução

A representação de gênero em obras culturais, como a literatura, desempenha um papel central na compreensão de como narrativas refletem, reproduzem e transformam estruturas sociais. Nesse contexto, abordagens computacionais têm se consolidado como ferramentas fundamentais para a análise em larga escala de textos literários. Em particular, a Análise de Redes Sociais (ARS) permite modelar personagens e suas interações como grafos, oferecendo uma perspectiva estrutural sobre a organização das narrativas e os padrões relacionais que emergem ao longo do enredo [Labatut and Bost 2019].

Diversos estudos demonstraram que personagens masculinos ocupam sistematicamente posições de maior destaque na ficção de língua inglesa, com uma proporção aproximada de dois homens para cada mulher no elenco de personagens, e que interações entre pares de mulheres são consistentemente raras [Kraicer and Piper 2019]. Trabalhos

recentes em larga escala revelaram ainda a existência de um *gender agency gap*, no qual personagens femininas tendem a ser retratadas de forma mais passiva, especialmente em obras de autores homens [Kreuzhage 2024]. Resultados semelhantes foram encontrados na literatura francesa, tanto em padrões de interação quanto na presença de estereótipos linguísticos associados a gênero [Vianne et al. 2023].

Embora existam muitos estudos sobre redes sociais na literatura estrangeira, notou-se uma escassez de trabalhos que abordem o contexto da literatura em língua portuguesa associado às discussões sobre gênero. Trabalhos recentes [Silva et al. 2023, Silva and Moro 2024, Silva 2025] estabeleceram bases metodológicas importantes para a extração de redes de personagens e a análise textual nesse idioma. No entanto, ainda é necessária uma investigação sistemática que explore não apenas os padrões estruturais de interação entre personagens, mas também o papel do gênero do autor na configuração dessas dinâmicas narrativas.

O presente trabalho tem dois objetivos complementares. Primeiro, analisar a presença de viés de gênero em redes de personagens extraídas de um corpus composto por 551 obras da literatura em língua portuguesa dos séculos XVIII a XX, analisando padrões de interação, centralidade e agência narrativa. Segundo, investigar o impacto do gênero do autor na construção dessas redes, a partir de um corpus balanceado de 58 obras, comparando sistematicamente narrativas escritas por autores e autoras. Os resultados revelam um desequilíbrio estrutural no corpus completo, com interações dominadas por personagens masculinos e concentração masculina nas posições de maior centralidade, e mostram que o gênero do autor influencia significativamente esses padrões: a composição da elite de centralidade e as dinâmicas de homofilia se invertem conforme o gênero autoral.

2. Trabalhos Relacionados

A análise computacional de textos científicos e literários tem se consolidado como uma área relevante nas humanidades digitais, especialmente com o uso de métodos de ARS. Na ciência, coautorias formam redes cujas relações podem ser analisadas sob diversas dimensões, revelando índices inesperados, e.g., [Brandão et al. 2016, Digiampietri et al. 2012, Lopes et al. 2010]. Na literatura, narrativas podem ser modeladas como grafos, nos quais personagens são representados como nós e suas interações como arestas, permitindo investigar a organização estrutural dos enredos e os padrões relacionais que emergem ao longo da narrativa [Labatut and Bost 2019, Silva et al. 2023]. Essa abordagem tem sido aplicada em diferentes tarefas, como sumarização, recomendação, classificação de textos e identificação de papéis narrativos.

Nos últimos anos, estudos utilizaram redes de personagens para investigar vieses sociais em larga escala. Na literatura de língua inglesa, [Kraicer and Piper 2019] demonstraram a existência de uma hierarquia de gênero, com predominância de personagens masculinos e baixa frequência de interações entre personagens femininas. De forma similar, [Kreuzhage 2024], ao analisar mais de 87 mil obras, identificou um *gender agency gap*, evidenciando que personagens femininas tendem a ocupar papéis mais passivos, especialmente em narrativas escritas por autores homens. Resultados semelhantes foram observados na literatura francesa por [Vianne et al. 2023], que destacaram a presença de estereótipos de gênero tanto em padrões de interação quanto em estruturas linguísticas.

No contexto brasileiro, a aplicação de ARS a narrativas tem sido explorada prin-

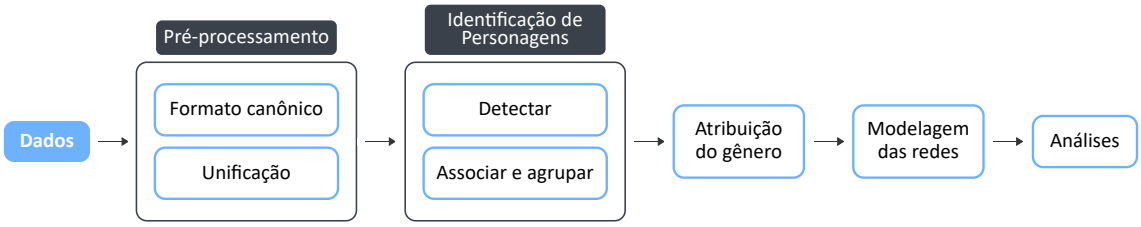

Figura 1. Metodologia para a construção e análise das redes de personagens.

cipalmente sob a perspectiva estrutural. [Aires et al. 2017] analisaram redes de personagens em adaptações cinematográficas, investigando métricas como grau, intermediação e PageRank, enquanto [Ribeiro et al. 2016] utilizaram redes literárias como um meio para estudar propriedades de redes complexas. Embora esses trabalhos demonstrem a viabilidade da ARS para análise de narrativas, eles não abordam explicitamente questões relacionadas a viés de gênero.

Mais recentemente, estudos focados na literatura em língua portuguesa começaram a integrar ARS e Processamento de Linguagem Natural (PLN). Por exemplo, [Silva et al. 2023] propuseram pipelines para extração e análise de redes de personagens, utilizando modelos como BERTimbau com CRF para reconhecimento de entidades nomeadas. Também existem métodos para atribuição de gênero e cálculo de agência narrativa [Silva and Moro 2024, Silva 2025]. Esses trabalhos estabelecem uma base metodológica importante para o estudo computacional da literatura em português.

No entanto, apesar desses avanços, a literatura ainda carece de investigações que articulem, de forma sistemática, a análise estrutural de redes de personagens com o papel do gênero do autor na configuração dessas redes. Em particular, permanece pouco explorado como o gênero autoral pode influenciar padrões de interação, centralidade e agência narrativa. Este trabalho avança nesse sentido ao combinar análise estrutural em larga escala com uma comparação controlada entre obras de autores e autoras.

3. Metodologia

Esta seção descreve o processo metodológico adotado neste estudo. O fluxo de trabalho foi implementado em *Python* e é composto por cinco etapas principais: (i) curadoria e pré-processamento do corpus, (ii) identificação e normalização de personagens, (iii) atribuição de gênero, (iv) modelagem das interações como redes sociais e (v) análise estrutural das redes. A Figura 1 apresenta uma visão geral do pipeline metodológico.

3.1. Dados

A base de dados utilizada neste estudo foi extraída do PPORTAL (Public Domain Portuguese Language Literature Dataset), um repositório digital que reúne obras literárias de domínio público [Silva et al. 2021, Silva et al. 2022]. O conjunto de dados inicial contém mais de 600 textos dos séculos XVIII a XX, com predominância de autores brasileiros.

Para analisar o efeito do gênero do autor, foi construído um corpus balanceado por gênero autoral. A partir do corpus final de 551 obras, foram selecionadas todas as obras escritas por mulheres (27 títulos). Em seguida, para cada autora, foi selecionado um autor homem com produtividade equivalente (i.e., com pelo menos o mesmo número de obras no corpus) e realizada uma amostragem aleatória da mesma quantidade de obras.

Esse procedimento resultou em um corpus balanceado de 58 obras (27 de autoras e 31 de autores), preservando a distribuição de produtividade autoral entre os grupos.

3.2. Pré-processamento

Para garantir a consistência dos dados, foi realizado um pré-processamento nos mesmos. Inicialmente, os nomes dos autores foram normalizados para um formato canônico (nome completo), permitindo o agrupamento correto das obras. Em seguida, diferentes edições da mesma obra foram identificadas e unificadas, mantendo-se apenas a versão mais antiga disponível, como aproximação da edição original. Esse processo evitou a super-representação de obras e resultou em um corpus final com 551 títulos únicos.

3.3. Identificação de Personagens

A identificação de personagens foi realizada utilizando o pipeline de PLN proposto por [Silva and Moro 2024, Silva 2025]. Menções a personagens foram detectadas por um modelo de Reconhecimento de Entidades Nomeadas (NER) baseado na arquitetura BERTimbau com camada CRF. Para consolidar diferentes menções ao mesmo personagem (e.g., "Bentinho" e "Bento Santiago"), aplicou-se normalização por similaridade textual com a biblioteca *thefuzz*, agrupando menções em entidades canônicas. Esse procedimento não realiza resolução completa de correferências (e.g., associar pronomes à entidade correspondente), tarefa que permanece desafiadora mesmo em outros idiomas [Krug et al. 2015].

3.4. Atribuição de Gênero

O gênero de cada personagem foi inferido por meio de um processo em duas etapas, baseado em [Silva 2025]. A primeira etapa analisa dependências sintáticas do texto (e.g., concordância nominal em artigos e adjetivos) para inferir o gênero. Para os casos não resolvidos, uma segunda etapa foi executada utilizando a biblioteca *gender-br* para inferir o gênero com base no primeiro nome do personagem.

A classificação adotada é estritamente binária (masculino/feminino), refletindo as convenções literárias predominantes no período do corpus (séculos XVIII a XX), no qual representações de identidades de gênero não-binárias são essencialmente ausentes. Além disso, as ferramentas de PLN utilizadas operam sobre a concordância gramatical da língua portuguesa, que se estrutura nessas duas categorias. Personagens cujo gênero não pôde ser inferido foram classificados como "desconhecido" e excluídos das análises comparativas.

3.5. Modelagem das Redes

Para cada obra, foi modelada uma rede de interação social utilizando a biblioteca *NetworkX*. Cada rede é representada por um grafo não direcionado e ponderado $G = (V, E)$, onde V corresponde aos 20 personagens canônicos mais frequentes de cada obra e E às interações entre eles. Uma aresta (u, v) é criada quando dois personagens coocorrem na mesma sentença,[1] com peso $w(u, v)$ igual ao número de coocorrências.

[1]A coocorrência em nível de sentença é utilizada como *proxy* para interação social. Essa abordagem, embora amplamente adotada na literatura [Labatut and Bost 2019], pode superestimar relações entre personagens que são apenas mencionados no mesmo contexto sem interação direta. As implicações dessa limitação são discutidas na Seção 6.

3.6. Análise de Redes Sociais

As redes construídas foram analisadas por meio de três dimensões complementares do viés de gênero: proporção de interações, importância estrutural dos personagens e agência narrativa.

Proporção de Interações. Foi calculada a proporção de arestas entre personagens do mesmo gênero (homofilia) e de gêneros distintos (heterofilia). As arestas foram classificadas em três categorias: *Homem–Homem*, *Mulher–Mulher* e *Homem–Mulher*.

Importância dos Personagens. A proeminência de personagens masculinos e femininos foi avaliada por meio de quatro métricas de centralidade: grau (*degree*), que mede o número de conexões diretas; intermediação (*betweenness*), que indica o papel de "ponte" entre grupos; proximidade (*closeness*), que captura a distância a todos os demais nós; e autovetor (*eigenvector*), que avalia a influência considerando a importância dos vizinhos.

Agência Narrativa. A agência de cada personagem foi quantificada com base em sua relação sintática com os verbos nas sentenças, utilizando o módulo de análise de dependências proposto por [Silva 2025]. Para cada personagem, foram contabilizados o *subject_count* (ocorrências como sujeito gramatical) e o *object_count* (ocorrências como objeto gramatical). A pontuação de agência (*agency_score*) é calculada como [Vianne et al. 2023]:

$$agency\_score = \frac{subject\_count - object\_count}{subject\_count + object\_count}$$

Essa métrica assume valores no intervalo $[-1, 1]$, em que valores próximos de 1 indicam predominância de papéis ativos e valores próximos de -1 indicam predominância de papéis passivos.

4. Resultados

Esta seção apresenta os resultados organizados em três partes. Inicialmente, são caracterizadas as redes construídas a partir do corpus (Seção 4.1). Em seguida, é analisado o viés de gênero no corpus completo de 551 obras (Seção 4.2). Por fim, é investigada a influência do gênero do autor na configuração dessas redes, com base no corpus balanceado de 58 obras (Seção 4.3).

4.1. Caracterização das Redes

O corpus completo é composto por 551 títulos únicos, abrangendo obras publicadas entre 1752 e 1970. O pipeline de extração identificou mais de 106 mil menções a personagens, das quais 51,4% foram classificadas como masculinas, 38,9% como femininas e 9,6% permaneceram com gênero desconhecido. Para cada obra, foi construída uma rede de coocorrência com os 20 personagens mais frequentes, totalizando 14.041 arestas.

No corpus balanceado (58 obras), redes de autoras são, em média, maiores (17,2 nós e 47,3 arestas) do que as de autores (13,9 nós e 34,1 arestas), embora apresentem densidade similar. Além disso, observa-se uma diferença consistente na composição do elenco: obras de autores homens apresentam predominância de personagens masculinos (61,1%), enquanto obras de autoras apresentam maioria feminina (56,7%) (Figura 2).

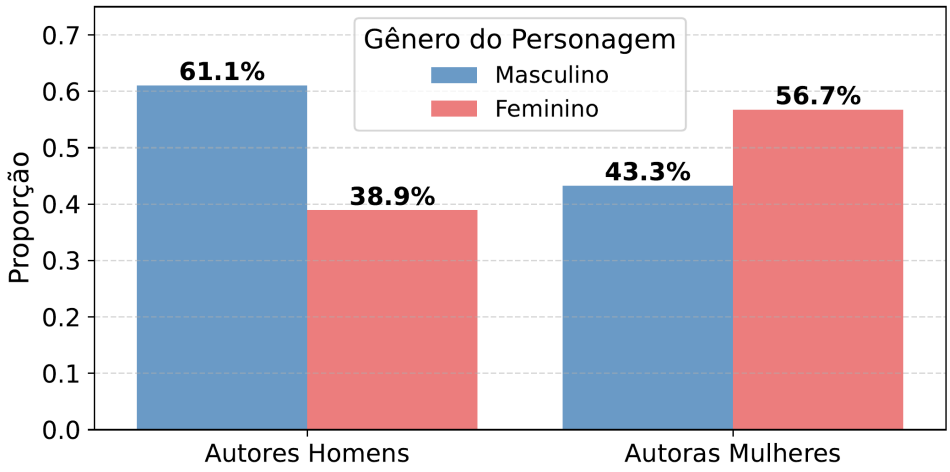

Figura 2. Proporção de personagens masculinos e femininos entre os 20 mais mencionados, por gênero do autor.

4.2. Análise Geral do Viés de Gênero

Esta seção investiga a presença de viés de gênero nas redes de personagens construídas a partir do corpus completo (551 obras). A análise é conduzida a partir de três dimensões complementares: padrões de interação, importância estrutural e agência narrativa.

4.2.1. Proporção de Interações

A Figura 3 apresenta a distribuição das proporções de interações por tipo em cada obra. Observa-se que as interações do tipo *Homem–Homem* apresentam maior concentração e densidade, com média de 48,5% e mediana de 45,9%, indicando que, em média, quase metade das interações ocorre entre personagens masculinos. Esse resultado evidencia a centralidade de núcleos masculinos na estrutura das narrativas.

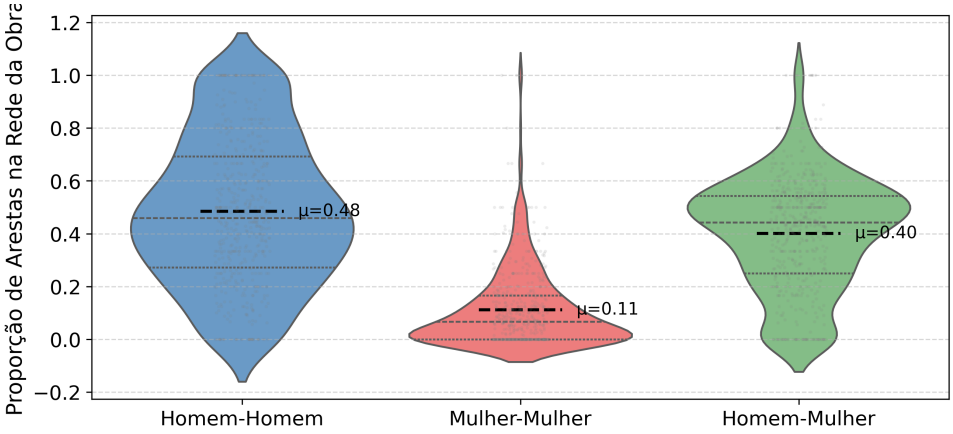

Figura 3. Distribuição das proporções de interação *Homem–Homem*, *Mulher–Mulher* e *Homem–Mulher* nas obras.

Por outro lado, interações *Mulher–Mulher* apresentam distribuição assimétrica e concentrada em valores baixos (média: 11,2%; mediana: 6,7%), indicando que redes exclusivamente femininas são pouco frequentes e ocupam posição periférica nas narrativas. As interações mistas (*Homem–Mulher*) apresentam maior dispersão, com concentração entre 0,4 e 0,6 (média: 40,2%), indicando que são comuns, mas raramente dominantes.

Um teste de Wilcoxon confirmou que a diferença entre as proporções *Homem–Homem* e *Mulher–Mulher* é estatisticamente significativa ($p < 0,00001$). Esses resultados sugerem a presença de homofilia de gênero, especialmente entre personagens masculinos, e indicam que a participação feminina ocorre majoritariamente em interações com personagens masculinos, sem formar núcleos estruturais equivalentes.

Adicionalmente, avalia-se a variação temporal desses padrões. A Figura 4 apresenta a distribuição das proporções de interação ao longo do tempo. Visualmente, observa-se uma aparente redução nas interações *Homem–Homem* e um aumento nas interações *Homem–Mulher*. Para verificar se essas tendências são estatisticamente sustentadas, aplicou-se o teste de correlação de Kendall, que avalia a presença de uma tendência monotônica entre duas variáveis — neste caso, o ano de publicação e a proporção de cada tipo de interação. Valores de τ próximos de zero indicam ausência de tendência consistente. Os resultados mostram que nenhuma das três proporções apresenta correlação significativa com o ano de publicação ($p > 0,05$ em todos os casos): *Homem–Homem* ($\tau = 0,023$; $p = 0,445$), *Mulher–Mulher* ($\tau = -0,018$; $p = 0,559$) e *Homem–Mulher* ($\tau = -0,049$; $p = 0,099$). Portanto, embora a visualização sugira variações ao longo do período, os dados não permitem confirmar a existência de uma tendência temporal sistemática nas proporções de interação.

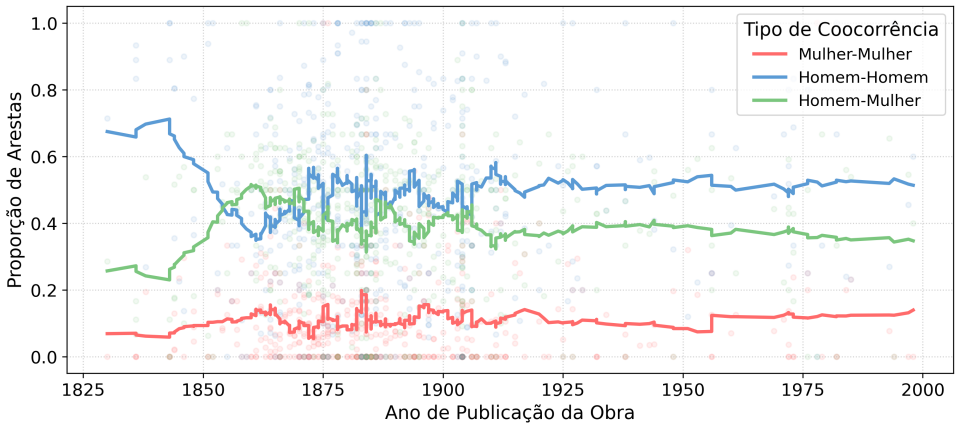

Figura 4. Tendência das proporções de interação por gênero ao longo do tempo.

4.2.2. Importância dos Personagens

A Figura 5 apresenta a distribuição das métricas de centralidade por gênero considerando todos os personagens do corpus. Os testes de Mann-Whitney U não identificaram diferenças estatisticamente significativas entre personagens masculinos e femininos em nenhuma das métricas analisadas — grau ($p = 0,364$), intermediação ($p = 0,259$), proximidade ($p = 0,785$) e autovetor ($p = 0,429$) —, indicando que, quando se considera a rede como um todo, a centralidade estrutural média é similar entre os gêneros.

Contudo, essa similaridade nas distribuições globais não implica ausência de viés: ela mascara uma assimetria que se torna visível quando o foco recai sobre as posições de maior destaque. Ao restringir a análise aos cinco personagens mais centrais de cada obra (top 5 em grau), observa-se que 68,5% desses personagens são masculinos, contra 31,5% femininos (Figura 6). Em outras palavras, personagens masculinos e femininos competem em condições estruturais médias comparáveis ao longo da rede, mas os papéis de maior influência narrativa são ocupados desproporcionalmente por homens. Esse padrão sugere

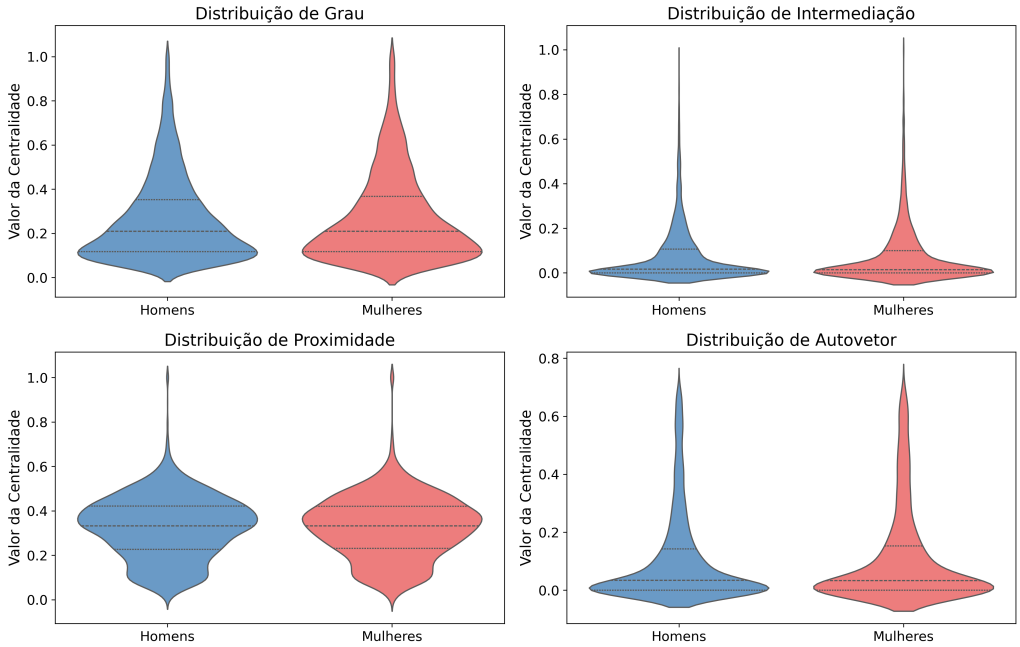

Figura 5. Distribuição das métricas de centralidade (Grau, Intermediação, Proximidade e Autovetor) para personagens masculinos e femininos.

que o viés de gênero na centralidade não é difuso, mas concentrado no topo da hierarquia estrutural — precisamente onde o impacto sobre a narrativa é mais pronunciado.

4.2.3. Agência Narrativa

A análise de agência, realizada com base nas relações de dependência sintática, revela que personagens masculinos apresentam *Agency Score* médio consistentemente superior ao das femininas, com diferença média de 0,094. Entre os personagens de maior agência em cada obra, 68,4% são masculinos. A análise temporal indica uma tendência de convergência: a disparidade, acentuada no início do século XIX, diminui progressivamente, com aumento da agência feminina em meados do século XX.

Em síntese, os resultados do corpus completo evidenciam um viés de gênero estrutural nas redes de personagens, com interações dominadas por homens, concentração masculina nas posições de maior centralidade e maior agência atribuída a personagens do gênero masculino.

4.3. Influência do Gênero do Autor

Esta seção explora o quanto os padrões de viés de gênero identificados na Seção 4.2 são influenciados pelo gênero do autor. Para isso, utiliza-se o corpus balanceado de 58 obras, permitindo uma comparação controlada entre narrativas escritas por autores e autoras.

4.3.1. Proporção de Interações por Gênero do Autor

A comparação entre obras de autores e autoras revela diferenças estruturais consistentes nos padrões de interação (Figuras 7 e 8). Em obras de autores homens, observa-se predominância de interações *Homem–Homem* (36,8%), enquanto em obras de autoras há aumento expressivo de interações *Mulher–Mulher* (32,2%) e redução das interações entre personagens masculinos (16,5%).

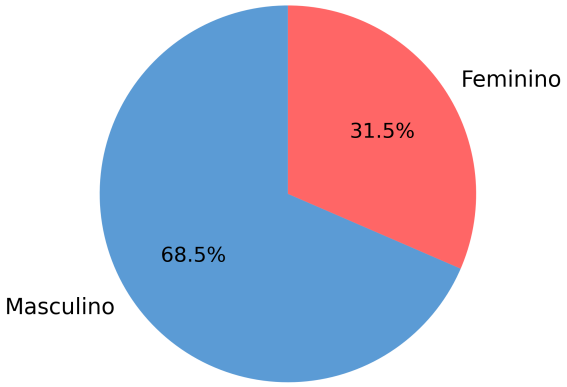

Figura 6. Composição de gênero do grupo de elite (top 5 personagens mais centrais por obra).

Testes de Mann-Whitney U indicam que as diferenças nas proporções de interações *Homem–Homem* ($p = 0,0004$) e *Mulher–Mulher* ($p = 0,0002$) são estatisticamente significativas, enquanto as interações *Homem–Mulher* não apresentam diferença significativa ($p = 0,672$). Esses resultados sugerem que as interações entre gêneros constituem um componente estrutural estável das narrativas, ao passo que os padrões de interação homofílica variam conforme o gênero do autor.

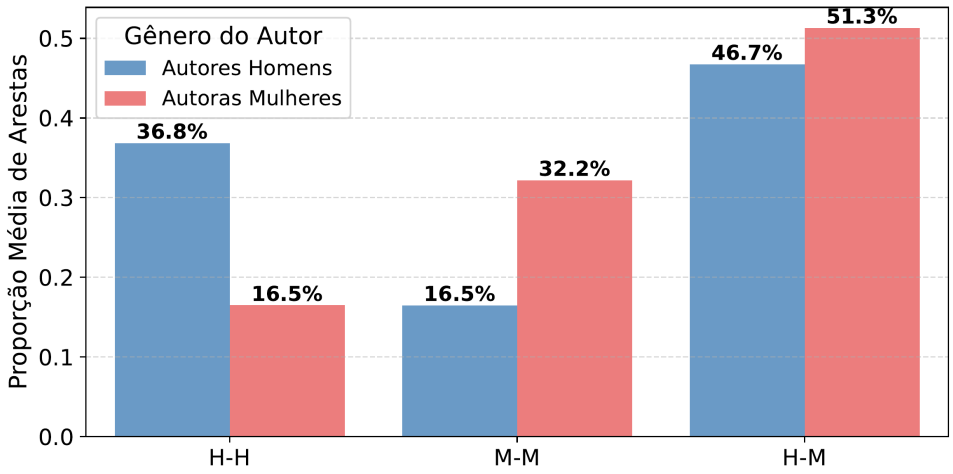

Figura 7. Proporção média de interações por tipo e gênero do autor.

4.3.2. Centralidade por Gênero do Autor

A composição dos personagens mais centrais apresenta variações sistemáticas conforme o gênero do autor (Figura 9). Em obras de autores homens, 56,3% dos cinco personagens mais centrais são masculinos e 38,7% femininos. Em obras de autoras, essa proporção se inverte: 56,3% são femininos e 43,0% masculinos.

Esse resultado contrasta com a análise global apresentada na Seção 4.2, na qual a elite de centralidade é majoritariamente masculina (68,5%). Tal diferença indica que a

Figura 8. Distribuição das proporções de interação por tipo, separadas por gênero do autor.

predominância masculina nas posições de destaque não é uma propriedade intrínseca das narrativas, mas reflete, ao menos em parte, a composição majoritariamente masculina do cânone analisado.

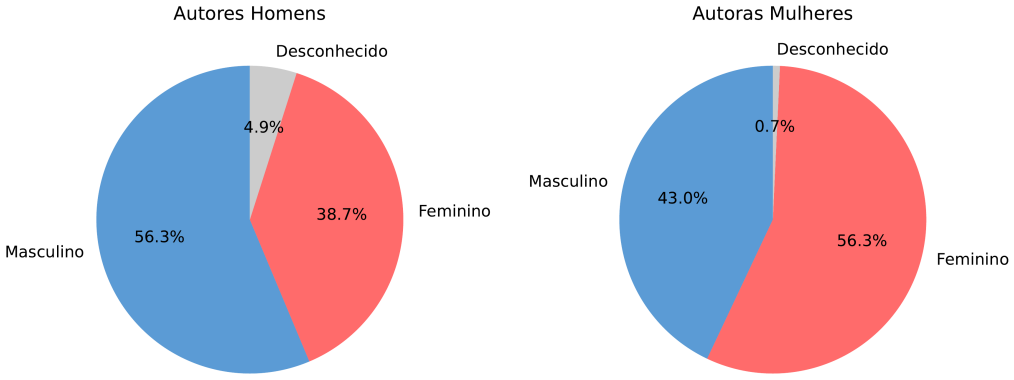

Figura 9. Composição de gênero dos cinco personagens mais centrais de cada obra, por gênero do autor.

4.3.3. Agência por Gênero do Autor

A análise de agência por gênero do autor é apresentada na Figura 10. Em obras de autores homens, a agência média dos personagens masculinos (0,456) é superior à das personagens femininas (0,396), uma diferença de 0,060 pontos. Em obras de autoras, os valores são praticamente simétricos: 0,453 para personagens masculinos e 0,461 para femininos (diferença de 0,008).

No entanto, o teste de Mann-Whitney U não identificou diferença estatisticamente significativa entre os dois grupos de autores ($p = 0{,}175$). Portanto, com base nos dados disponíveis, não é possível afirmar que o gênero do autor afeta a distribuição de agência narrativa. Essa questão permanece em aberto e requer investigação em corpus de maior escala para que se possa avaliar se a diferença descritiva observada reflete um padrão real ou apenas variação amostral.

5. Discussão

Os resultados obtidos permitem discutir o viés de gênero na literatura em língua portuguesa a partir de duas perspectivas complementares: a estrutura das narrativas em larga

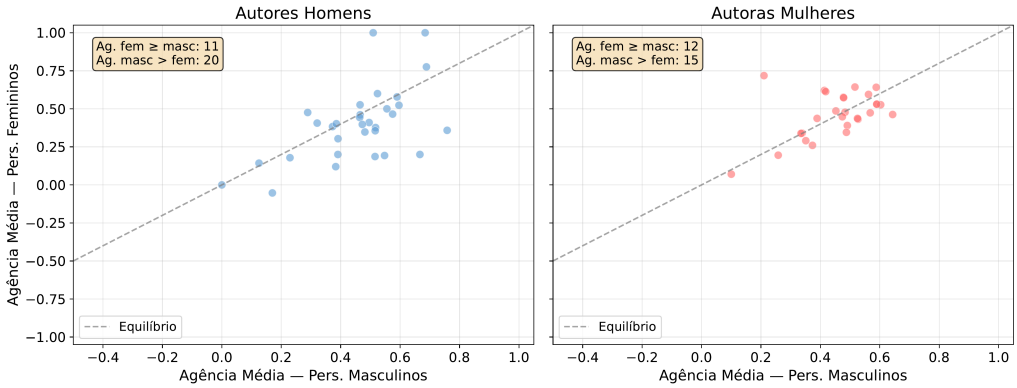

Figura 10. Agência média de personagens masculinos vs. femininos por obra, separada por gênero do autor. Pontos abaixo da linha indicam maior agência masculina.

escala e o papel do gênero autoral na configuração dessas estruturas.

A análise do corpus completo confirma que padrões de desigualdade de gênero documentados na literatura anglófona [Kraicer and Piper 2019] e francesa [Vianne et al. 2023] se reproduzem na literatura em língua portuguesa: a predominância de interações *Homem–Homem*, a concentração masculina nas posições de maior centralidade e a maior agência atribuída a personagens masculinos são consistentes com os achados de Kreuzhage [Kreuzhage 2024] em um corpus de 87 mil obras. Essa convergência entre contextos culturais e linguísticos distintos é sugestiva de padrões estruturais mais amplos, embora uma afirmação mais forte de universalidade requeira análise comparativa sistemática entre tradições literárias — o que está além do escopo deste trabalho.

A principal contribuição deste trabalho, contudo, emerge da análise do corpus balanceado. Ao controlar o gênero do autor, evidencia-se que os padrões estruturais observados em larga escala não são uma propriedade intrínseca das narrativas, mas refletem, ao menos em parte, a composição majoritariamente masculina do cânone analisado. Essa interpretação é sustentada por dois achados estatisticamente significativos. Primeiro, autores e autoras constroem redes com dinâmicas de homofilia opostas, cada grupo privilegiando interações entre personagens do seu próprio gênero ($p = 0{,}0004$ para *Homem–Homem*; $p = 0{,}0002$ para *Mulher–Mulher*). Segundo, a elite de centralidade — majoritariamente masculina (68,5%) na análise global — se inverte nas obras de autoras, onde personagens femininas ocupam 56,3% das posições de destaque. No que diz respeito à agência narrativa, observou-se uma diferença descritiva entre os grupos — maior disparidade de agência nas obras de autores homens do que nas de autoras —, porém essa diferença não atingiu significância estatística ($p = 0{,}175$), de modo que a influência do gênero autoral sobre a agência permanece uma questão em aberto.

Esses achados têm implicações metodológicas diretas para estudos computacionais em larga escala sobre literatura. Análises que não controlem o gênero autoral correm o risco de atribuir à estrutura narrativa em si um viés que é, ao menos parcialmente, um artefato da sub-representação de autoras nos corpora analisados. Nesse sentido, a composição de gênero dos corpora literários não é um detalhe metodológico neutro: é uma variável que condiciona as conclusões sobre desigualdade estrutural na ficção e deve ser reportada e controlada em estudos futuros.

6. Conclusão

Este trabalho investigou o viés de representação de gênero na literatura em língua portuguesa por meio de duas análises complementares. A análise do corpus amplo (551 obras) confirmou a existência de um desequilíbrio estrutural consistente com padrões reportados internacionalmente. A análise do corpus balanceado (58 obras) revelou que o gênero do autor exerce influência significativa sobre os padrões de interação e de centralidade: autores homens constroem redes centradas em personagens masculinos, enquanto autoras produzem narrativas mais equilibradas ou com maior protagonismo feminino. No que se refere à agência narrativa, a influência do gênero autoral não foi confirmada estatisticamente e permanece como questão em aberto.

A principal contribuição deste trabalho é evidenciar que o viés de gênero na literatura não deve ser interpretado apenas como uma propriedade intrínseca das narrativas, mas também como um reflexo de quem as escreve — e, por extensão, de quem é publicado e canonizado. Esse resultado tem implicações diretas para estudos em larga escala: a composição de gênero dos corpora literários é uma variável metodológica que condiciona as conclusões sobre desigualdade estrutural na ficção.

Limitações. A modelagem de interações pode superestimar relações, pois baseia-se na coocorrência de personagens na mesma sentença. Essa limitação pode afetar de forma desigual as comparações de gênero: se personagens femininas são mencionadas com maior frequência em cenas coletivas ou descrições passivas, suas interações podem ser superestimadas em relação às de personagens masculinos, que tendem a aparecer em cenas diádicas com foco na ação. A inferência de gênero, por sua vez, depende de heurísticas baseadas em concordância gramatical e dicionário de nomes, e adota uma classificação estritamente binária, sem revisão manual dos 9,6% de personagens cujo gênero não pôde ser determinado automaticamente, o que pode introduzir viés nas análises comparativas. A medida de agência captura apenas a relação sintática entre personagens e verbos, sem considerar dimensões semânticas como o tipo de ação realizada ou o contexto pragmático da cena. Por fim, o corpus é restrito a obras de domínio público anteriores a meados do século XX, o que impede comparações com a literatura contemporânea, e o tamanho reduzido do corpus balanceado ($n = 58$) limita o poder estatístico de algumas análises.

Trabalhos futuros. Pretende-se expandir o corpus balanceado, o que permitiria revisitar a relação entre gênero autoral e agência narrativa com maior poder estatístico. Planeja-se também incorporar Modelos de Linguagem de Grande Escala (LLMs) para uma inferência de gênero mais robusta. Outra direção envolve a exploração de redes direcionadas baseadas em dependências sintáticas, visando uma estimativa mais precisa da agência narrativa. Por fim, pretende-se investigar abordagens de classificação de gênero que superem o modelo binário adotado neste estudo.

Código e dados. O código-fonte, os scripts de processamento e análise, bem como os dados derivados utilizados neste trabalho, estão disponíveis publicamente no repositório do projeto: `https://github.com/babimartins/literary-networks-portuguese`.

Referências

Aires, V. P., Martins, P. R., and Nakamura, F. (2017). Construção e análise das redes sociais de personagens dos filmes da franquia O Senhor dos Anéis. In *Anais do Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*, pages 599–610. SBC.

Brandão, M., Diniz, M., and Moro, M. (2016). Using topological properties to measure the strength of co-authorship ties. In *Anais do V Brazilian Workshop on Social Network Analysis and Mining*, pages 199–210, Porto Alegre, RS, Brasil. SBC.

Digiampietri, L., Mena-Chalco, J., Silva, G., et al. (2012). Dinâmica das relações de coautoria nos programas de pós-graduação em computação no brasil. In *Anais do I Brazilian Workshop on Social Network Analysis and Mining*, pages 105–116, Porto Alegre, RS, Brasil. SBC.

Kraicer, E. and Piper, A. (2019). Social characters: The hierarchy of gender in contemporary English-language fiction. *Journal of Cultural Analytics*, 3(2).

Kreuzhage, L. (2024). The gender agency gap in fiction writing (1850 to 2010). *Proceedings of the National Academy of Sciences*, 121(29).

Krug, M., Puppe, F., Jannidis, F., Macharowsky, L., Reger, I., and Weimar, L. (2015). Rule-based coreference resolution in German historic novels. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 98–104. Association for Computational Linguistics.

Labatut, V. and Bost, X. (2019). Extraction and analysis of fictional character networks: A survey. *ACM Computing Surveys*, 52(5).

Lopes, G. R., Moro, M. M., Wives, L. K., and de Oliveira, J. P. M. (2010). Collaboration recommendation on academic social networks. In *Advances in Conceptual Modeling – Applications and Challenges*, pages 190–199, Berlin, Heidelberg. Springer Berlin Heidelberg.

Ribeiro, M. A. et al. (2016). A rede social complexa de O Senhor dos Anéis. *Revista Brasileira de Ensino de Física*, 38(1).

Silva, M. d. O. S. (2025). *A computational framework for measuring and analyzing gender bias in Portuguese-language literary texts*. PhD thesis, Universidade Federal de Minas Gerais.

Silva, M. O. and Moro, M. M. (2024). NLP pipeline for gender bias detection in Portuguese literature. In *Proceedings of the Seminário Integrado de Software e Hardware (SEMISH)*. SBC.

Silva, M. O., Oliveira, G. P., and Moro, M. M. (2023). Analyzing character networks in Portuguese-language literary works. In *Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia)*. SBC.

Silva, M. O., Scofield, C., de Melo-Gomes, L., et al. (2022). Cross-collection dataset of public domain Portuguese-language works. *Journal of Information and Data Management*, 13(1).

Silva, M. O., Scofield, C., and Moro, M. M. (2021). PPORTAL: Public Domain Portuguese-language Literature Dataset. In *Anais do III Dataset Showcase Workshop*, pages 77–88, Rio de Janeiro, Brazil. SBC.

Vianne, L., Dupont, Y., and Barré, J. (2023). Gender Bias in French Literature. In *Conference on Computational Humanities Research CHR2023*.