Proposta recebida em Maio 2023 e aceite para publicação em Junho 2023.

Desafios e vantagens do processo de identificação automática do gênero e das profissões das personagens no DIP

Challenges and advantages of the automatic identification of character gender and professions in DIP

Emanoel Pires
Universidade Estadual do Maranhão

Marcia Caetano Langfeldt

Rebeca Schumacher Fuão

Resumo

O desenvolvimento de sistemas para identificação automática de personagens e de algumas de suas características é o objetivo central do projeto Desafio de Identificação de Personagens (DIP) desenvolvido junto à Linguateca. Dentre essas características, trataremos neste artigo da identificação do gênero e das profissões das personagens. Primeiramente, justificaremos a nossa escolha em trabalhar com esses dois dados, apresentando os diferentes caminhos que trilhamos para estabelecer diretrizes para a identificação dos mesmos. A identificação manual do gênero e da profissão é exaustiva e passível de falhas, sendo cada vez mais comum o uso de sistemas computacionais para essa tarefa. A análise das profissões permitiria refletir sobre questões como a definição de profissão, sua frequência em obras brasileiras e portuguesas, e possíveis relações com os gêneros literários. Em seguida, apresentaremos alguns resultados provenientes da leitura distante e da leitura próxima de um grupo de obras. Contrastaremos esses resultados e comentaremos os desafios e as vantagens que encontramos ao longo dessa tarefa e que parecem reforçar a nossa hipótese de preferência por um esforço combinado de sistemas automáticos e interpretação humana na identificação de personagens.

Palavras chave

leitura distante, identificação de personagens, gênero, profissão

Abstract

The development of systems for automatic identification of characters and some of their characteristics is the central objective of the Character Identification Challenge (DIP) project developed in conjunction with Linguateca. Among these characteristics, 2 this article will focus on the identification of gender and professions of the characters. Firstly, we will justify our choice to work with these two data sets, presenting the different paths we have taken to establish guidelines for their identification. Manual identification of gender and profession is exhaustive and susceptible to errors, making the use of computer systems increasingly common for this task. The analysis of professions would allow reflection on issues such as the definition of a profession, its frequency in Brazilian and Portuguese works, and possible relationships with literary genres. We present some results from distant and close reading of a group of works, contrast these results and comment on the challenges and advantages we encountered throughout this task, which seem to reinforce our hypothesis of a preference for a combined effort of automatic systems and human interpretation in character identification.

Keywords

distant reading, character identification, gender, profession

1. Introdução

Identificar personagens a partir da aplicação de sistemas automáticos num corpus de literatura em prosa de língua portuguesa é a questão central do projeto desenvolvido junto à Linguateca intitulado Desafio de Identificação de Personagens (DIP) (Santos et al., 2022). Esta tarefa tem nos revelado novos dados impossíveis de serem identificados numa leitura próxima (*close reading*), primeiramente porque a velocidade de identificação é imensamente inferior àquela realizada pelos sistemas e, em seguida, porque o volume de obras passíveis de serem trabalhadas é infinitamente inferior àquele que os sistemas podem dar conta. Já no início dessa atividade, um primeiro desafio foi levantado pelo grupo. Era preciso retomar o conceito de personagem e estabelecer diretrizes

DOI: 10.21814/lm.15.1.401

para o que os sistemas considerariam e o que descartariam como sendo personagens. Além disso, era preciso igualmente apontar as variadas formas pelas quais as personagens são 3referidas de modo a evitar o máximo possível as perdas que os sistemas poderiam ter em suas identificações.

Isso nos levou a renovar uma discussão sobre o que são as personagens literárias. Questionamentos como: o que faríamos com personagens que são animais? E com as figuras históricas mencionadas no interior da obra? Esta análise pode ser lida na íntegra no artigo sobre personagens publicado na Linguateca (Langfeldt et al., 2021). O que fica de importante para este artigo é que decidimos de que não haveria identificação de personagens que não contribuíam para o desenvolvimento da narrativa, como por exemplo personagens históricas que apenas são citadas no texto. Isso porque o nosso interesse justamente era o de observar personagens construídas dentro dessas obras e alguns elementos que faziam parte das escolhas dos autores nessas construções e que poderiam, portanto, indicar tendências de diferentes épocas, países e estilos nos textos analisados.

Dentre as categorias de nosso interesse, falaremos aqui sobre o gênero e a profissão das personagens. A identificação do gênero das personagens é um aspecto crucial da análise literária, uma vez que o modo como um gênero é apresentado na prosa ficcional pode revelar aspectos sobre o contexto cultural e histórico da obra, bem como as intenções e modos de pensar do seu autor. Entretanto, identificar manualmente o gênero das personagens de um largo conjunto de dados pode ser um processo exaustivo e passível de falhas e omissões. Neste sentido, o uso de sistemas de computador para identificar e analisar o gênero das personagens literárias tem se tornado cada vez mais comum na pesquisa literária (Bamman et al., 2014; Elsner, 2012).

A identificação das profissões viria complementar essa análise nos permitindo refletir sobre algumas outras questões como: o que seria considerado como profissão e o que seria considerado como uma mera ocupação ou estatuto social? Seria possível identificar profissões mais frequentes em obras brasileiras que portuguesas? E entre os gêneros, haveria alguns gêneros privilegiados em determinadas profissões? Encontraríamos alguma constância e/ou aspectos reveladores nos resultados?

1.1. A questão do gênero

O gênero na literatura se assemelha e se diferencia do seu conceito para- textual, no sentido em que, embora as personagens possam ser classificadas em gêneros mais ou menos precisos, o contexto literário no qual estão inseridos pode deslocá-las desta posição. Talvez o caso mais emblemático da literatura brasileira de uma personagem que muda de gênero seja o de Diadorim/Reinaldo, do romance *Grande Sertão: Veredas* (1956), de João Guimarães Rosa (1908–1967), em que a personagem Diadorim se traveste do jagunço Reinaldo, e mais recentemente, exemplos como os ilustrados no romance *Stella Manhattan* (1985), de Silviano Santiago e no livro reportagem *Ricardo e Vânia* (2019). No primeiro, não há exatamente uma mudança de gênero da personagem, mas a criação, pela personagem Eduardo da Costa e Silva, de um *alter ego* chamado Stella Manhattan. No último exemplo, Chico Felitti acrescenta ao conhecido relato sobre Ricardo a história de Vânia, que antes já tinha se chamado de Vagner.

É importante distinguir entre o gênero de uma personagem literária e o gênero ou o sexo de uma pessoa no mundo real, pois se tratam de dois conceitos diversos. Quando utilizamos sistemas de computação para identificar o gênero de uma personagem literária, estamos nos referindo à sua representação textual em uma obra literária, e não ao gênero ou sexo na vida real.

O gênero de uma personagem se refere ao modo como ela é representada em uma obra literária, inclusive os traços de personalidade, os comportamentos, as ações e os estereótipos de gênero. O gênero de uma personagem literária é, portanto, uma construção social que pode ser influenciada pela cultura, o período histórico e a intenção do autor.

1.2. A questão da profissão

A preocupação em identificar a ocupação surgiu do mesmo interesse em recolher informações que julgávamos básicas para caracterizar uma personagem. No início das nossas reflexões, logo nos demos conta que a escolha de um termo "guarda-chuva," como profissão ou ocupação, poderia apresentar complicações. Antes da equipe realizar a leitura das 43 obras que comporiam a coleção dourada e que depois passariam pelos sistemas que testariam as identificações, vários membros do grupo realizaram a leitura de quatro obras: *Dom Casmurro* (1899), *As Pupilas do Senhor Reitor* , *Quincas Borba* e *Dramas da Côrte* (1905). Essas leituras foram seguidas de

discussões que nos ajudaram a perceber o tipo de dificuldade que teríamos com as demais obras. Foi o caso, por exemplo, dos escravizados que apareceram em Dom Casmurro e do agregado, personagem que se faz passar por médico homeopata e depois assume que havia mentido. Em ambos os casos, decidiu-se rapidamente que não caberia classificar os escravizados ou o agregado como profissão ou ocupação, mas como algo que desse conta da condição social dos mesmos. Foi então que pensamos no termo estatuto social.

Em *Dramas da Côrte*, por sua vez, percebemos rapidamente outro desafio que os romances históricos, haviam vários do século XIX na nossa coleção, apresentariam: as personagens que possuíam títulos da nobreza ou da realeza ("príncipe", "duque", "conde", "baronesa"). Pensamos que igualmente o termo estatuto social poderia dar conta.

Restava, ainda, a diferença entre o que consideraríamos profissão ou ocupação. Havia muitos casos como "colecionador", "ladrão" ou "curandeiro." Esses pareciam ser mais atividades que as personagens exerciam para ocupar o seu tempo e que, por terem recebido a mesma, eram informações relevantes para o papel da personagem no desenvolvimento da narrativa.

Foi assim que decidimos arranjar os dados nesses três termos: profissão, ocupação e estatuto social. Para dar embasamento a essas escolhas, confirmamos a definição desses termos nos dicionários. Iniciando por profissão, na Infopédia, encontramos a seguinte definição: "exercício habitual de uma atividade econômica como meio de vida; ofício; mister; emprego; ocupação." No Dicionário de usos do Português Borba (2002), encontramos uma definição que complementava ainda este sentido "atividade ou ocupação especializada, e que supõe determinado preparo; ofício."

Já o termo ocupação apresentava como sinônimo também o termo profissão, mas apresentava a possibilidade de ser igualmente uma "atividade ou serviço em que se gasta algum tempo," o que ia ao encontro do que pensávamos sobre a distinção entre profissão e ocupação. Para terminar, procuramos a definição de "estatuto" e encontramos em uma das possibilidades "situação social; condição ou posição hierárquica; status." Decidimos, assim, que "estatuto social" daria conta de classificações ligadas a títulos de nobreza e funções dentro de um reino. Deixamos ainda uma quarta possibilidade para aquelas classificações que tinham sido realizadas de forma equivocada, agrupadas no termo NP, não profissão. Além disso, com relação às personagens que tinham tido uma profissão identificada, dividimos as mesmas em tipos, a fim de verificar se haveria grupos de profissões que eram mais frequentes e, também, grupos de profissões que eram mais comuns para personagens femininas ou masculinas. Dividimos as mesmas em profissões serviçais (PS), por exemplo "camareira", "ama", "cabra", "motorista"; liberais (PL), como "padeiro", "peixeiro", "pianista"; militares (PM), como "general", "infante" e "marechal"; e, por fim, as religiosas (PR) ao exemplo de "vigário", "arcebispo" e "padre."

A partir dessas classificações, passamos a aplicar os sistemas para obter alguns resultados que poderiam nos indicar algumas características de cada época ou literatura. Por exemplo, seria alguma profissão, ocupação ou estatuto social (POES) mais comum para as mulheres que para os homens? Haveria alguma diferença nas POES identificadas na literatura brasileira e portuguesa? Quais profissões eram mais frequentes em cada século?

Para responder a essas questões, julgamos que também seria interessante classificar as próprias profissões em tipos: profissões liberais, profissões militares, profissões religiosas e pessoal serviçal. Comentaremos os resultados dos sistemas aplicados tanto na CD quanto no PALAVRAS na Seção 4 "Alguns resultados do DIP."

2. Possibilidades e ganhos na identificação do gênero e das POES por sistemas automáticos

A primeira evidente vantagem do uso de sistemas automáticos na identificação de gênero e das POES das personagens é a velocidade da máquina. Imensamente maior do que aquela que os leitores humanos podem fazer, ainda que se tome aqui um grupo numeroso e experiente de pesquisadores.

A segunda grande qualidade desta abordagem, decorrente da primeira, é que ela permite que se identifique o gênero e a POES das personagens em um grande conjunto de dados, possibilitando aos pesquisadores que se concentrem em outros aspectos, tais como a análise do resultado encontrado. Isso significa não apenas uma maior amplitude do corpus a ser pesquisado, como um aprofundamento da análise em relação a um dado período, estilo de época ou conjunto de autores (possibilitando a comparação entre autores de países diversos, por exemplo).

É importante notar que a virada do estudo em larga escala proporcionada pela análise quantitativa dos sistemas computacionais gerou no-

vas formas de abordagens críticas, antes inimaginadas pela teoria literária (Piper et al., 2017). Tais análises demandam diversas formas de investigação e modelos, com a utilização de métricas, mapas, árvores e padrões sistematizados, menos comuns dentro da teoria literária até os anos 2000 (Moretti, 2005). Ou seja, o uso sistemático desta forma de abordagem necessariamente traz amplitude e um olhar diverso sobre as obras de um dado período ou região.

Além disso, a capacidade de identificar personagens automaticamente pode ter implicações significativas para os estudos literários, tais como a facilitação na análise da evolução da personagem na narrativa, assim como das suas relações, possibilitando estudos em larga escala de tendências e padrões, o que certamente abre novas avenidas para a crítica literária lusófona.

Neste sentido, o resultado desta abordagem pode ser utilizado para identificar tendências em relação às redes sociais e aos ambientes em que dados gêneros são descritos no texto, tais como quais personagens são mais centrais para a narrativa, quais são mais interconectadas do que outras etc. Associado a outros tópicos explorados também pelo DIP, como as relações familiares e a ocupação, estes insights fornecidos pelos sistemas automáticos podem contribuir para se lançar um novo olhar para a estrutura e a dinâmica das narrativas literárias.

3. Desafios

Em um primeiro momento, é preciso discutir a complexidade da tarefa de identificar uma personagem através de sistemas automáticos. Uma distinção primordial é entre os atos de ≪identificar≫ e ≪perceber≫, fundamental no que se refere a esta tarefa, já que o leitor humano realiza as duas operações simultaneamente. No Dicionário Online de Português,[1] a palavra identificar tem as seguintes definições: "conseguir comprovar ou definir a identidade de; saber quem é", "distinguir ou ter a capacidade de reconhecer (alguém ou alguma coisa)", enquanto perceber é definido assim: "entender o significado de algo através da inteligência". Falando de modo muito simplificado, a percepção implica a pré-existência de um dado objeto, enquanto a representação (literária) refere-se necessariamente a um elemento ausente, mas que aparece em cena graças a ela. A percepção de uma personagem literária ocorre quando o leitor produz uma imagem mental dela, ou seja, a personagem literária não existe a priori, tampouco ela é uma pura criação do seu autor, mas ela é necessariamente formada a partir da percepção do leitor, ela é sujeita a uma perspectiva, a um ponto de vista, a um "equipamento" intelectual e cultural, como observa Jouve (1998):

> É preciso observar que a identidade da personagem apenas pode ser concebida como o resultado de uma cooperação produtiva entre o texto e o sujeito leitor. O romance não dispõe dos meios, por conta própria, de dar uma percepção global da personagem. As razões são claramente formuladas por Umberto Eco na sua análise dos mundos narrativos. O universo induzido por um romance se caracteriza, em efeito, por uma ausência de autonomia, na medida em que: 1. de um ponto de vista formal, o texto não pode descrever exaustivamente um mundo; 2. de um ponto de vista semiótico, é inimaginável: a. estabelecer um mundo alternativo completo; b. Descrever como completo o mundo "real". Os universos narrativos, incapazes de constituir por eles mesmos os mundos possíveis, são obrigados de tomar emprestado algumas de suas propriedades do mundo de referência do leitor.[2]

Por consequência, uma personagem não é apenas um ator romanesco criado pelo autor, mas ela é também tudo que o leitor lhe atribui. Em uma pesquisa que pode se aproximar do DIP em alguns aspectos (embora com objetivos diversos), intitulada "Extração e análise de redes de personagens ficcionais", Labatut & Bost (2019) se depararam com alguns dilemas, no que diz respeito à identificação de personagens em literatura a partir de sistemas automáticos, em comparação à mesma tarefa realizada com textos não ficcionais, que seria interessante contemplar aqui. Segundo os autores, o mais desenvolvido programa

[1] https://www.dicio.com.br/identificar/ Acesso em 15/06/2023.

[2] Il convient de remarquer que l'identité du personnage ne peut se concevoir que comme le résultat d'une coopération productive entre le texte et le sujet lisant. Le roman n'a pas, à lui seul, les moyens de donner une perception globale du personnage. Les raisons en sont clairement formulées par Umberto Eco dans sons analyse des mondes narratifs. L'univers induit par un roman se caractérise, en effet, par une absence d'autonomie dans la mesure où 1/ d'un point de vue formel, le texte ne peut décrire exhaustivement un monde ; 2/ d'un point de vue sémiotique, il est inimaginable a/ d'établir un monde alternatif complet, b/ de décrire comme complet le monde ≪ réel ≫. Les univers narratifs, incapables de constituer par eux-mêmes des mondes possibles, sont obligés d'emprunter certaines de leurs propriétés au monde de référence du lecteur.

não pode produzir uma análise significativa de um texto literário, pois os dados precisam ser analisados por humanos (e, em particular, especialistas na área em questão). Além do mais, a intervenção humana é também necessária antes da execução da máquina, pois muitas vezes é preciso "preparar" os dados a serem analisados pelo computador. O que está completamente alinhado aos postulados do DIP, uma vez que a organização do desafio contava não apenas com especialistas na área de informática e processamento de linguagem natural, mas também com pesquisadores da área de literatura. Ainda segundo os autores Labatut & Bost (2019):

> Nos textos, a prosa literária é considerada mais complexa do que a prosa jornalística, ainda mais quando a obra é mais antiga. [...] Por exemplo, para detecção de personagens em romances: muitos personagens são parentes e compartilham o mesmo sobrenome; eles carregam apelidos; alguns personagens fictícios são objetos inanimados na vida real; escritores usam honoríficos específicos correspondentes a convenções sociais complexas, possivelmente desatualizadas e até imaginárias; e eles criam nomes para transmitir certo significado ou função. Para resolução de co-referência, o problema está relacionado a sentenças mais longas, uso mais frequente de pronomes e discurso direto, cadeias de co-referência mais numerosas e curtas.[3]

Além disso, há alguns outros desafios no uso de sistemas de computador para identificar o gênero na literatura. Um deles é a ambiguidade da identificação. Muitas obras literárias apresentam personagens que possuem características que podem ser interpretadas tanto como masculinas quanto femininas, ou que desafiam intencionalmente as normas de gênero. Nesses casos, os modelos de aprendizado de máquina e os que utilizam regras podem ter dificuldades para identificar com precisão o gênero da personagem.

Por exemplo, no livro *Ensaio sobre a Cegueira* (1995), do escritor português José Saramago (1922–2010), o gênero de várias personagens não é explicitamente declarado, deixando-o aberto à interpretação. Essa ambiguidade, também presente na leitura humana, pode tornar desafiador para os sistemas identificarem com precisão o gênero das personagens literárias.

Outro importante desafio de identificação das personagens é a variabilidade dos nomes e referências. As personagens podem ser referidas por nomes ou apelidos diferentes no decorrer da narrativa e podem ser identificadas através de inúmeras referências indiretas, tais como pronomes ou descrições. Além disso, se a única referência ao gênero da personagem for o nome e esta tiver um nome ambíguo, tal como ocorre com certos nomes de personagens indígenas, então não é possível determinar com precisão o gênero de uma personagem. Pensando nesses casos, que são ambíguos tanto para sistemas quanto para leitores reais, o DIP estabeleceu uma categoria em que uma personagem poderia ter os dois gêneros ou mesmo gênero nenhum (0). Os sistemas que poderiam concorrer ao desafio tiveram acesso a esse tipo de informação através das Perguntas Técnicas, disponibilizadas no site da avaliação conjunta.[4]

Do mesmo modo, em língua portuguesa, há inúmeras palavras que, sendo atributos de uma personagem ou a sua definição, não permitem estabelecer um gênero preciso. Este é o caso do substantivo sobrecomum, que é um substantivo uniforme, que apresenta apenas um termo para os dois gêneros que existem em língua portuguesa (masculino e feminino). São palavras tais como: a criança (para menino ou menina), o anjo (Maria é um anjo), cônjuge, defunto (o defunto era Maria), a estrela (de cinema), pessoa, monstro, testemunha, vítima, etc.

Além disso, as personagens podem ser introduzidas gradualmente no decorrer da narrativa e as suas relações evoluem, apresentando igualmente um caráter ambíguo e aberto à interpretação. Em outras palavras, como notou Mark Algee-Hewitt em sua pesquisa sobre a identificação sistemática de personagens dramáticas da literatura inglesa, é importante que o pesquisador entenda o que se deve medir através das análises quantitativas (Piper et al., 2017). O gênero das personagens, por exemplo, é um dos aspectos a se considerar quando se propõe a investigar as redes de relacionamento constituídas na narrativa.

Um outro aspecto limitante a se levar em

[3] In texts, literary prose is considered as more complex than journalistic prose, and even more so when the work is older. [...] For instance, for character detection in novels: many characters are relatives and share the same last name; they bear nicknames; some fictional characters are inanimate objects in real life; writers use specific honorifics corresponding to complex, possibly outdated, and even imaginary social conventions; and they craft names to convey certain meaning or function. For co-reference resolution, the problem comes from longer sentences, more frequent use of pronouns and direct speech, more numerous and shorter co-reference chains.

[4] https://www.linguateca.pt/aval_conjunta/dip/pjr.html#ptecnicas

conta é a interseccionalidade de identidade. Personagens literárias podem ter múltiplas intersecções identitárias, tais como raça e classe social, além do gênero. Modelos de pesquisa automática que apenas identifiquem o gênero, por exemplo, podem deixar escapar importantes aspectos de uma personagem que podem afetar a leitura dela.

Por exemplo, no romance brasileiro *O Bom-Crioulo* (1895), de Adolfo Caminha, a personagem Amaro é um marinheiro, negro, ex-escravizado, gay, descrito principalmente pela sua força física, "uma massa bruta de músculos", "um animal inteiro era o que ele era!". Um sistema automático que apenas identificasse o gênero de Amaro poderia perder uma importante interseccionalidade da identidade da personagem e o modo como isto afeta a sua representação no texto. Portanto, é de se destacar que apenas a identificação de um aspecto da personagem traz necessariamente um resultado incompleto e precisa ser colocado em relação a outros eixos que constituem a personagem ficcional. Essa foi a abordagem do DIP.

4. Alguns resultados do DIP

Descrevemos aqui os resultados obtidos por meio do único sistema que se candidatou à avaliação conjunta, o PALAVRAS-DIP (Bick, 2023), dando destaque para o gênero e a POES. Para outras categorias analisadas, conferir os outros artigos do volume.

4.1. Alguns resultados do DIP: o gênero das personagens

Confirmando aquilo que já conhecíamos em virtude da nossa prática de leitura próxima de literatura, a imensa maioria das obras literárias presentes no DIP traz um número maior de personagens masculinas, ou seja, mesmo observando um número considerável de textos, muitos dos quais distantes do cânone e do conhecimento de pesquisadores experientes da área de Letras, a constância da presença masculina nas obras literárias é mantida.

Mesmo em obras com um grande apelo a uma personagem feminina, como *Úrsula* (1859), *A Escrava Isaura* (1875), *A viuvinha* (1857), entre outras, a constância se mantém. O que reflete não só a realidade histórica, com o destaque do sujeito masculino quantitativa e qualitativamente, esse último, no que diz respeito à profissão, ocupação social ou estatuto social, como reforça a estereotipia de gênero, uma vez que a representação numérica masculina superior carrega, também, a ideia de que os homens são mais importantes e desenvolvem mais papeis do que as mulheres. Um estudo das personagens protagonistas, muito provavelmente, confirmaria o que aqui se levanta enquanto hipótese sobre o número total de personagens.

O PALAVRAS-DIP também identificou que em obras publicadas mais recentemente, continua havendo uma menção maior às personagens masculinas, ainda que essa diferença pareça diminuir, o que apenas seria possível de comprovar se tivéssemos no corpus um número maior de obras contemporâneas (o que pode ser feito em trabalhos futuros).

A nossa hipótese para o decréscimo no número de personagens em obras mais recentes está ligada ao gênero do texto: os romances históricos, muito comuns no século XIX e com uma grande quantidade de personagens, perdem força nos séculos seguintes, dando lugar a romances realistas/naturalistas, em que o foco recaía sobre os tipos sociais (Lukács, 2000 [1916]).

O gráfico abaixo, por exemplo, demonstra o peso que os romances históricos têm quando tomamos como base o número de personagens. Autores como Antônio José Coelho Lousada, Zeferino Norberto Gonçalves Brandão e Diogo de Macedo puxam para cima o número de personagens nas obras portuguesas, o mesmo acontecendo com a obra do brasileiro Franklin Távora.[5] Ademais, para alguns comentários sobre a escolha das obras, consultar Santos et al. (2023) neste mesmo volume.

Do ponto de vista da análise histórico-literária, alguns motivos justificam o grande número de personagens nos romances históricos, entre eles:

- A retratação de períodos históricos significativos para uma nação envolve uma gama ampla de atores (políticos, heróis nacionais, a realeza, as classes camponesas e militares, figuras importantes para o período, etc.);
- O tempo narrado costuma ser longo, o que, na maioria das vezes, envolve distintas gerações de personagens;
- A presença de intrigas políticas/culturais/ religiosas/étnicas que enriquecem a narrativa e ajudam a dar complexidade ao enredo.

[5] A lista completa com as obras usadas pelo DIP está em https://www.linguateca.pt/aval_conjunta/dip/colecao.html.

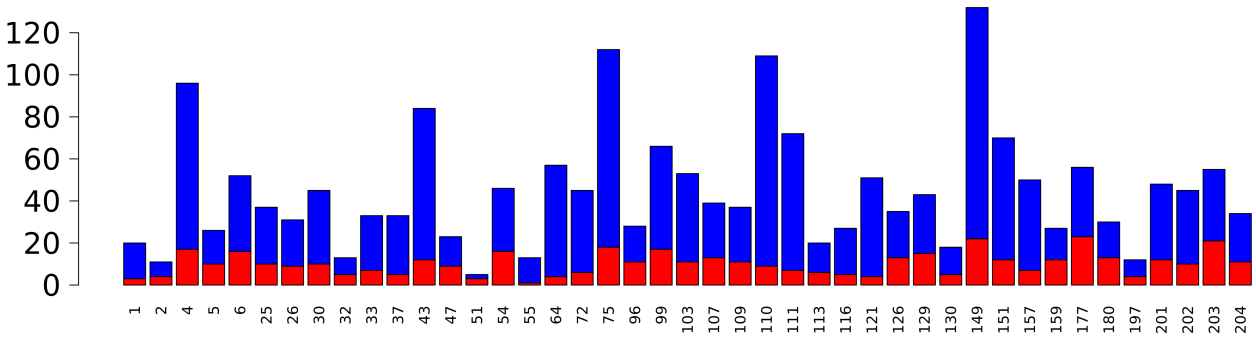

Figura 1: A distribuição de personagens por género na coleção dourada total: a vermelho, as personagens femininas; a azul, as masculinas

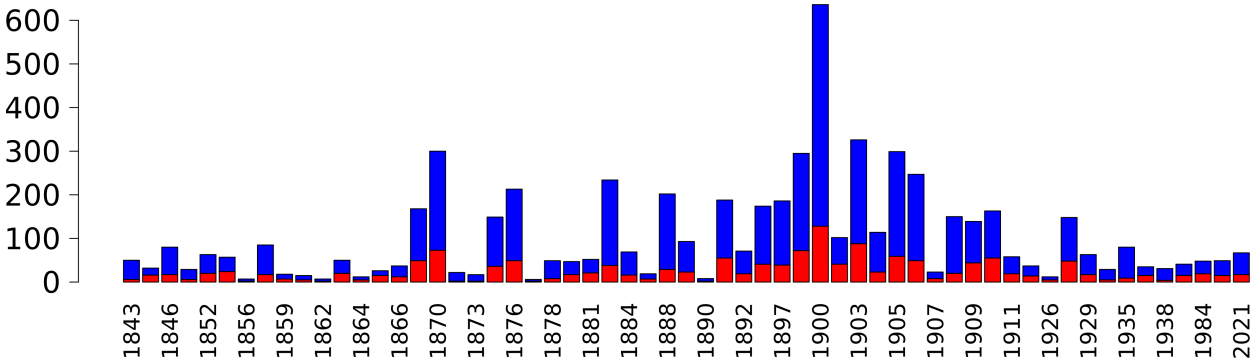

Figura 2: Distribuição do gênero das personagens ao longo do tempo, na resposta do PALAVRAS-DIP

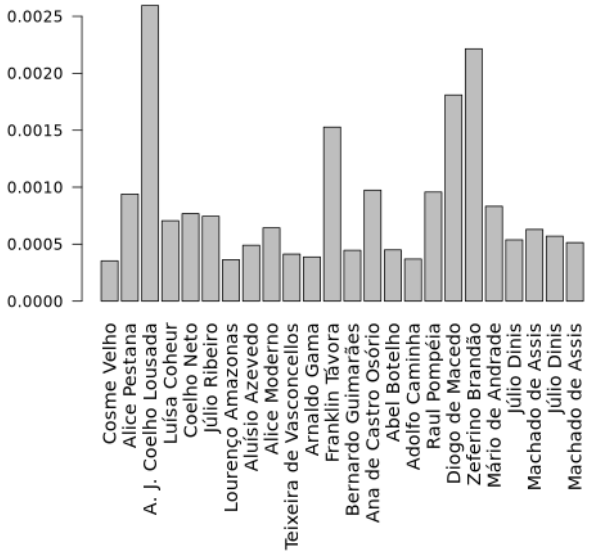

Figura 3: A densidade relativa do número de personagens por obra, na coleção dourada total

A segunda constatação, a diminuição não só de personagens, mas na distância entre o número de personagens masculinas e femininas nas obras mais recentes, pode ser explicada pela expansão social dos direitos femininos ao longo dos séculos XIX e XX, ainda que à personagem feminina recai criada como a POES mais frequente, reforçando a ideia de que a identidade de gênero está diretamente relacionada com outras esferas (estatal, institucional, trabalhista, educativa, doméstica, afetiva, sexual) (Carson, 1995).

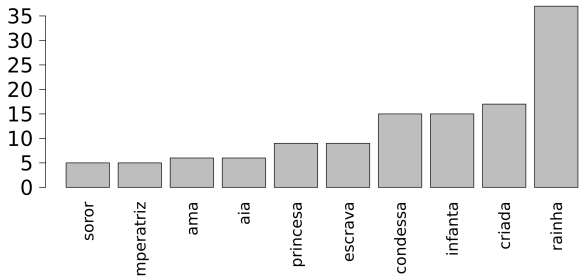

Figura 4: Profissões femininas com mais de 4 ocorrências, na resposta do PALAVRAS-DIP.

Mesmo que os autores, homens na sua grande maioria, tivessem as mulheres como provável público leitor, real ou imaginado, as personagens femininas das obras raramente conquistam posições de prestígio social, e, quando isso é feito, estão sujeitas a tios, maridos ou outra figura masculina, como no caso da personagem Aurélia Camargo, da obra *Senhora* (1875), de José de Alen-

car. Sobre esse público leitor feminino, Candido (2011, p. 94) afirma que:

> Daí um amaneiramento bastante acentuado que pegou em muito estilo; um tom de crônica, de fácil humorismo, de pieguice, que está em Macedo, Alencar e até Machado de Assis. Poucas literaturas terão sofrido tanto quanto a nossa, em seus melhores níveis, esta influência caseira e dengosa, que leva o escritor a prefigurar um público feminino e a ele se ajustar

Ou seja, apesar do leitor intencionado ser a figura feminina, as personagens masculinas são, não só as com maior peso numérico nas narrativas, mas, também, as que desenvolvem papéis de maior destaque social, relegando à mulher o ambiente doméstico.

Ainda no que diz respeito ao tamanho das obras e sua relação com a quantidade de personagens, a pesquisa realizada no contexto do DIP revela a tendência, imaginada, mas demonstrada agora de forma empírica em um grande número de obras, que o número de personagens tende a aumentar à medida que as obras se tornam mais extensas, mesmo que não se possa falar em uma tendência absoluta, uma vez que o gráfico abaixo nos demonstra obras que não apresentam esse padrão (*outliers*).

Também percebemos que as obras portuguesas não só possuem um número maior de personagens como contam com mais personagens femininas que as obras brasileiras, ainda que se repitam aqui POES de menor prestígio social. Como é muito raro um estudo sobre as obras portuguesas que levem em conta um número considerável de obras para, daí, tirar as suas conclusões, os resultados do DIP podem, por exemplo, ajudar a confirmar ou não o que diz Barreira (1986) sobre a personagem portuguesa oitocentista (tomando em conta apenas dois autores, Almeida Garret e Eça de Queiroz). Para a autora, a personagem portuguesa desse período é, em geral, fútil, vazia, quase ridícula, além de estar submissa aos caprichos e desejos masculinos.

4.2. Alguns resultados do DIP: a POES

Nesses primeiros dois gráficos (figuras 8 e 9) relativos a POES, vemos que tanto a classificação manual (CD) quanto a automática (PALAVRAS-DIP) identificaram em maior número ocupações que eram profissões que outros tipos de classificações, chegando ambas as avaliações em resultados muito semelhantes. Esse tipo de resultado nos dá confiança na aplicação dos sistemas, desde que direcionemos os mesmos de forma correta.

No grupo de figuras 10 a 15 vemos primeiramente que ambas as classificações chegaram a resultados muito semelhantes no que diz respeito ao volume de profissões classificadas em cada categoria (PL, PS, PM e PR). O que podemos perceber, e aí analisando a POES juntamente com o gênero, é a clara diferença entre as POES masculinas e femininas, tanto na coleção dourada, que foi atribuída manualmente, quanto pelo PALAVRAS-DIP. Se olharmos para profissões classificadas como serviçais (PS), ao exemplo de criada, vemos que é a POES feminina mais mencionada tanto pelo PALAVRAS-DIP quanto na CD. As outras posições com o maior número de ocorrências dizem respeito a POES de mais prestígio e/ou poder de controle social, como padres, reis, capitães e coronéis e são claramente mais atribuídas a personagens masculinas.

5. Conclusões e apontamentos para o futuro

Conforme visto no DIP, o uso de sistemas computacionais para a identificação do gênero e da POES em literaturas de língua portuguesa apresenta diversas vantagens e desafios. Enquanto a ambiguidade dos atributos, principalmente na identificação do gênero, e a falta de uma grande quantidade de dados anotados podem significar desafios consideráveis, a velocidade, a objetividade, a amplitude e o poder de análise que os sistemas computacionais apresentam os tornam uma ferramenta imprescindível para a pesquisa literária. Em síntese, o gênero das personagens pode ser colocado em perspectiva em relação a outros marcadores do DIP, tais como o gênero e as relações familiares, o gênero e as profissões/estatutos sociais (como demonstramos neste artigo), o gênero e os nomes (nomes que se repetem por gerações, por exemplo).

Com o avanço das pesquisas dedicadas à leitura distante e o desenvolvimento de ferramentas de pesquisa automáticas, esta abordagem na identificação dos diversos atributos de uma personagem literária irá cada vez mais proporcionar novas possibilidades e perspectivas à crítica que se refere à questão do gênero na representação literária lusófona. Proposta para o qual esta iniciativa do DIP vem a ser uma importante etapa fundadora.

Além das dificuldades inerentes à execução da tarefa, cabe ainda abordar a pertinência de tais estudos dentro do âmbito da literatura. Ora,

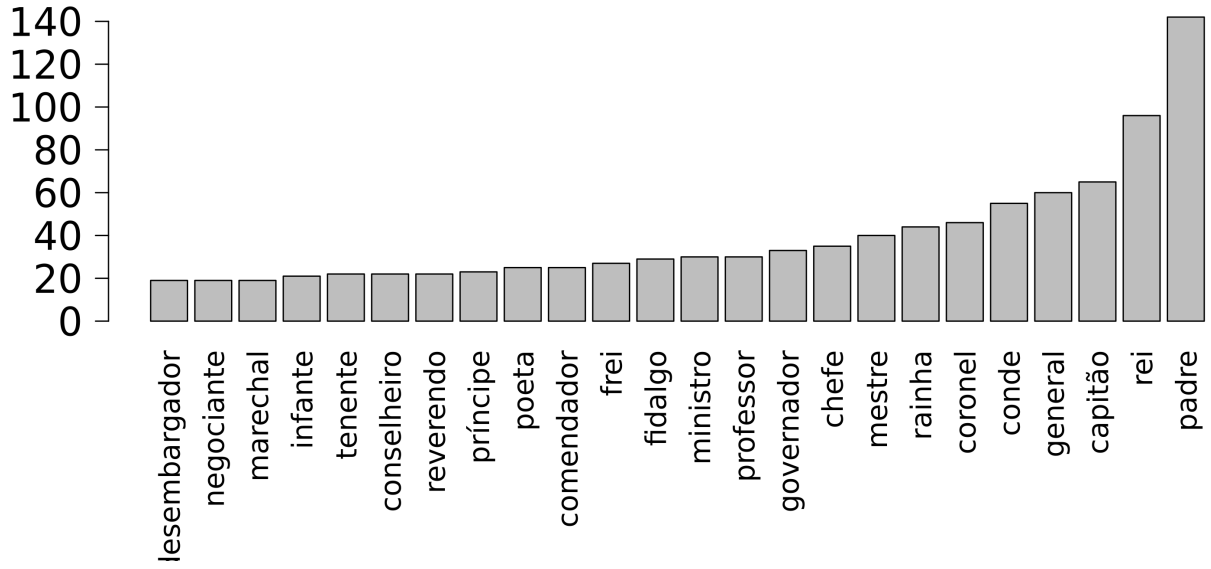

Figura 5: Profissões masculinas e femininas com mais de 18 ocorrências, na resposta do PALAVRAS-DIP.

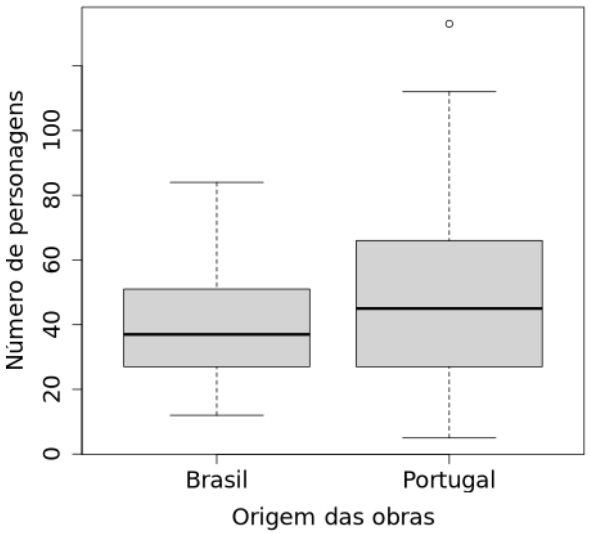

Figura 6: O número de personagens por obra, nas 43 obras brasileiras e portuguesas a que atribuímos uma solução.

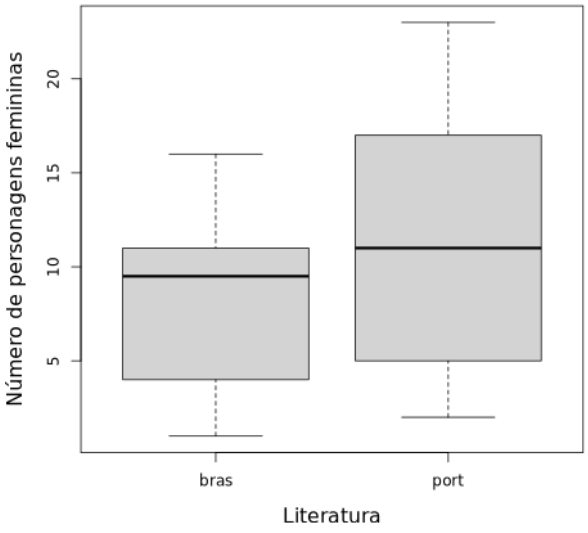

Figura 7: O número de personagens femininas por obra, por literatura.

neste aspecto, Véronique Parenteau traz uma constatação fundamental. As pessoas que geralmente se interessam pelo uso de sistemas automáticos na análise literária não são especialistas de literatura. Embora haja alguns que atuem de fato como professores de literatura em universidades, a grande maioria que se interessa pelo distant reading são não-especialistas de literatura, mas matemáticos, físicos, engenheiros de TI, psicólogos, linguistas, entre outros. Os teóricos da literatura por seu lado ainda fazem pouco uso da informática em suas pesquisas e frequentemente põem esta iniciativa em questão.

Parenteau (1998) afirma sobre este aspecto:

> A análise dos textos literários pelo computador é marginalizada pelos literatos. A maior parte deles não acredita que a informática possa lhes trazer uma ajuda efetiva nos seus trabalhos e parecem não ter a curiosidade de descobrir as possibilidades desta ferramenta. É preciso dizer que uma boa parte dos textos dentro da área da análise de textos por computação são bastante técnicas e um tanto rebarbativas para quem não é muito familiarizado com as estatísticas e a informática. Por outro lado, os especialistas em análise de textos li-

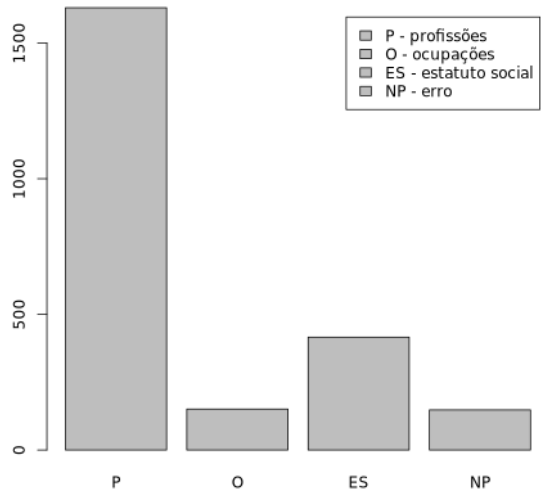

Figura 8: Número de profissões, ocupações e estatutos sociais por grupo na resposta do PALAVRAS-DIP.

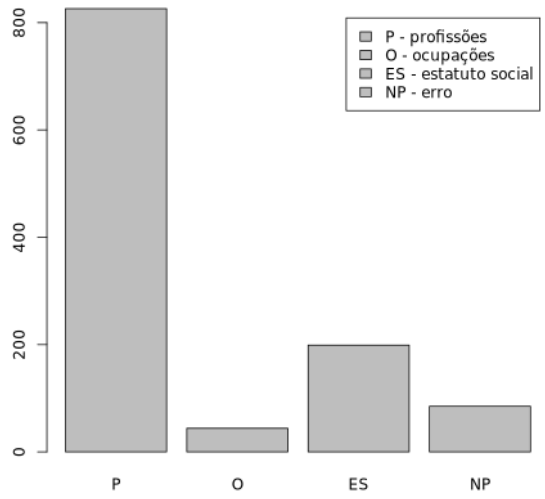

Figura 9: Número de profissões, ocupações e estatutos sociais por grupo na CD total.

> terários assistida por computador não fazem sempre um uso muito pertinente das ferramentas computacionais. Uma grande parte dos estudos se limitam à análise de aspectos muito simples, tais como o tamanho das palavras e das frases, a frequência de certas palavras etc. Em si mesmos, estes resultados de tais análises não são suficientemente interessantes de um ponto de vista estritamente literário. Por outro lado, eles podem ser práticos quando utilizados para fins comparativos, à condição, certo, que a comparação seja pertinente, que o seu autor tenha um objetivo específico.[6]

[6] L'analyse de textes littéraires par ordinateur est marginalisée par les littéraires. La plupart d'entre eux ne croient pas que l'informatique puisse leur apporter une aide réelle dans leurs travaux et ne semblent pas avoir la curiosité de découvrir les possibilités de cet outil. Il faut dire qu'une bonne partie des écrits dans le domaine de l'analyse de textes par ordinateur sont assez techniques et quelquefois rébarbatifs pour qui n'est pas très familier avec les statistiques et l'informatique. D'un autre côté, les experts en analyse de textes littéraires assistée par ordinateur ne font pas toujours un usage très pertinent des outils informatiques. Bien des études se limitent à l'analyse d'aspects très simples comme la longueur des mots et des phrases, la fréquence de certains mots, etc. En eux-mêmes, les résultats de telles analyses ne sont pas très intéressants d'un point de vue strictement littéraire. Par contre, ils peuvent être pratiques lorsqu'utilisés pour fins de comparaison; à la condition, bien sûr, que la comparaison soit pertinente, que son auteur ait un objectif précis.

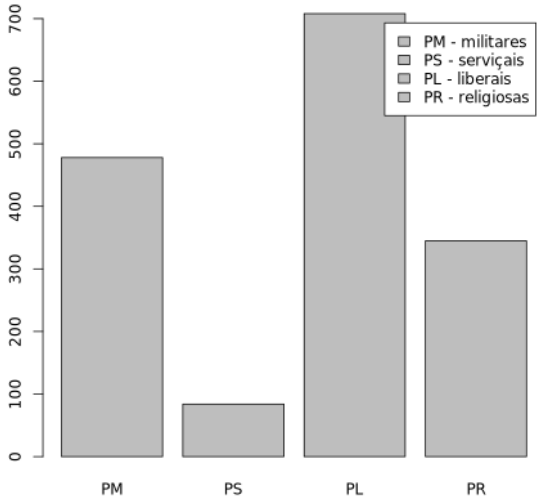

Figura 10: Número de profissões por subgrupo na resposta do PALAVRAS-DIP.

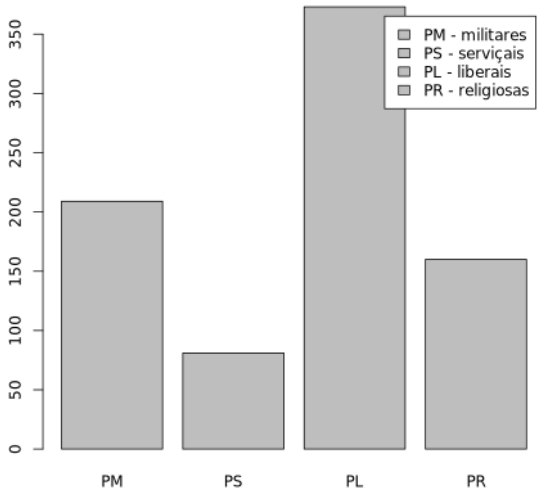

Figura 11: Número de profissões por subgrupo na coleção dourada total.

Dentre as utilidades que Parenteau enumera para o uso da computação nos estudos literários, a comparação é, de fato, a primeira qualidade.

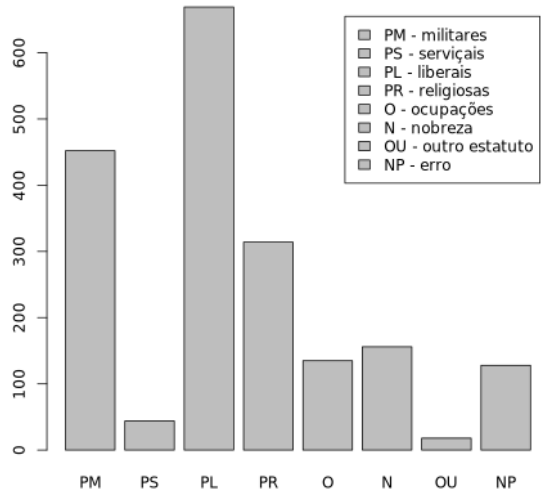

Figura 12: Número de profissões masculinas por subgrupo na resposta do PALAVRAS-DIP.

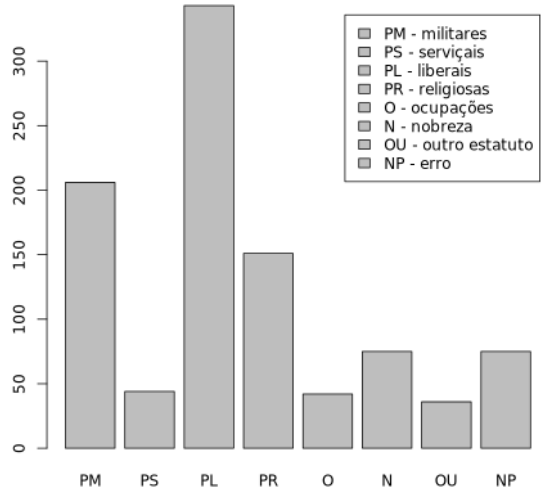

Figura 13: Número de profissões masculinas por subgrupo na coleção dourada total.

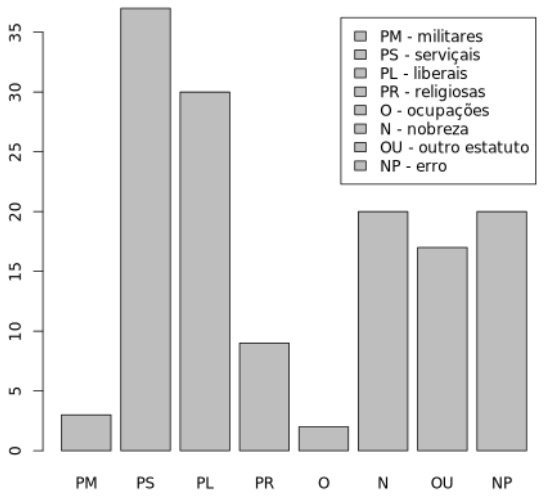

Figura 14: Número de POES femininas por grupo na resposta do PALAVRAS-DIP.

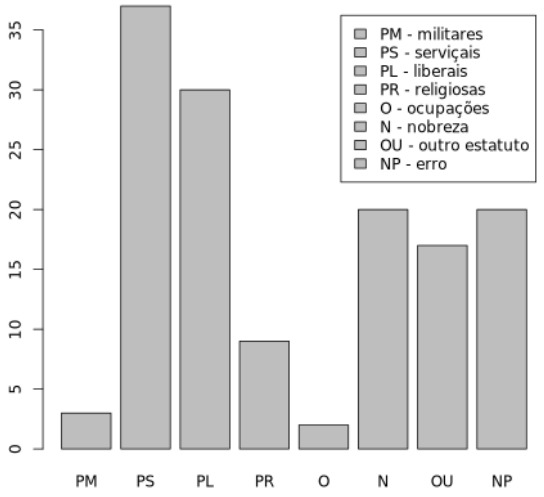

Figura 15: Número de POES femininas por subgrupo na coleção dourada total.

Além desta, ela cita também que estes estudos podem contribuir a: determinar a paternidade de um texto, diferenciar as imitações das obras autênticas, estudar os motivos rítmicos em versos e descobrir as marcas de um autor na evolução da língua. Ou seja, estes são alguns temas que o uso das ferramentas de computação pode contribuir com novas pesquisas.

Seguindo estas indicações de Parenteau, parece-nos que o DIP ainda poderá oferecer muitas oportunidades de novos caminhos dentro dos estudos literários no que se refere às personagens de romances em literatura em língua portuguesa. Vejamos a seguir alguns exemplos que podem vir a ser explorados.

- O nome da personagem como indicativo de informações relevantes à narrativa em questão. Tomemos como exemplo o autor brasileiro Machado de Assis, cuja obra encontra-se em domínio público, digitalizada e disponível, portanto, para o uso de ferramentas computacionais. Em várias de suas obras, os nomes das personagens são indicativos do que irá se desenrolar na narrativa. Por exemplo: em Dom Casmurro (1889), a personagem Capitu seria aquela que capitula? Bentinho seria aquele que crê sem pensamento crítico? Em Memórias póstumas de Brás Cubas (1881), por que a personagem teria o nome de um famoso bandeirante brasileiro? Sendo o bandeirante aquela figura considerada heroica em alguns manuais de história do Brasil que, porém, era um caçador de escravos fugitivos e escravizador de indígenas. Seria coincidência que a personagem de Machado de Assis chicoteia seu escravo desde pequeno e é um tipo de protótipo

de "dono do mundo," no universo machadiano? Será que encontraríamos outros casos de personagens deste tipo, em que o nome tem um duplo sentido com um personagem histórico e com a etimologia das palavras dentro de toda a imensa obra em prosa deste autor? Outra questão interessante é a do livro Esaú e Jacó (1904), em que o título do livro e o nome dos protagonistas — Pedro e Paulo — remetem à bíblia. Teria esta escolha recaído sobre o fato de que o perfil do leitor deste autor seriam mulheres, de classe média, cuja principal leitura, além dos romances, era a bíblia? Este é apenas um caso, mas poderíamos ampliar para outros, como o autor contemporâneo amazonense Milton Hatoum, cuja obra é repleta de nomes de personagens que significam algo na história.

Ora, ocorre que um autor dá nome a seus personagens de um modo diverso como um pai dá a seu filho. Um personagem literário é geralmente "batizado" segundo alguma intenção relacionada à narrativa. Neste sentido, o DIP teria muitas possibilidades a serem exploradas no estudo de inúmeros autores de língua portuguesa.

- Os nomes das personagens como atributos de um dado estilo literário. Aqui pensamos, por exemplo, em um tipo de obra tal como *O Paroara* (1889), de Rodolfo Teófilo, em que a personagem principal da narrativa tem a alcunha de paroara, que dá título ao livro. Vejamos que no caso desta obra, típica do naturalismo brasileiro, os nomes dão frequentemente lugar a apelidos, indicativos de "tipos". Estes tipos por si só sintetizam uma história, como o paroara — aquele retirante do Ceará que emigra para o Amazonas em busca de fazer fortuna na época da borracha e retorna a sua terra natal. Pois bem, haveria muitos outros tipos dentro da literatura brasileira do século XIX? Seria isso uma marca da literatura brasileira ou isso poderia ser encontrado em obras de outros países lusófonos? Está aqui mais uma instigante relação entre narrativa e estudo dos estilos literários que poderia ser explorada com a ajuda do DIP.
- O texto e seu autor — aqui os nomes das personagens, suas relações de parentesco e suas ocupações/posições sociais teriam também muito a revelar. Por exemplo, dentro das relações familiares ou a ocupação das personagens, haveria muitas pistas a explorar sobre as condições de realização da obra literária e das intenções do seu autor. Personagens que têm nomes curtos e simples, gente simples do povo, poderiam indicar, para além do estilo literário (naturalista ou realista), o engajamento do seu autor e a sua intenção em denunciar algum tipo de injustiça social, sua ligação com ideias políticas ou movimentos sociais etc. A simples busca através do sistema desenvolvido pelo DIP possibilitaria traçar um mapeamento deste ramo da literatura dentro do universo lusófono. Inversamente, relações de parentesco com pessoas de classes mais altas ou de profissões de prestígio à época (como o pároco ou o prefeito), todas são informações preciosas e que poderiam ser utilizadas para a análise literária com a ajuda do DIP, que contribuíram muito para uma compreensão mais refinada das obras e de seus autores.

Enfim, assim como indicou Parenteau, haveria certamente muitas outras oportunidades abertas por esta iniciativa precursora dentro dos estudos literários em língua portuguesa no mundo. O fundamental é perceber que o DIP foi uma etapa capital no desenvolvimento de uma nova ferramenta de pesquisa dentro dos estudos literários, que amplia as pesquisas que já vêm sendo realizadas no âmbito das personagens para um número muito maior de obras e que abre novas portas para indagações e caminhos futuros dentro dos estudos de literatura em língua portuguesa.

Agradecimentos

Agradecemos o apoio da FAPEMA pelo financiamento de uma bolsa de pós-doutorado a Emanoel Pires.

Referências

Bamman, David, Ted Underwood & Noah A. Smith. 2014. A Bayesian mixed effects model of literary character. Em *52nd Annual Meeting of the Association for Computational Linguistics*, 370–379. doi 10.3115/v1/P14-1035.

Barreira, Cecília. 1986. Imagens da mulher na literatura portuguesa oitocentista. *Análise Social* 22(92/93). 521–525.

Bick, Eckhard. 2023. Extraction of literary character information in Portuguese. *Linguamática* 15(1). 31–40. doi 10.21814/lm.15.1.397.

Borba, Francisco da Silva. 2002. *Dicionário dos usos do português no Brasil*. Ática.

Candido, Antonio. 2011. *Literatura e sociedade*. Ouro sobre Azul.

Carson, Alejandro Cervantes. 1995. Entrelaçando consensos: reflexões sobre a dimensão social da identidade de gênero da mulher. *Cadernos Pagu* 4. 187–218.

Elsner, Micha. 2012. Character-based kernels for novelistic plot structure. Em *13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 634–644.

Jouve, Vincent. 1998. Le personnage comme produit de l'interaction texte/lecteur. Em *L'effet-personnage dans le roman*, 25–39. Presses universitaires de France.

Labatut, Vincent & Xavier Bost. 2019. Extraction and analysis of fictional character networks: A survey. *ACM Computing Surveys* 52(5). 1–40. 10.1145/3344548.

Langfeldt, Marcia Caetano, Emanoel Pires, Rebeca Schumacher Fuão & Ricardo Gaiotto. 2021. Considerações sobre a personagem literária. https://www.linguateca.pt/aval_conjunta/dip/personagem.html.

Lukács, Georg. 2000 [1916]. *A teoria do romance*. Duas Cidades/Editora 34. Traduzido por José Marcos Mariani de Macedo.

Moretti, Franco. 2005. *Graphs, maps, trees: Abstract models for a literary history*. Verso.

Parenteau, Véronique. 1998. L'analyse de textes littéraires assistée par ordinateur: une introduction. *Cursus* 4(1). em linha.

Piper, Andrew, Mark Algee-Hewitt, Koustuv Sinha, Derek Ruths & Hardik Vala. 2017. Studying literary characters and character networks. Em *Digital Humanities*, 119–121.

Santos, Diana, Cristina Mota, Emanoel Pires, Marcia Caetano Langfeldt, Rebeca Schumacher Fuão & Roberto Willrich. 2023. DIP - desafio de identificação de personagens: objectivo, organização, recursos e resultados. *Linguamática* 15(1). 3–30. 10.21814/lm.15.1.399.

Santos, Diana, Roberto Willrich, Marcia Langfeldt, Ricardo Gaiotto de Moraes, Cristina Mota, Emanoel Pires, Rebeca Schumacher & Paulo Silva Pereira. 2022. Identifying literary characters in Portuguese: Challenges of an international shared task. Em *Computational Processing of the Portuguese Language (PROPOR)*, 413–419.