http://dx.doi.org/10.5007/1984-8412.2023.e86749

BELA, RECATADA E DO LAR: O QUE A MINERAÇÃO DE TEXTOS LITERÁRIOS NOS DIZ SOBRE A CARACTERIZAÇÃO DE PERSONAGENS FEMININAS E MASCULINAS

BELLA, RECATADA Y HOGAREÑA: QUÉ NOS DICE LA MINERÍA DE TEXTOS LITERARIOS SOBRE LA CARACTERIZACIÓN DE LOS PERSONAJES FEMENINOS Y MASCULINOS

BEAUTIFUL, MODEST AND HOUSEWIFE: WHAT LITERARY TEXT MINING TELL US ABOUT MALE AND FEMALE CHARACTERIZATION

Cláudia Freitas*

Flávia Martins**

Pontifícia Universidade Católica | Rio de Janeiro

RESUMO: Este artigo apresenta os resultados de uma pesquisa que articula métodos quantitativos e qualitativos sobre representações discursivas que tematizam a questão do gênero. Nosso objetivo é identificar como personagens masculinas e femininas são caracterizadas em textos literários, e, para isso, tomamos como objeto de exploração um corpus de obras da literatura brasileira com cerca de 5 milhões de palavras, anotado semântica e morfossintaticamente. O estudo se dá em duas frentes: observando os predicadores na descrição das personagens e as ações desempenhadas por elas. Como resultado, (i) indicamos como a metodologia utilizada permite diferentes perspectivas sobre os dados, indo além da análise baseada em formas e listas de frequência, e (ii) ratificamos a construção estereotipada dos gêneros masculino e feminino na literatura dos séculos XIX e XX, com o feminino caracterizando-se sobretudo pela aparência, especialmente pela beleza, e pela esfera doméstica.

PALAVRAS-CHAVE: Humanidades Digitais. Linguística de Corpus. Mineração de textos. Estudos de gênero.

RESUMEN: Este artículo presenta resultados de una investigación que articula métodos cuantitativos y cualitativos sobre representaciones discursivas que tematizan la cuestión del género. Nuestro objetivo es identificar cómo se caracterizan los personajes masculinos y femeninos en los textos literarios, y para ello tomamos como objeto de exploración un corpus de obras de la literatura brasileña con cerca de 5 millones de palabras, anotado semántica y morfosintácticamente. El estudio se desarrolla en dos frentes: observar los predicativos en la descripción de los personajes y las acciones realizadas por ellos. Como resultado, (i) señalamos cómo la metodología utilizada permite diferentes perspectivas sobre los datos, yendo más allá del análisis basado en listas de frecuencia, y (ii) ratificamos la construcción estereotipada de los géneros masculino y femenino en la literatura de los siglos XIX y XX, donde lo femenino se caracteriza principalmente por la apariencia, especialmente la belleza, y por el ámbito doméstico.

PALABRAS-CLAVE: Humanidades Digitales. Lingüística de Corpus. Minería de textos. Estudios de género.

* No momento da submissão do artigo, Cláudia Freitas era Professora e pesquisadora do Programa de Pós-Graduação em Estudos da Linguagem (PPGEL) da Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio). Atualmente é pesquisadora do C4IA-ICMC-USP; E-mail: claudiafreitas@usp.br .

** Mestre em Estudos da Linguagem pela PUC-Rio. E-mail: flaviamrps@gmail.com.

ABSTRACT: This paper presents the results of a research that combines quantitative and qualitative methods on discursive representations that focus on gender. Our goal is to identify how male and female characters are characterized in literary texts, and for that we explore a corpus of Brazilian literature with 5 million words, annotated with semantic and morphosyntactic information. The study is conducted along two directions: observing the predicates present in characters' description and the actions performed by them. As a result, (i) we indicate how the methodology used allows different perspectives on the data, going beyond the analysis based on forms and frequency lists, and (ii) we ratify the stereotyped construction of male and female in the literature of the 19th and 20th centuries, with the female being characterized mainly by appearance, especially beauty, and by the domestic sphere.
KEYWORDS: Digital Humanities. Corpus Linguistics.. Text Mining. Gender Studies.

1 INTRODUÇÃO

A exploração do conteúdo de grandes acervos textuais, cuja leitura convencional na íntegra seria inviável em um tempo de vida, tem sido uma das tarefas da Linguística Computacional/Processamento de Linguagem Natural (PLN), sendo alvo de interesse das subáreas Mineração de Textos e Extração de Informação. Nestas, o que está em jogo é a detecção de padrões textuais capazes de indicar a presença de algum tipo de informação – implícita ou explícita – a partir do que está materializado no texto. Assim, ambas almejam a explicitação, por meio de métodos e ferramentas computacionais, de informação nova por meio da extração automática (de informação) em diferentes fontes textuais. Essa exploração dos dados textuais, por sua vez, pode ser feita tanto com o objetivo de levantar pontos até então desconhecidos quanto para procurar tendências e padrões gerais nos dados (HEARST, 2003).

Até recentemente, no entanto, esta forma de trabalhar automaticamente com textos não costumava ser aplicada a conjuntos de textos de interesse das Humanidades – exceto no caso dos próprios estudos linguísticos, como nos mostra Santos (2019). Com a popularização de métodos estatísticos aplicados a diversas áreas, inclusive humanas, a grande área das Humanidades Digitais (HDs) tem ganhado espaço, trazendo com ela um sintagma cada vez mais frequente nos trabalhos que lidam com textos: leitura distante (tradução do inglês "*distant reading*"). O termo, cunhado por Franco Moretti no âmbito dos estudos literários, pode ser entendido como "[...] uma forma de trabalhar em que a distância não é um obstáculo, mas uma forma específica de conhecimento" (MORETTI, 2008, p. 7). Se de longe não é possível ter acesso a detalhes, por outro lado trata-se de um ângulo que possibilita a observação e descoberta de relações e padrões nos textos, de uma maneira que a leitura convencional não seria capaz de capturar, assim como na Mineração de Textos.

O crescimento e popularização das HDs, com incorporação de diferentes métodos e aplicações para além dos estudos literários, traz também práticas que complexificam a própria ideia de contagem de palavras: textos passam a ser entendidos como um espaço de relações que deverão ser modeladas pelos especialistas, conectando variáveis linguísticas a variáveis sociais. Em consequência, uma das mudanças vivenciadas nas HDs é a maior facilidade na combinação de dados quantitativos e qualitativos (UNDERWOOD, 2016).

É neste espaço que trazemos um modo de trabalhar que combina aspectos da mineração de textos, dos estudos linguísticos baseados em corpus e de uma leitura distante e crítica dos dados. Para tanto, delineamos como objetivo identificar de que maneira personagens humanas se constituem no discurso, e como se dá a diferenciação entre os gêneros feminino e masculino, avançando com o estudo piloto apresentado em Freitas *et al.* (2022). Utilizamos um *corpus* composto por obras de literatura brasileira em domínio público e direcionamos nossa atenção para dois pontos principais: (i) os predicadores utilizados para caracterizar personagens e (ii) as ações que estas executam. Conectando variáveis linguísticas (predicações e ações verbais) com variáveis sociais (gênero), buscamos elaborar uma visão geral de como homens e mulheres são construídos através da linguagem, e acreditamos que seja possível identificar nos resultados obtidos alguns vieses de gênero presentes na estrutura social brasileira.

2 ALGUMAS PALAVRAS SOBRE GÊNERO E REPRESENTATIVIDADE NAS ARTES

Para nós, o gênero é performativo; uma forma de se apresentar no mundo, um traço identitário. Assim, nosso entendimento sobre gênero alinha-se com Butler (2019), que define gênero como os significados culturais assumidos pelo corpo sexuado. Homens e

mulheres, cis ou trans, apresentam-se ao mundo de determinada forma que os faça serem reconhecidos como pertencentes a um ou outro gênero. Neste trabalho, assumimos uma perspectiva binária de gênero apenas por uma escolha metodológica, visto que os textos que compõem nosso corpus partem desse pressuposto, e entendemos que "homem" e "mulher" são categorias não naturais, criadas por nós.

Em artigo de 1973, Laura Mulvey investigou o prazer visual do cinema sobre a mulher. Uma vez que a maioria dos diretores era composta por homens, a mulher na tela apareceria para-ser-olhada. Mulvey percebe não só a agentividade do homem nos filmes, quando são estes que fazem a história avançar, deflagrando os acontecimentos, mas também a objetificação e glamourização do corpo feminino, indicando que as características "glamourosas" de um astro masculino e feminino são diferentes.

Mais recentemente, uma forma de medir a representatividade feminina em obras de ficção é por meio do teste de Bechdel, que verifica se a obra tem ao menos duas personagens femininas que falem entre si sobre um assunto que não seja um homem. Segundo pesquisa do Geena Davis *Institute on Gender in Media*[1] (SMITH *et al*., 2014), realizada com base em roteiros de 120 filmes produzidos em 11 países lançados entre 2010 e 2013, as mulheres continuam sendo estereotipadas e sexualizadas nas telas, sendo representadas em papeis sociais de pouca importância. A maioria das personagens nomeadas nas telas é masculina (69% dos quase 6 mil personagens), e a presença de personagens femininas nos filmes aumenta de acordo não apenas com quem os escreve ou realiza, mas também de acordo com o gênero dos filmes. A pesquisa conclui que as personagens femininas nas telas têm 5 vezes mais chances de receberem comentários baseados na aparência do que as masculinas.

No que se refere à literatura, nosso objeto de análise, é esperado que, enquanto produto cultural, também (re)produza o comportamento social e reflita ideologia de quem a escreve, como indica Cameron (2012). É desse lugar que partimos com o objetivo de explorar a caracterização de personagens masculinas e femininas da literatura brasileira dos séculos XIX e XX.

3 CONSIDERAÇÕES TEÓRICO-METODOLÓGICAS

Um corpus, nos estudos linguísticos, é um conjunto de enunciados ou textos, em formato eletrônico, produzido naturalmente e compilado com alguma intenção. Segundo Santos (2008, p. 45), um corpus é "[...] uma colecção classificada de objectos linguísticos para uso em Processamento de Linguagem Natural/Linguística Computacional/Linguística". O termo *uso*, nesse caso, pode se referir a estudos, avaliações, testes, por exemplo, e *objetos linguísticos* podem ser textos, frases, palavras, erros ortográficos, entradas de dicionário, citações, pareceres jurídicos, traduções, dentre outros. Já *classificada* indica que o material não é uma compilação feita de maneira aleatória, mas que contém algum tipo de organização (ou classificação) em algum aspecto, como a escolha do material – uma compilação exaustiva de algo (um autor, todas as edições de um ano de um determinado jornal), ou apenas uma amostra – ou o nível dos fenômenos que se deseja observar – um determinado gênero de texto, um tipo de tradução, por exemplo. Ainda conforme Santos (2008, p. 46), um corpus "[...] não é o objeto de estudo do que em inglês se chama *corpus linguistics*, mas sim a ferramenta, o utensílio com que se faz linguística". Estamos, portanto, tratando de um estudo linguístico com base em um conjunto de textos em formato eletrônico e que conta com o auxílio de ferramentas computacionais para sua exploração. Este estudo é, ainda, compatível com diferentes perspectivas teóricas da linguagem – veja-se o amplo levantamento de Mcenery e Hardie (2012) e Baker (2010) para aproximações com estudos sociolinguísticos, e Sampson (2001) para uma "linguística empírica", por exemplo.

Com o auxílio de corpora linguísticos é possível tanto identificar comportamentos linguísticos mais comuns como olhar para as variações e casos raros, muitas vezes não tão acessíveis através da introspecção. Cada corpus oferece inúmeras possibilidades de exploração e análise, cabendo a quem pesquisa avaliar quais serão usadas tendo em vista as questões de interesse. Por outro lado, e de maneira complementar, se os corpora contêm enunciados produzidos naturalmente, eles também têm potencial de nos dizer algo sobre os valores da sociedade em que surgiram (e por isso um bom corpus, além de *adequado* às necessidades de pesquisa, é um corpus bem *documentado*). Especificamente, as repetições e padrões que sobressaem na imensa quantidade de dados podem explicitar traços *discursivos* que confirmem ou refutem nossas intuições (BAKER, 2010).

[1] Organização que trabalha colaborativamente com a indústria do entretenimento para ajudar a conseguir equilíbrio nas representações de gênero e a reduzir os estereótipos prejudiciais na caracterização das personagens.

Na aproximação entre *corpus* e análise do discurso, seguimos os passos de Baker *et al*. (2008); Mautner (2009); Baker (2010) e Cameron e Panović (2014), dentre outros, para quem o acréscimo de uma dimensão quantitativa à análise de discurso pode oferecer uma outra – complementar – perspectiva sobre os dados. Baker *et al*., (2008), por exemplo, combinaram explorações em grandes corpora e Análise Crítica do Discurso (ACD) a fim de identificar como artigos jornalísticos ingleses construíram a imagem dos imigrantes e refugiados num corpus de 140 milhões de palavras. Os pesquisadores fizeram, primeiramente, suas análises de forma independente, valendo-se dos usuais métodos das duas distintas áreas – Baker, Gabrielatos e McEnery trabalharam com os métodos de corpus, Khosravinik, Krzyzanowski e Wodak da ACD – e, em seguida, avaliaram que benefícios cada abordagem poderia ter a partir da combinação dos métodos. Iniciaram suas análises por listas de frequências e padrões lexicais estatisticamente relevantes complementados pela observação das respectivas linhas de concordância. Interessaram-se pelas preferências semânticas e pelas coocorrências, que seriam capazes de indicar a posição ideológica do autor dos textos em relação a migrantes e solicitantes de refúgio.

Apesar de a valorização do texto autêntico ser compartilhada tanto pelos adeptos do trabalho com corpora quanto pela ACD, essa convergência não indica que todos que utilizam corpora em seus estudos compartilhem da mesma visão do que seja um texto ou da intenção de analisar o discurso nele presente. Como aponta Baker (2010), uma análise tradicional baseada em corpus não consegue apontar as razões pelas quais certos padrões são encontrados uma vez que não leva em consideração o contexto social, político, histórico e cultural dos dados.

Uma das críticas comumente feitas aos analistas do discurso é que suas amostras são, muitas vezes, cuidadosamente escolhidas para comprovar um ponto de acordo com uma agenda política ou ideológica (MAUTNER, 2009; BAKER *et al*., 2008). Nesse sentido, a utilização de métodos estatísticos sobre um corpus reduziria as possibilidades de que as análises ou amostras "[...] seja[m] tendenciosa[s], no sentido de [inconscientemente] o analista produzir frases que dariam jeito para uma determinada teoria" (SANTOS, 2008, p. 48), ou apenas para se comprovar algo em que se acredita. Reforçamos, entretanto, que não nos colocamos na posição ingênua que assume que números e tecnologias são neutras. Como já afirmamos em trabalhos anteriores, não nos alinhamos com um posicionamento segundo o qual os dados emergem do corpus, fruto de tecnologias assépticas e prontos para serem analisados por um/a observador/a neutro/a ou bem treinado/a. Pelo contrário, defendemos que se trata de objetos construídos segundo a perspectiva de quem pesquisa. Do mesmo modo, quando nos referimos a abordagens qualitativas e quantitativas estamos lidando com *escolhas metodológicas*, e não *epistemológicas*. A associação usualmente feita nas ciências humanas entre opções quantitativas e paradigmas epistemológicos objetivistas, por um lado, e métodos qualitativos e práticas interpretativas situadas que localizam *quem* pesquisa como parte da produção do conhecimento, por outro, é arbitrária.

Reconhecemos que na articulação entre dados quantitativos e discurso está um tipo de trabalho conhecido como "análise de conteúdo", definido por Bardin (1977, p.42) como uma técnica para obtenção de "[...] indicadores para inferências em relação ao contexto de produção e recepção de textos", e que faz uso de "procedimentos sistemáticos, objetivos e descritivos", tais como classificação e análise da frequência de presença ou da ausência de certos itens, para ordenar um conjunto de dados textuais. No entanto, explicitamos que temos formas distintas de operar com tais procedimentos, como também já detalhamos em Freitas *et al.*, 2022.

Como mencionamos, tem sido cada vez mais frequente, no âmbito das Humanidades, que abordagens baseadas em dados sejam quantitativas e qualitativas, como ilustram, por exemplo, os trabalhos de Soni *et al*.(2021) e Underwood *et al*. (2018). O primeiro aborda o papel de jornais abolicionistas estadunidenses publicados no século XIX na formação de opinião pública relativa à abolição – especificamente, os autores investigam o papel de vanguarda das mulheres no movimento abolicionista, bem como o papel da imprensa negra, considerando um acervo composto pelas edições de 11 jornais (a maioria dos títulos com periodicidade semanal) publicados entre 1827 e 1865. A partir desse acervo, os autores investigam os deslocamentos de sentido relativos às palavras *liberdade* (*freedom*) e *justiça* (*justice*), baseados na técnica de vetores de palavras (*word embeddings*) e levando em conta informações de raça e gênero associadas aos editores de cada jornal.

Underwood *et al.* (2018), assim como o presente trabalho, tematizam questões de gênero em uma coleção de obras literárias (de língua inglesa). O recorte temporal da pesquisa é amplo, com obras que vão do final do século XVIII ao início do século XXI. Ao

investigar descrições de personagens, os autores concluem que as diferenças entre personagens masculinas e femininas ficaram menos acentuadas nos últimos 170 anos: se, em meados do século XIX, a diferença na linguagem usada para descrever homens e mulheres fictícios é bem marcada, há um enfraquecimento à medida que o tempo avança. Por outro lado, no que se refere à autoria das obras, a pesquisa encontra o movimento oposto: a proporção de obras ficcionais escritas por mulheres cai pela metade (de cerca de 50% dos títulos para cerca de 25%) no período que vai de 1850 a 1950, e, com isso, diminui-se também o número de personagens femininas.

Apesar das aproximações, o trabalho de Underwood *et al.* (2018) difere do nosso de um modo substancial: o tamanho do acervo é muito superior ao nosso, o que também impacta na metodologia utilizada – os autores analisam centenas de milhares de obras, e nós trabalhamos "apenas" com centenas de obras. Underwood *et al.* (2018) usam uma estratégia automática para verificar se a diferenciação entre gênero se sustenta ao longo do tempo. Inicialmente, é criado um modelo para prever, a partir da linguagem usada na descrição das personagens, se se trata de personagem masculina ou feminina. Para a criação do modelo, foram selecionadas, aleatoriamente, 800 personagens femininas e 800 masculinas[2]. Para caracterizar cada grupo e criar o modelo de diferenciação automática, foram selecionadas as 2200 palavras mais frequentemente associadas às personagens, como as ações que realiza, ações das quais é objeto, adjetivos que a modificam ou substantivos que governa[3]. Com isso, espera-se que o modelo "aprenda" o que significa ser "masculino" ou "feminino" apenas observando correlações implícitas expressas nas palavras associadas às personagens. O passo seguinte é usar os padrões "aprendidos" pelo modelo para fazer previsões sobre personagens que ainda não foram vistos. Se o modelo tiver um bom desempenho na diferenciação, é possível concluir que as caracterizações foram organizadas por uma concepção binária de gênero. Ou seja, quanto mais acertadamente o modelo consegue distinguir entre personagens masculinas e femininas, maior seria a diferenciação presente entre os gêneros. Os resultados mostraram que, levando em conta o recorte temporal, as diferenças entre personagens masculinos e femininos ficam cada vez mais difíceis de discernir desde meados do século XIX até o início do século XXI.

No presente trabalho, utilizando um acervo menor e fazendo uso de padrões léxico-sintáticos, investigamos qualitativamente essa diferenciação, partindo de uma caracterização em quatro eixos: aparência, caráter, emoção e papel social. De maneira complementar, realizamos ainda uma avaliação da atribuição automática de gênero aos nomes próprios de personagem para a língua portuguesa.

4 METODOLOGIA

Para identificar os termos utilizados na caracterização de seres humanos no corpus, desenvolvemos uma metodologia em cinco etapas, listadas a seguir:

1. Busca por padrões: nesta etapa, inspiradas por métodos da extração de informação e mineração de textos, buscamos estruturas linguísticas indicadoras de predicação e de ação, referentes a personagens do discurso. Este é um processo interativo, de tentativa e erro, até a obtenção dos padrões mais precisos. As buscas foram feitas por meio do AC/DC (ver seção 4.1).
2. Geração e análise de listas de distribuição derivadas dos resultados dos padrões, tendo em vista a distribuição por gênero.
3. Organização das ocorrências (apenas da predicação) em classes. Nesta etapa, buscamos atribuir sentidos mais gerais às ocorrências (formas) obtidas em (2).
4. Validação das classes criadas em (3).
5. Análise dos resultados com base nas frequências, formas (palavras) e classes.

Em Freitas *et al.* (2022) apresentamos explorações iniciais relacionadas à predicação, propondo uma classificação inicial em quatro eixos, que retomaremos aqui. Neste trabalho, (i) aperfeiçoamos as expressões de busca, e com isso trazemos de forma adicional novas explorações e análises sobre os dados, (ii) avaliamos e corrigimos a atribuição automática de gênero dos nomes próprios e (iii) realizamos explorações e análises associadas às ações exercidas pelas personagens.

[2] A identificação de uma personagem como masculina ou feminina é feita por meio do gênero gramatical atribuído ao nome próprio associado a personagem.

[3] Foram excluídas palavras indicativas de papéis de gênero como menina/menino, esposa/marido e pronomes pessoais.

4.1 O CORPUS

Levando em conta os objetivos de nossas explorações, utilizamos o *corpus* OBras (SANTOS; FREITAS; BICK, 2018), que compreende um acervo de obras da literatura brasileira em domínio público. Trata-se de um corpus dinâmico – novas obras estão sempre sendo adicionadas – e a versão com que trabalhamos contém 248 obras. O OBras integra o acervo da Literateca (SANTOS, 2019), e é parte do projeto AC/DC (SANTOS; BICK, 2000), um serviço de busca em corpora criado e mantido pela Linguateca, que disponibiliza corpora anotados sintática e semanticamente. A anotação sintática é fornecida pelo *parser* PALAVRAS (BICK, 2014) e a anotação semântica é feita pelo PALAVRAS e por colaboradores da Linguateca. Todo material é público e está acessível para buscas complexas por meio interface AC/DC e para download na página do projeto[4].

A assimetria entre os gêneros masculino e feminino é visível desde a composição do corpus, pois há apenas três obras de autoria feminina, de duas autoras: Maria Firmina dos Reis e Júlia Lopes de Almeida[5]. Quanto a isto, é importante registrar a dificuldade de acesso a obras em domínio público em formato de texto digital acessível. A Biblioteca Nacional disponibiliza obras em formato fac-símile com acessibilidade textual propiciada pela tecnologia OCR (*Optical Character Recognition*), mas pouquíssimas de autoria feminina. O contraste com o material utilizado por Underwood *et al.* (2018) é evidente: o trabalho utiliza um acervo de milhares de obras e equilibrado quanto ao gênero dos autores, o que permite investigar o papel do gênero também no que se refere à autoria, o que não pudemos fazer.

4.2 PADRÕES DE BUSCA

Para identificar personagens humanas buscamos por nomes próprios associados à etiqueta semântica de pessoa, pronomes pessoais e uma lista de substantivos indicativos de pessoa (*mulher, rapariga, moça, menina, senhora, irmã, mãe, prima* etc.)[6]. Para identificar as caracterizações associadas às personagens, buscamos por estruturas sintáticas indicativas de predicação: predicativos de sujeito, apostos, adjuntos adnominais. Após algumas experimentações, e em um processo iterativo, chegamos ao conjunto final de 27 padrões de busca para cada gênero, que foram usados de forma concatenada em uma única busca.

O Quadro 1 traz uma das regras, que deve ser lida como: "uma pessoa do gênero feminino e com a função sintática de *sujeito*, seguida pelos verbos *ser* ou *estar*, seguida opcionalmente por um *advérbio*, seguida por um elemento com função de *predicativo do sujeito* que seja *adjetivo, nome* ou *verbo*, desde que não esteja na voz passiva"[7].

```
[sema=".*Pessoa.*" & gen="F" & func="SUBJ>"] [lema="ser|estar"] [pos="ADV.*"]* @[temcagr!=".*PASS.*"
& pos="ADJ|N|V" & func="<SC"]
```

Quadro 1: Exemplo de regra usada na busca por predicações

Fonte: Silva (2020)

Para a caracterização de personagens por meio de suas ações, chegamos a quatro regras para cada gênero. O Quadro 2 traz uma das regras, que deve ser lida como: "uma pessoa do gênero feminino e com a função sintática de *sujeito*, seguida opcionalmente por um advérbio que não seja *não* ou *nunca*, seguida por um verbo que não seja *ser*, *estar* ou *haver*, e que não esteja na *voz passiva* ou em uma *forma participial*."

[4] http://www.linguateca.pt/OBRAS/OBRAS.html

[5] Na versão atual, 10.22, conta com 279 obras, e as mesmas duas autoras, mas são cinco obras femininas, e não três.

[6] Temos listas diferentes para personagens masculinas, com palavras como *homem; rapaz, pai* etc.

[7] Alguns elementos dos padrões podem parecer redundantes, mas têm a função de minimizar erros da análise sintática automática.

```
[sema=".*Pessoa.*" & gen="F" & func="SUBJ>"] [pos="ADV.*" & lema!="não|nunca"]* @[pos="V" &
lema!="ser|estar|haver" & temcagr!="PASS|PCP"]
```

Quadro 2: Exemplo de regra usada na busca por ações

Fonte: Silva (2020)

Ambas as expressões dependem de uma correta atribuição de gênero gramatical aos nomes próprios[8]. Para a atribuição de gênero aos nomes próprios, o já referido trabalho de Underwood *et al.* (2018) utiliza a ferramenta BookNLP (BAMMAN *et al.*, 2014), que oferece um índice de precisão e abrangência de 94.7% e 83.1%, respectivamente, para homens, e 91.3% e 85.7% para mulheres. Para a língua portuguesa, verificamos, em explorações preliminares, que a atribuição automática de gênero aos nomes próprios humanos poderia impactar nos resultados da caracterização. A fim de compreender o papel desta etapa, realizamos um estudo com o objetivo de medir o grau de dificuldade na atribuição automática de gênero e, consequentemente, o quanto podemos confiar na análise automática quando não é possível contar com alguma revisão humana.

4.3 AVALIAÇÃO E REVISÃO DA ATRIBUIÇÃO DE GÊNERO AOS NOMES PRÓPRIOS

Para a avaliação da atribuição automática de gênero, utilizamos uma porção do OBras um pouco menor que aquela utilizada para as caracterizações: 226 obras diferentes, totalizando quase 5.5 milhões de *tokens*. Todo o material disponível no AC/DC – tanto pela própria interface, como para *download* – já passou por revisões, dentre elas a revisão de gênero gramatical que relatamos aqui. Para a avaliação da anotação automática de gênero, tivemos acesso a uma versão do OBras logo após a análise do PALAVRAS, e, portanto, anterior à revisão. Para a presente avaliação, fizemos uma segunda rodada de revisão dos nomes próprios, tendo em vista a detecção de erros não corrigidos anteriormente. Nesta etapa, mais de 20 mil ocorrências de nomes próprios de pessoa tiveram o gênero gramatical modificado, o que corresponde a cerca de 25% do material e, indica, portanto, um acerto na atribuição automática de gênero em torno de 75%[9]. Todas as revisões foram incorporadas ao material, e os resultados a seguir foram obtidos já com os dados revistos.

5 RESULTADOS: PREDICAÇÃO

Foram encontradas 5262 predicações no total, 2937 (56%) atribuídas a personagens masculinas e 2325 (44%) a personagens femininas[10]. Quanto à variedade lexical, encontramos, para o masculino, 819 lemas diferentes e, para o feminino, 712. No entanto, apesar da diferença de mais de mil predicadores para personagens masculinas, vemos na Tabela 1 que o predicador mais usado para descrição de personagens masculinas, *sério*, aparece 74 vezes, praticamente a metade do número de ocorrências do predicador feminino mais frequente, *bonito* (146 vezes). Ou seja, de início vemos que os dados reforçam o que outros estudos já haviam indicado: a caracterização de mulheres a partir da beleza (além de *bonita*, temos *bela* e *formosa* também entre os predicadores mais frequentes). Por outro lado, as predicações masculinas mais frequentes referem-se ao caráter ou personalidade – *sério*, *bom*.

Femininos	**Qtd**	**Masculinos**	**Qtd**
bonito	146	sério	74
belo	68	bom	60
casado	58	alto	57
amado	46	pobre	46

[8] Essa informação está codificada indiretamente nas expressões de busca: a etiqueta *Pessoa* presente no atributo *sema* é atribuída apenas a nomes próprios.

[9] Notamos que boa parte dos erros de gênero se deve a nomes próprios estrangeiros, como *Metternich* ou *Estherhazy*.

[10] Constatamos que menos de 10% das predicações encontradas com os padrões continha algum tipo de erro, que poderia ser decorrência de erro do processamento linguístico automático (erro de anotação) ou de o padrão léxico-sintático ter levado a uma estrutura não-predicativa, ou que não predique pessoas, como *cheio*, em "Pois vocês não estão vendo que **ela está cheia** de açucenas?".

formoso	44	rico	41
honesto	42	capaz	40
velho	40	bonito	34
pálido	36	pálido	32
solteiro	34	forte	29
bom	31	cândido	29

Tabela 1: Frequência dos lemas predicadores mais comuns por gênero

Fonte: Silva (2020)

No entanto, uma imensa parcela dos predicadores tem uma frequência baixíssima. Independentemente do gênero, quase 85% das predicações têm três ocorrências ou menos, e 59,44% dos predicadores aparece apenas uma vez: se fôssemos analisar apenas as predicações mais frequentes, a grande maioria dos predicadores ficaria de fora. Reconhecemos que é importante olhar para ocorrências mais frequentes, e que a repetição pode estar associada à presença de estereótipos e de discursos dominantes, que tendem a apresentar um traço de repetição na linguagem (BAKER, 2010), mas interessa-nos igualmente olhar para a longa lista de predicadores que aparece menos vezes, pois, de fato, são a maioria.

Mas não é possível observar padrões, e, portanto, fazer uma *leitura distante*, se o que temos são ocorrências singulares. A fim de atribuir sentido a essas ocorrências distribuídas, criando então novos padrões de visualização desses mesmos dados, classificamos as predicações em quatro categorias não excludentes. As categorias refletem nosso interesse na caracterização humana, e foram identificadas a partir da leitura e análise dos dados. São elas:

> Aparência – traços externos associado sobretudo ao corpo, como alto, velho, nua.
> Caráter – traços psicológicos como honesto, burro.
> Emoção – traços emocionais como triste, corajosa.
> Papel social – traços "sociais", associados à ocupação ou papel na sociedade, como padre, escravo, mãe, vagabundo, doutor, brasileiro.

A análise inicial dos dados mostrou ainda que os padrões também podem indicar estados temporários, expressos em formas participiais como "sentado" ou "sacudido". Porque não têm a ver com caracterizações, essas formas, que correspondem a cerca de 5% dos dados, não foram levadas em conta neste estudo.

5.1 VALIDAÇÃO DAS ANÁLISES: UM ESTUDO DA CONCORDÂNCIA ENTRE ANOTADORES[

Inicialmente, os dados foram manualmente classificados por uma das autoras do artigo[11]. Para validar a classificação, nos inspiramos na *concordância interanotadores,* prática da Linguística Computacional que visa avaliar a confiança que se pode ter nas anotações (e toda anotação é uma classificação) de um *corpus* (ARSTEIN, 2017). A confiança depende da reprodutibilidade, isto é, da possibilidade de, analisando um mesmo conjunto de dados e seguindo as mesmas instruções ou critérios, diferentes pessoas chegarem às mesmas análises. Realizamos um estudo utilizando a ferramenta Rêve[12], criada justamente para avaliar convergências e divergências de análises linguísticas baseadas em corpus (SANTOS *et al*., 2015).

O estudo contou com 17 anotadores, que receberam como instruções da tarefa uma breve explicação e exemplos de cada categoria, além das alternativas "Outro" e "Não sei". Também informamos que seria possível selecionar mais de uma opção, em consonância com a possibilidade de classificação múltipla[13]. Selecionamos 40 frases do *corpus*, incluindo exemplos que consideramos mais difíceis de classificar.

[11] Em diversas situações foi necessária a consulta às linhas de concordância para confirmar os sentidos com que cada predicador ocorre.

[12] O Rêve é uma ferramenta da *Gramateca* para elaboração e realização de testes online utilizando os mesmos corpora anotados disponibilizados no AC/DC.

[13] Na frase "Os homens são **encantadores**, o homem é insuportável." (*Turbilhão*, Coelho Neto, 1904) *encantador* pode ser igualmente classificado como *aparência* ou *caráter*.

Não houve frase em que apenas uma única classe tenha sido selecionada pelos anotadores e, com frequência, foi selecionada mais de uma alternativa para cada frase. Esta situação não surpreende quando lembramos que estamos no terreno das classificações/anotações semânticas – por exemplo veja-se, para a língua portuguesa, a tentativa de classificação semântica dos verbos de elocução feita em Costa e Freitas (2017). Os resultados reforçam, ainda, uma das vantagens de trabalhar com material anotado (e documentado), em oposição à leitura exclusiva de linhas de concordância: com a anotação temos análises explícitas, públicas, que podem ser discutidas e refutadas; já a leitura e análise das linhas de concordância lida com interpretações implícitas, nem sempre cientes de que podem não ser intepretações únicas ou inquestionáveis (McENERY; HARDIE, 2012).

Para a avaliação dos resultados da concordância, sempre que um participante selecionasse mais de uma resposta, acrescentamos um ponto para cada classe. Consideramos concordância total os casos em que nossa classificação coincidiu com a escolha da maioria dos participantes e concordância parcial os casos em que uma das respostas da maioria coincidiu com a nossa[14].

Na maioria das frases/predicações houve uma categoria que se destacou, com aproximadamente o dobro das respostas obtidas nas outras categorias. Nos 40 predicadores analisados, tivemos 80% de concordância total e 17% de concordância parcial (97% de concordância), um índice adequado no contexto de concordância interanotadores – que não é exatamente o que fizemos, mas serve como um parâmetro relativo ao que esperar dos resultados. O único caso de divergência completa refere-se à palavra *trigueira*, e temos consciência de que parte da classificação fornecida pelos participantes estava errada, possivelmente devido ao pouco uso do adjetivo no português contemporâneo falado no Brasil. Esta etapa permitiu ainda aperfeiçoar nossa própria percepção dos sentidos tomados pelos predicadores no contexto em que ocorrem. Assim, para além de validar a classificação dos dados, esta fase serviu para melhorarmos nossa própria classificação nos seguintes casos: *ardente, caro, virgem, indispensável, ocioso* e *cabeçudo* (uma descrição detalhada desta etapa, que inclui uma análise dos casos divergentes, encontra-se em Silva (2021).

5.2 PREDICAÇÕES HUMANAS VISTAS DE LONGE E DE PERTO

Após a classificação, a distribuição dos dados assume as formas indicadas na Figura 1. Se consideramos, por outro lado, apenas as 20 predicações mais frequentes para cada gênero, temos as formas da Figura 2.

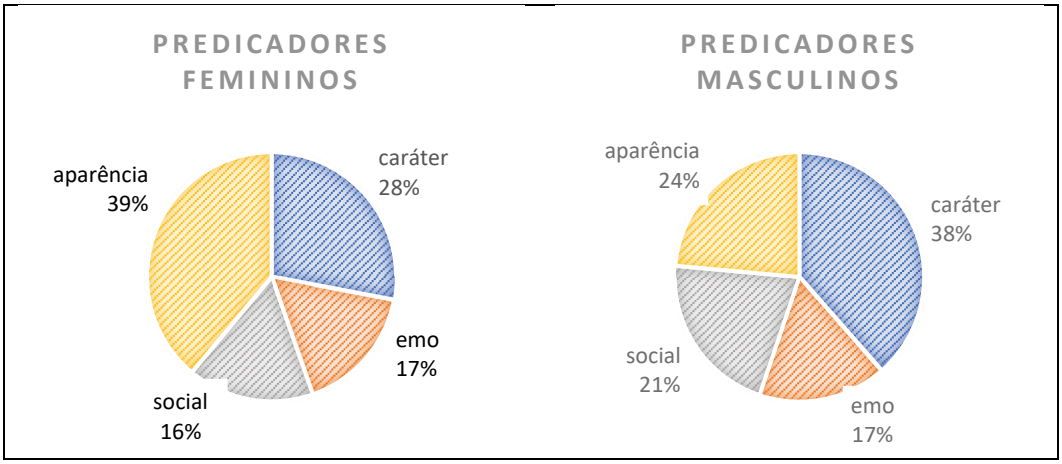

Figura 1: Distribuição dos predicadores por gênero e por eixo, considerando todos os dados do corpus Obras

Fonte: Silva (2020)

[14] Todas as frases do estudo foram analisadas, de maneira convergente, pelas autoras do artigo, consistindo em uma espécie de *gabarito* da classificação.

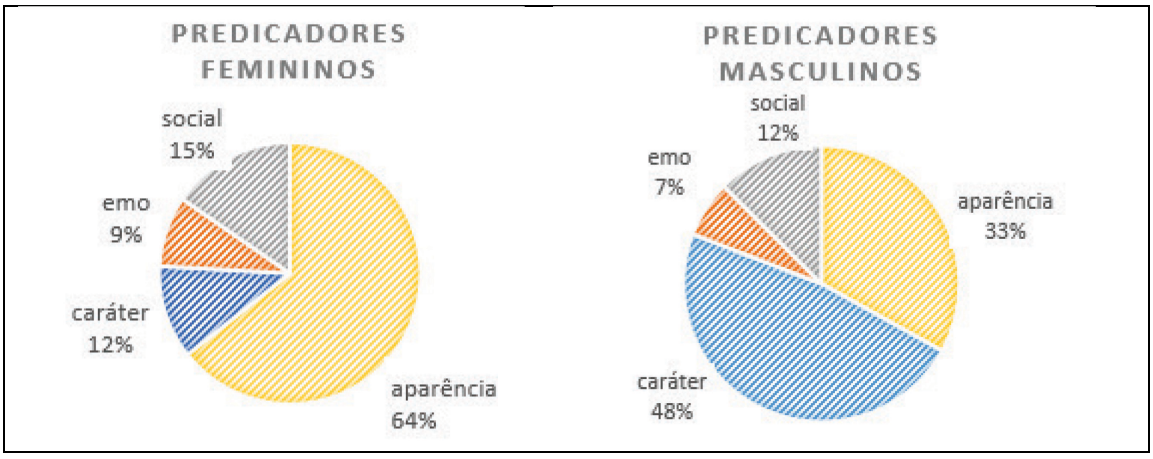

Figura 2: Distribuição dos predicadores por eixo e gênero, considerando apenas os 20 mais frequentes no corpus OBras

Fonte: Silva (2020)

Como podemos notar comparando as Figuras 1 e 2, o contraste entre predicações femininas e masculinas é mais visível quando levamos em conta apenas as predicações mais frequentes, e a diferença é especialmente nítida para a classe aparência: quando levamos em conta apenas os lemas mais frequentes, a aparência feminina responde por 64% das caracterizações, contra 33% das caracterizações masculinas. Quando consideramos todos os dados (figura 1), a aparência continua se destacando como a classe mais tipicamente feminina, mas não chega a 40%, e a diferença entre predicações masculinas e femininas, que antes era de quase 30%, cai para 15%. Do mesmo modo, considerando apenas as predicações mais frequentes (figura 2), traços de caráter são pouco utilizados na representação feminina (é a segunda predicação menos comum), mas passa à segunda mais frequente quando consideramos todo o material (figura 1), reforçando a importância de se considerar a totalidade dos dados, e não apenas os mais frequentes.

Voltando à Figura 1 (e daqui em diante só levaremos em conta todas as predicações, e não apenas as mais frequentes), podemos observar tendências e em seguida escolher lugares para analisar de perto, articulando leitura distante e aproximada. Tomando distância, vemos que cerca de 40% das predicações femininas são relacionadas à aparência, e cerca de 40% das masculinas, ao caráter. Predicações associadas ao papel social são mais frequentemente associadas aos homens (21%) que às mulheres (16%) e emoções têm uma distribuição parecida para ambos os gêneros (16% e 17%). Aproximando o olhar, como veremos a seguir, temos predicações qualitativamente distintas.

Diminuindo a distância para uma análise qualitativa, temos, conforme as Tabelas 2 e 3, informações relativas a predicadores usados exclusivamente para cada um dos gêneros, levando em conta as quatro classes de análise. Vemos, então, que somente personagens masculinas são retratadas como *valentes, honrados, maduros* e *públicos*[15], e apenas personagens femininas são *ciumentas, gentis, formosas e íntimas*[16].

Tal como *bonita* se destaca no total de predicadores, ocorrendo 146 vezes no *corpus*, entre os exclusivos para um gênero também vemos *formosa* como o ponto fora da curva entre todos os predicadores femininos exclusivos, seguido em seu eixo por *linda* e *encantadora*, com aproximadamente metade das ocorrências de cada um. A descrição da aparência é tão predominante para as personagens femininas que, dentre todos os predicadores exclusivos femininos, os três mais frequentes são desta classe.

[15] Isto se deve à expressão *homem público*, que, segundo o Dicionário Houaiss online (acesso em abril de 2022), indica "indivíduo que ocupa um alto cargo do Estado", e que não tem um equivalente para pessoas do gênero feminino – "mulher pública".

[16] Na anotação feita pelo analisador PALAVRAS, adjetivos têm lema na forma do masculino singular e substantivos no singular, visto que substantivos não flexionam em gênero. Adjetivos, como *bonito* e *bonita*, têm como lema a versão masculina singular, e substantivos, como *empregado* e *empregada*, têm lemas distintos para a forma masculina e a feminina (no singular), por isso vemos ambos os nomes na lista de predicações únicas para cada gênero.

emoção masculina		caráter masculino		aparência masculina		social masculino		estado masculino	
predicador	**qtd**	**predicador**	**qtd**	**predicador**	**qtd**	**predicador**	**qtd**	**predicador**	**qtd**
valente	15	honrado	24	maduro	8	público	20	ferido	6
entusiasmado	6	ilustre	10	barbado	5	filho	15	arruinado	3
furioso	6	sábio	9	calvo	4	político	11	esquecido	2
corajoso	5	rude	7	criança	4	notável	10	deitado	3
deslumbrado	4	malcriado	7	baixinho	3	importante	10	desinteressado	2
misericordioso	4	seguro	6	teso	3	capitão	9	diferente	2
feroz	4	amável	6	rijo	3	português	8	incomodado	2
apressado	3	sagaz	5	imberbe	3	empregado	8	inerte	2
vexado	3	morigerado	5	rapaz	3	primitivo	7	estremunhado	2
apreensivo	3	útil	5	vesgo	2	pai	7	atorrear	1
pacato	3	indiferente	5	quadragenário	2	feudal	6	internado	1
hediondo	3	extraordinário	5	galhardo	2	formado	6	endefluxado	1
impassível	3	franco	5	esguio	2	negociante	6	entretidíssimo	1
insolente	3	singular	5	mascarado	2	célebre	6	mortinho	1
bravo	3	medíocre	4	brioso	2	oficial	5	contorsionado[sic]	1
humilde	3	probo	4	espadaúdo	2	poderoso	5	lavadinho	1
maravilhado	3	tenaz	4	ágil	2	rústico	4	detido	1
respeitoso	3	moderno	4	tisnado	2	sócio	4	equipado	1
folgazão	3	polido	4	miúdo	2	trabalhador	4	agitante	1
audacioso	2	decente	4	bonitinho	2	poeta	4	rojado	1

Tabela 2: Lista das vinte predicações exclusivas (lemas) mais frequentes para personagens masculinas

Fonte: Silva (2020)

emoção feminina		caráter feminino		aparência feminina		social feminino		estado feminino	
predicador	**qtd**	**predicador**	**qtd**	**predicador**	**qtd**	**predicador**	**qtd**	**predicador**	**qtd**
ciumento	6	gentil	8	formoso	38	filha	15	desmaiado	4
faceiro	6	má	6	lindo	18	mãe	14	enganado	4

indiferente	5	namoradeiro	5	encantador	17	íntimo	4	embebido	3
adorável	5	loureira	5	grávido	6	primo	3	confuso	3
choroso	4	trabalhadeira	4	moça	6	empregada	3	ciente	2
desgraçado	4	senhora	4	maltrapilho	5	solteira	3	suspeito	2
ofendido	3	atento	4	esplêndido	5	rainha	3	cônscio	2
espavorido	3	insensível	4	galante	5	donzela	3	estendido	2
amuado	2	ingênuo	4	ardente	5	comum	3	transido	2
agradecido	2	angélico	4	tísico	4	supremo	3	inexperiente	2
afortunado	2	romanesco	3	cansado	4	núbil	2	desfalecido	2
culpado	2	ímpio	3	sereno	4	chic	2	próximo	2
carinho	2	frívolo	3	jovem	4	cozinheira	2	ultrapassado	1
amoroso	2	trêfego	3	másculo	3	parenta	2	recém-aparecer	1
envergonhado	2	bondoso	3	são	3	inglês	2	retido	1
fúria	1	divino	3	sujo	3	desquitado	2	entretido	1
desventuroso	1	morfético	2	preto	3	noiva	2	desprevenido	1
agastado	1	prendado	2	alvo	3	espanhol	2	salteado	1
chorão	1	rebelde	2	quarentão	2	desamparado	2	debruçado	1
ciumentíssimo	1	bem-educado	2	setuagenário	2	escrava	2	predisposto	1

Tabela 3: Lista das vinte predicações exclusivas (lemas) mais frequentes para personagens femininas
Fonte: Silva (2020)

A classe das *emoções* contém muitos predicadores comuns aos dois gêneros, por isso são poucos aqueles surgindo entre as predicações exclusivas. Ainda assim destacamos *valente* (que pode ser *emoção* ou *caráter)*, que não aparece entre as predicações femininas. Quando nos detemos no *caráter* masculino, vemos quão frequentemente homens são *honrados, ilustres* e *sábios* – e surpreendentemente percebemos que são predicações exclusivas masculinas – enquanto mulheres são *gentis*. Neste ponto, voltamos a lembrar que as expressões de busca no corpus dão conta de muitos casos, ao menos entre os contextos sintáticos mais comuns na nossa língua, mas não de todos. Com isso, a exclusividade de alguns predicadores pode ser fruto também de erros da análise sintática automática, e é possível que as mulheres *honradas, valentes* e *sábias* estejam mencionadas no *corpus*; mas não resta dúvida, após esta análise, de que apareçam menos do que os homens com estas qualidades.

Além da *exclusividade*, analisamos os predicadores também em termos de *preferência* quanto à associação a um determinado gênero. O que chamamos de "preferência" é um fator comparativo de proporção entre a quantidade de vezes em que cada predicador ocorre para cada gênero, em relação ao total de ocorrências desse predicador. Tomemos como exemplo o predicador *belo*, que ocorre 71 vezes no *corpus*, sendo 66 vezes para caracterização de personagens femininas – o que corresponde a 93% do total de ocorrências desse predicador. Isto faz com que *belo* seja *preferencialmente* utilizado para predicações femininas. Nas Tabelas 4 e 5 vemos os trinta predicadores comuns aos dois gêneros, em ordem de preferência por gênero.

Os dados mostram que o predicador mais feminino é *belo*, coincidentemente, um sinônimo do predicador mais frequente dentre todos os encontrados para personagens femininas, *bonito*, e assim mais um elemento da classe *aparência*. Também identificamos que elas são mais caracterizadas como *medrosas* (80%), *infelizes* (78%) e *assustadas* (89%) e *idosas* (88%) do que eles. Já eles costumam ser mais *sérios* (90%), *absolutos*[17] (90%), *resolutos* (88%), *robustos* (87%) e *teimosos* (88%). Do mesmo modo, caracterizações ligadas à esfera doméstica são mais comuns entre as predicações femininas, e entre as masculinas observam-se mais frequentemente caracterizações referentes a relações sociais fora da esfera familiar.

Preferência feminina					**Preferência masculina**			
predicador	**eixo**	**fem**	**%**		**predicador**	**eixo**	**masc**	**%**
belo	aparência	66	93%		mau	caráter	29	94%
assustado	emo	8	89%		grave	caráter	21	91%
idoso	aparência	22	88%		cândido	caráter	29	91%
viúvo	social	21	88%		sério	caráter	74	90%
amado	emo	43	81%		absoluto	caráter	9	90%
bonito	aparência	146	81%		civilizado	social	7	88%
agradável	emo	4	80%		sertanejo	social	7	88%
medroso	emo	4	80%		teimoso	caráter	7	88%
agitado	emo	4	80%		robusto	aparência	20	87%
mudo	aparência	4	80%		contente	emo	12	86%
travesso	caráter	11	79%		ilustrado	social	6	86%
amante	social	7	78%		admirável	caráter	6	86%
infeliz	emo	7	78%		Ativo	social	6	86%
feio	aparência	17	77%		respeitável	caráter	11	85%
contentíssimo	emo	3	75%		calmo	caráter	10	83%
corado	emo	3	75%		abastado	social	5	83%
guerreiro	social	3	75%		útil	caráter	5	83%

Tabelas 4 e 5: Lista dos vinte predicadores mais frequentes comuns aos dois gêneros, listados por preferência e por eixo

Fonte: Silva (2020)

Associando as preferências e exclusividades por gênero a cada uma das classes, vemos que as predicações preferencialmente femininas são do tipo *emoção* e *aparência*, e as predicações preferencialmente masculinas dizem respeito ao *caráter* e ao *papel social*, em um quadro diferente daquele que apresenta as predicações por classe sem levar em conta as preferências (Figura 1).

[17] Presente nas expressões "senhor absoluto" e "senhora absoluta".

Diminuindo ainda mais a distância, vemos que certos predicadores, como aqueles associados à beleza no que se refere ao eixo *aparência*, e associados a medo (*medrosa, assustada*) tristeza (*infeliz, chorosa, chorona*) e ciúme, no eixo *emoção*, aparecem como tipicamente femininos. Por outro lado, para personagens masculinas, o quadro é inverso: em oposição ao medo, a temos a coragem (*valente, corajoso, audacioso, bravo, feroz*); em oposição à tristeza, temos a alegria/satisfação (*contente* é a única emoção preferencial masculina dentre as mais frequentes). Ainda como tipicamente masculinas temos a retidão de caráter (*grave, sério, respeitável, honrado*), que se opõe à frivolidade feminina (*travessa* é única predicação da classe caráter que é tipicamente feminina, e *frívola* e *namoradeira* são exclusivamente femininas).

6 CARACTERIZAÇÃO PELAS AÇÕES: RESULTADOS

Utilizando os padrões de busca para ações, encontramos mais de 26 mil ocorrências, com 1.355 verbos distintos. Em uma situação diferente daquela das predicações, a maioria dos verbos associados a personagens tem seu sentido especificado pelos seus complementos. Verbos são classes de palavras reconhecidamente polissêmicas e, por esse motivo, a etapa de distribuição em classes não foi realizada. Se *chorar* ou *sorrir*, verbos intransitivos, permitem um entendimento sem a necessidade de contextos mais amplos, o mesmo não pode ser dito de verbos como *ganhar* ou *exercer*, que podem assumir diferentes feições conforme o que se ganha e o que se exerce. A Tabela 6 apresenta as informações quantitativas relativas aos verbos[18], e a Tabela 7 traz os verbos mais frequentes por gênero.

	Total de lemas distintos	**Total de ocorrências**
Verbos associados exclusivamente a	355	499
Verbos associados exclusivamente a	355	913
Verbos associados a ambos	645	24.876
Total de verbos associados	1.355	26.288

Tabela 6: Distribuição dos verbos que têm como sujeito personagens no corpus Obras
Fonte: Silva (2020)

Femininos				**Masculinos**			
verbo	qtd	verbo	qtd	verbo	qtd	verbo	qtd
ter	537	voltar	139	ter	825	olhar	211
ficar	279	sentir	138	fazer	487	voltar	200
fazer	272	saber	111	ficar	378	vir	197
querer	197	ouvir	109	querer	342	sentir	194
dar	178	sair	107	dar	292	chegar	192
dizer	167	ver	101	entrar	276	saber	184
entrar	160	falar	99	sair	272	falar	184
olhar	156	vir	98	dizer	243	passar	164

[18] Lembramos que já foram excluídos dos padrões de busca, e da contagem, verbos que a gramática tradicional chama de verbos de ligação, como *ser*, *estar*, *parecer* etc.

sorrir	153	estremecer	92	ver	241	sorrir	160
ir	146	passar	90	ir	226	tomar	158

Tabela 7: Lista dos 20 verbos mais frequentes associados a personagens, por gênero
Fonte: Silva (2020)

Como podemos perceber, e de forma esperada, a distribuição dos verbos traz, dentre aqueles mais frequentes, verbos-suporte ou auxiliares, independentes do gênero. Tais verbos carregam pouco ou nenhum sentido, sendo não sendo possível analisá-los sem levar em conta seus complementos (como dar em *dar um grito*; *dar fé*; *dar a perceber*).

As Tabelas 6 e 7 mostram que, embora a maioria dos verbos esteja associada aos dois gêneros, há muitos que são exclusivos de um determinado gênero. As Tabelas 8 e 9 detalham essa informação, apresentando os verbos exclusivos e os verbos preferenciais para cada gênero, respectivamente.

A análise dos verbos exclusivos mostra que somente eles "montam" e apenas elas "cavalgam". O Quadro 3 traz exemplos de concordâncias para cada um dos casos.

Exclusivos femininos				**Exclusivos masculinos**			
verbo	qtd	verbo	qtd	verbo	qtd	verbo	qtd
cogitar	5	saudar	3	montar	12	plantar	7
verter	5	elevar	3	fumar	11	determinar	7
alimentar	5	repassar	3	obter	11	rejeitar	7
ostentar	4	dignar	3	marcar	10	demonstrar	6
ressonar	4	obtemperar	3	depositar	9	regressar	6
enrolar	4	cavalgar	3	coçar	9	salvar	6
enrubescer	4	rodar	3	travar	9	empurrar	6
agonizar	4	renunciar	3	discutir	9	atalhar	6
amamentar	4	tapar	3	desistir	8	respeitar	6
talhar	4	dissimular	3	construir	8	aconselhar	6
definhar	4	enfeitar	3	roubar	8	velar	6
derrear	3	arfar	3	governar	8	depor	6
instar	3	reprovar	3	impor	7	galgar	6
ansiar	3	proteger	3	combater	7	bastar	6

Tabela 8: Distribuição dos verbos exclusivos por gênero
Fonte: Silva (2020)

O ato de montar a cavalo e de cavalgar podem soar como sinônimos e a escolha lexical pode ser mera coincidência. No entanto, *montar* e *cavalgar* podem ter uma conotação erótica, que poderia justificar a escolha lexical; o primeiro, associado à virilidade e poder – na pecuária, o termo *monta* designa o ato reprodutivo natural do gado. O uso de *montar*, no lugar de *cavalgar*, atua como

um eufemismo para "diluir", de certa forma, o ato de sentar-se sobre algo com as pernas abertas, inconveniente a um estereótipo de masculinidade; reservariam portanto, o *cavalgar* à fêmea [19].

Outras ações exclusivas dos homens (no corpus) são *determinar, demonstrar, governar, impor, combater, construir*: ações que colocam os homens em posição de demonstração de poder; *roubar* e *fumar* também são ações apenas presentes entre eles.

Das ações exclusivas das mulheres, vemos *menear, verter, derrear, enrubescer, definhar, obtemperar*, verbos que demonstram posição de submissão, timidez ou fraqueza; *amamentar* e *alimentar*, associados a atribuições familiares*; ostentar, enfeitar* e *dissimular*, associados a vaidade e futilidade. Esses verbos, escolhidos pelos autores para caracterizar cada personagem, são um indício de que a literatura retrata a desigualdade existente nos papeis de gênero em nossa sociedade, e/ou ainda atua na reprodução de estereótipos.

Cavalgar:
id="Recordações_do_escrivão_Isaías_Caminha Prosa:romance LB 1909 realismo masc ": Em torno da mesa, uma mulher **cavalgava** uma espécie de tapir ou de anta.
id="Helena Prosa:romance MdA 1876 romantismo masc ": Quando chegou à porta da cavalariça, viu aparelhados dois animais, o cavalo de seus passeios da manhã, e a égua que a tia **cavalgava** uma ou outra vez.
Montar:
id="A_Escrava_Isaura Prosa:romance BG 1875 romantismo masc ": Depois do almoço Leôncio **montou** a cavalo, percorreu as roças e cafezais, coisa que bem raras vezes fazia, e ao descambar do Sol voltou para casa, jantou com o maior sossego e apetite, e depois foi para o salão, onde, repoltreando-se em macio e fresco sofá, pôs-se a fumar tranqüilamente o seu havana.
id="O_seminarista Prosa:romance BG 1872 romantismo_regionalismo masc ": À tardinha desse mesmo dia, o rapaz **montou** a cavalo, e tomou o caminho da vila, mas lá não chegou.
id="O_gaúcho Prosa:romance JdA 1870 romantismo_regionalismo masc ": Pela primeira vez **montou ele** o soberbo ginete, e deu algumas voltas pelo campo.

Quadro 3: Exemplos de ocorrências dos verbos *cavalgar* e *montar*

Fonte: Silva (2020)

Outras ações exclusivas dos homens (no corpus) são *determinar, demonstrar, governar, impor, combater, construir*: ações que colocam os homens em posição de demonstração de poder; *roubar* e *fumar* também são ações apenas presentes entre eles.

Das ações exclusivas das mulheres, vemos *menear, verter, derrear, enrubescer, definhar, obtemperar*, verbos que demonstram posição de submissão, timidez ou fraqueza; *amamentar* e *alimentar*, associados a atribuições familiares*; ostentar, enfeitar* e *dissimular*, associados a vaidade e futilidade. Estes verbos, escolhidos pelos autores para caracterizar cada personagem, são um indício de que a literatura retrata a desigualdade existente nos papeis de gênero em nossa sociedade, e / ou ainda atua na reprodução de estereótipos.

Por fim, dos verbos comuns a ambos os gêneros (os mais frequentes estão na Tabela 9), há 15.324 ocorrências masculinas e 9.552 femininas, que correspondem a uma proporção de 62% e 38%. Levando em conta o padrão de busca utilizado, podemos ler os números da seguinte maneira: em 62% dos casos, personagens masculinas são sujeitos sintáticos, isto é, são agentes. No que se refere às personagens femininas, a agentividade cai para 38%. A fim de comparar os verbos comuns aos dois gêneros, retomamos a análise segundo as preferências (Tabela 9).

Do lado masculino, o verbo *ganhar* ocupa a primeira posição, com 91% das ocorrências totais. A partir da leitura das linhas de concordância, percebemos que os homens ganham *reputação, partidas, rios de dinheiro* e *batalhas*, enquanto as mulheres ganham *na sorte*, ou *em estar ao lado de alguém*, ou ganham *o processo de divórcio*, mais uma vez relegando-as ao ambiente familiar. Já vimos que somente os homens *fumam* (Tabela 8) e, por esse motivo, *acender* é um verbo mais comum entre eles do que entre elas. Elas

[19] Ainda que o verbo *montar* seja mais polissêmico que cavalgar, notamos que todas as ocorrências encontradas de *montar* no padrão buscado (verbo *montar* que tem como sujeito seres humanos) têm *cavalo* como objeto.

acendem *velas*, o *gás* ou *um fogo em suas veias*, enquanto eles também acendem *candelabros, tochas, cigarros* e *charutos*, como nos mostram os exemplos do Quadro 4.

Preferência feminina					
verbos	**masc**	**%**	**fem**	**%**	**total**
exercer	3	19%	13	81%	16
soluçar	5	22%	18	78%	23
coser	2	25%	6	75%	8
atender	2	25%	6	75%	8
suprir	1	25%	3	75%	4
corar	14	26%	40	74%	54
baixar	10	26%	28	74%	38
cobrir	3	27%	8	73%	11
derramar	3	27%	8	73%	11
enviuvar	2	29%	5	71%	7
despedir	2	29%	5	71%	7
perdoar	2	29%	5	71%	7
cessar	2	29%	5	71%	7
relancear	3	30%	7	70%	10
saborear	3	30%	7	70%	10
esconder	7	30%	16	70%	23
recear	5	31%	11	69%	16
chorar	35	32%	74	68%	109
tocar	17	33%	34	67%	51
desatar	8	33%	16	67%	24
adoecer	7	33%	14	67%	21
trajar	5	33%	10	67%	15
rolar	4	33%	8	67%	12
lavar	3	33%	6	67%	9
inspirar	3	33%	6	67%	9
invejar	2	33%	4	67%	6
mover	2	33%	4	67%	6
cantarolar	2	33%	4	67%	6
cheirar	2	33%	4	67%	6
revelar	2	33%	4	67%	6
aquiescer	1	33%	2	67%	3

Preferência masculina					
verbos	**masc**	**%**	**fem**	**%**	**total**
ganhar	30	91%	3	9%	33
espreitar	9	90%	1	10%	10
estabelecer	9	90%	1	10%	10
considerar	16	89%	2	11%	18
suspeitar	8	89%	1	11%	9
piscar	8	89%	1	11%	9
acender	38	88%	5	12%	43
expor	15	88%	2	12%	17
agradecer	49	88%	7	13%	56
suar	7	88%	1	13%	8
visitar	7	88%	1	13%	8
penetrar	19	86%	3	14%	22
comprar	25	86%	4	14%	29
conseguir	62	86%	10	14%	72
embarcar	18	86%	3	14%	21
jogar	18	86%	3	14%	21
vender	12	86%	2	14%	14
cortejar	6	86%	1	14%	7
berrar	6	86%	1	14%	7
tossir	6	86%	1	14%	7
entreabrir	6	86%	1	14%	7
aproveitar	41	85%	7	15%	48
atribuir	17	85%	3	15%	20
cortar	17	85%	3	15%	20
resmungar	17	85%	3	15%	20
suportar	11	85%	2	15%	13
pular	16	84%	3	16%	19
acenar	10	83%	2	17%	12
campear	5	83%	1	17%	6
atentar	5	83%	1	17%	6
caçar	5	83%	1	17%	6

Tabela 9: Lista dos verbos comuns, distribuídos por gênero e preferência (apenas os 30 verbos mais frequentes)

Fonte: Silva (2020)

Quadro 4: Exemplos de ocorrências do verbo *acender*

Masculinos:
id="Contos_fora_da_moda Prosa:conto ArtA 1894 masc ": Ele **acendeu** uma lamparina e apagou o gás.
id="Água_de_Juventa Prosa:contos CN 1905 realismo masc ": Eduardo **acendeu** um cigarro e, enquanto a mulher se revia ao espelho, recompondo os cabelos esvoaçantes, pôs-se a mirar as unhas, limpando-as, polindo-as com um pequenino estilete de prata
Femininos:
id="O_Mulato Prosa:romance AA 1881 naturalismo masc ": Ana Rosa **acendera** uma vela a São Manuel do Buraco e Maria Bárbara prometera uma bochecha de cera a Santa Rita dos Milagres.
id="Romanceiro Prosa:contos CN 1898 realismo masc ": Quando o médico a declarou perdida as senhoras **acenderam** velas no oratório e rezaram para que Deus mandasse lágrimas à mísera como se fazem preces, nos campos, para que venham chuvas.
id="Turbilhão Prosa:romance CN 1904 realismo masc ": Ritinha **acendeu** uma vela e colocou-a à mesa de cabeceira, ao lado de um pequeno crucifixo.
id="O_monstro_e_outros_contos Prosa:conto HC 1932 masc ": Desde o momento em que o filho partiu, **acendera** ela uma lamparina de azeite em frente ao oratório tosco, forrado de azul, onde a Senhora das Dores chorava, o coração transpassado por uma espada.
id="Ubirajara Prosa:romance JdA 1874 indianismo_romantismo masc ": Ela **acende** em suas veias um fogo mais generoso que o do cauim, e prepara para seu corpo o repouso da cabana.

Fonte: Silva (2020)

Além da exclusividade masculina para o ato de fumar, e, portanto, de acender cigarros e afins, chama a atenção um único caso em que o "acender" feminino não se refere a uma vela ou lampião, mas ao "generoso fogo" que "ela acende em suas veias", mais uma situação envolta pela sensualidade feminina. Não vemos essa conotação em nenhum dos 38 casos em que "acender" surge como ação de personagens masculinas. Além disso, as velas acesas por elas não são tanto para iluminar, mas estão em um contexto de religiosidade, que não aparece no ambiente masculino. Outro verbo interessante, porque surge como o de maior preferência para personagens femininas, é "exercer". O Quadro 5 traz exemplos do que eles e elas *exercem*.

Masculinos:
id="Contos_fora_da_moda Prosa:conto ArtA 1894 masc ": Maurício **exercia** na Alfândega um modesto emprego de escriturário, e, como residisse nas proximidades do Passeio Público, e era por natureza comodista e ordenado, tomava sistematicamente, às nove horas, o bondinho que contornava parte do morro do castelo, e ia despejar-lo no Carceler, perto da repartição.
id="A_falência Prosa:romance JLdA 1901 naturalismo_realismo fem ": Era a filha mais velha e a mais instruída: pilhara os Tempos das vacas gordas, quando o pai **exercia** um cargo lucrativo
Femininos:
id="O_Cortiço Prosa:romance AA 1890 naturalismo masc ": Que estranho poder era esse, que a mulher **exercia** sobre eles, a tal ponto, que os infelizes, carregados de desonra e de ludibrio, ainda vinham covardes e suplicantes mendigar-lhe o perdão pelo mal que ela lhes fizera? ...
id="Água_de_Juventa Prosa:contos CN 1905 realismo masc ": Ela fez-me sinal para que me não movesse e eu sentia-me dominado -- era uma fascinação que aquela mulher **exercia** sobre mim ou era o assombro de tão insólita cena que me tolhia e avassalava.
id="Antes_que_cases Prosa:conto MdA 1875 masc ": A influência que a mulher **exercia** nele não podia ser mais decisiva
id="Comentários_da_semana Prosa:crônica MdA 1861 masc ": Na apoteose dos talentos, bem como no conforto dos que padecem, a mulher **exerce** sempre a sua alta missão; tanto galardoa como consola.
id="Helena Prosa:romance MdA 1876 romantismo masc ": Estácio conhecia já o domínio que a moça **exercia** sobre si mesma; a tranqüilidade não o convenceu.

Quadro 5: Exemplos de ocorrências do verbo *exercer*

Fonte: Silva (2020)

Aqui temos um cenário inverso do que vimos até então. Eles exercem cargos (profissionais) ou caridade. E elas, o que exercem é alguma forma de poder, domínio, influência, comando, pressão. Há apenas um caso em que exercer se refere a uma função – a de pianista.

7 CONSIDERAÇÕES FINAIS

Neste trabalho, apresentamos os resultados de uma pesquisa de viés quantitativo e qualitativo, tomando por base um grande corpus composto por obras da literatura brasileira em domínio público, majoritariamente dos séculos XIX e XX. De maneira complementar, também buscamos desenvolver uma metodologia que incentivasse a colaboração entre estudos de discurso e análise com grandes corpora, fazendo um duplo uso da anotação linguística: por um lado, é a anotação do próprio corpus (anotação morfossintática e semântica) que permite um tipo de busca mais profundo, que vai além dos padrões de ocorrência; por outro, utilizamos uma estratégia associada ao processo de anotação – a concordância interanotadores – para validar um processo inicial de classificação dos dados, sobre o qual se desenvolveu boa parte da análise apresentada.

O trabalho com corpus não precisa olhar apenas para o que é quantitativamente relevante; dados raros só serão descartados se o/a pesquisador/a decidir ignorar ocorrências singulares. De maneira complementar, um tratamento quantitativo-qualitativo oferece diferentes ângulos de análise: os números têm algo a nos mostrar, e o que é ou não relevante depende do que se busca.

Assim, a distribuição das palavras em quatro eixos de análise permitiu ir além da pura frequência das formas, e associada à análise baseada em exclusividades e preferências levou a resultados que, se por um lado não são exatamente uma novidade, por outro trazem uma materialidade inegável para a discussão. A análise dos dados brutos relativos às palavras mais frequentes das predicações ratifica a associação entre o feminino e o corpo – especificamente, a beleza – em um quadro diferente do masculino, que traz caracterizações mais equilibradas entre papel social e caráter, com pouco destaque para a aparência. Quando levamos em conta todos os predicadores utilizados e a categorização em eixos, vemos um cenário um pouco mais equilibrado, mas ainda com predominância da aparência para o feminino, e sobretudo da aparência associada à beleza. Quantitativamente, o eixo emoção é simétrico, com uma distribuição idêntica. No entanto, levando em conta preferências e exclusividades, o quadro se transforma mais uma vez: medo, tristeza, ciúme e vaidade aparecem como tipicamente femininos; coragem e contentamento são tipicamente masculinos.

Reconhecemos, também, que apesar de o *corpus* anotado nos fornecer diferentes camadas de análise, há espaço para ajustes e melhorias. Sabemos que nem todas as predicações se realizam linguisticamente conforme as estruturas propostas. Sabemos, também, que quase 40% das frases do OBras não têm um sujeito explícito, conforme levantamento preliminar feito por Freitas e Souza (2021). Quando nos damos conta de que tanto os padrões de predicação quanto os padrões associados às ações pressupõem um sujeito explícito na frase, somos forçadas a admitir que podemos não ter capturado uma parcela dos dados do corpus (e vale notar que esta é uma preocupação inexistente em trabalhos de língua inglesa).

Por fim, o estudo aqui apresentado serviu de base para a implementação de um esquema de anotação das predicações em diversas obras literárias, acessível para consulta[20], permitindo assim que mais pesquisas sejam feitas levando em conta os quatro eixos semânticos propostos.

REFERÊNCIAS

ARSTEIN, R. Inter-Annotator Agreement. *In*: IDE, N.; PUSTEJOVSKY, J. (ed.).*Handbook of Linguistic Annotation*. Dordrecht: Springer, p. 297-313, 2017.

BAKER, P. *Sociolinguistics and Corpus Linguistics*. Edinburgh: Edinburgh University Press, 2010.

[20] Caracterização Humana: http://www.linguateca.pt/Gramateca/PredicacaoHumana.html

BAKER, P.; GABRIELATOS, C.; KHOSRAVINIK, M. *et al.* A Useful Methodological Synergy? Combining Critical Discourse Analysis and Corpus Linguistics to Examine Discourses of Refugees and Asylum Seekers in the UK Press. *Discourse & Society*, v. 19, n. 3, 273-306, 2008.

BARDIN, L. *Análise de conteúdo* Lisboa: Edições 70, 1977.

BICK, E. PALAVRAS, a Constraint Grammar-based Parsing System for Portuguese. *In:* SARDINHA, T.B. ; FERREIRA, T. (ed.). *Working with Portuguese Corpora*. London/New York: Bloomsburry Academic, 2014. p. 279-302.

BUTLER, J. *Problemas de gênero:* feminismo e subversão de identidade. 17. ed. Rio de Janeiro, Civilização Brasileira, 2019.

CAMERON, D. *More Heat Than Light?* Sex-difference Science & the Study of Language (Garnett Sedgewick Memorial Lecture) Vancouver: Ronsdale Press, 2012.

CAMERON, D.; PANOVIĆ, I. *Working with Written Discourse*. London: Sage, 2014.
COSTA, B.; FREITAS, C. Um léxico de verbos do dizer para tradutores – e considerações sobre a classificação dos verbos de elocução. *Calidoscópio*, *[S. l.]*, v. 17, n. 3, p. 494-512, 2019.

FREITAS, C.; MARTINS, F.; BIAR, L. Um 'olhar discursivo' sobre predicação e gênero: aproximações metodológicas entre corpus e discurso. *Texto Livre*, Belo Horizonte - MG, v. 15, p. e36213, 2022.

FREITAS, C.; SOUZA, E. de. Sujeito Oculto às claras: uma abordagem descritivo computacional. *Revista de estudos da linguagem*, [S.l.], v. 29, n. 2, p. 1033-1058, mar. 2021.

HEARST, M. A. Automatic acquisition of hyponyms from large text corpora. *In:* INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, 14, 1992, Nantes. *Proceedings* […]. Nantes, 1992.

MAUTNER, G. Corpora and critical discourse analysis. *In*: BAKER, P. (ed.). *Contemporary Corpus Linguistics*. London: Continuum, 2009.

McENERY, T.; HARDIE, A. *Corpus Linguistics: Method, theory and practice*. Cambridge: Cambridge University Press, 2012.

McENERY, T.; WILSON, A. *Corpus Linguistics*: An Introduction. Edinburgh University Press, 2001.

MORETTI, F. Conjectures on world literature. *New Left review*. Vol. 1, 54-68, jan.-feb 2000.

MORETTI, F. *A Literatura vista de longe*. Trad. Anselmo Pessoa Neto. Porto Alegre: Arquipélago, 2008 [2005].

MULVEY, L. Prazer Visual e Cinema Narrativo. *In*: XAVIER, I (org.). *A experiência do cinema*. Rio de Janeiro: Edições Graal, 1973. p. 437-454.

SAMPSON, G. *Empirical Linguistics*. London: Continuum, 2001.

SANTOS, D. Literature studies in Literateca: between digital humanities and corpus linguistics. *In*: DOERR, M.; EIDE, Ø; GRØNVIK, O; KJELSVIK, B. (ed.). *Humanists and the digital toolbox:* In honour of Christian-Emil Smith Ore. Oslo: Novus Forlag, 2019. p. 89-109.

SANTOS, D. Corporizando algumas questões. *In*: TAGNIN, S. E. O.; VALE, O. A. (ed.). *Avanços da Lingüística de Corpus no Brasil.* São Paulo: Editora Humanitas/FFLCH/USP, 2008, p. 41-66.

SANTOS, D.; BICK, E. Providing Internet access to Portuguese corpora: the AC/DC project. *In*: GAVRILIDOU, M.; CARAYANNIS, G.; MARKANTONATOU, S.; PIPERIDIS, S.; STAINHAUER, G. (ed.). *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000).* Atenas, Grécia, 31 de Maio a 2 de Junho de 2000. p. 205-210.

SANTOS, D.; MARQUES, R.; FREITAS, C.; SIMÕES, A.; MOTA, C. Comparando anotações linguísticas na Gramateca: filosofia, ferramentas e exemplos. *Domínios de Lingu@gem*, *[S. l.]*, v. 9, n. 2, p. 11-26, 2015. Disponível em: https://seer.ufu.br/index.php/dominiosdelinguagem/article/view/30798. Acesso em: 25 mar. 2022.

SANTOS, D.; FREITAS, C.; BICK, E. OBras: a fully annotated and partially human-revised corpus of Brazilian literary works in the public domain. *OpenCor*, Canela, RGS, Brasil, 24 de setembro de 2018.

SILVA, F. M. Representações de gênero na caracterização de personagens: uma proposta metodológica e primeiros resultados. Rio de Janeiro, 2021. 169 p. Dissertação (Mestrado em Letras) – Departamento de Letras, Pontifícia Universidade Católica do Rio de Janeiro, 2021.

SMITH, S. CHOUEITI, M.; PIEPER, K. *Gender Bias Without Borders*: An Investigation of Female Characters in Popular Films Across 11 Countries. The Geena Davis Institute on Gender and Media and the Social Change Initiative at USC Annenberg, 2014. Disponível em: https://seejane.org/wp-content/uploads/genderbias-without-borders-executive-summary.pdf Acesso em: 16 set. 2019.

SONI, S.; KLEIN, L.; EISENSTEIN, J. Abolitionist Networks: Modeling Language Change in Nineteenth-Century Activist Newspapers. *Journal of Cultural Analytics* (I), p. 1-43, 2021.

UNDERWOOD, T.; BAMMAN, D.; LEE, S. The transformation of gender in English-language fiction. *Journal of Cultural Analytics*, v. 3, n. 2, 2018.DOI: https://doi.org/10.22148/16.019

UNDERWOOD, T. Distant Reading and Recent Intellectual History. *In*: GOLD, M.; KLEIN, L. (ed.). *Debates in the Digital Humanities 2016*. University of Minnesota Press, Minneapolis, 2016.

Recebido em 03/04/2022. Aceito em 30/05/2022.