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

DIP - Desafio de Identificação de Personagens: objectivo, organização, recursos e resultados

DIP - Character Identification Challenge: goal, setup, resources and results

Diana Santos
Linguateca & ILOS, UiO

Cristina Mota
INESC-ID & Linguateca

Emanoel Pires
UEMA/UFPI

Marcia Langfeldt

Rebeca Schumacher Fuão

Roberto Willrich
Universidade Federal de Santa Catarina

Resumo

Este artigo apresenta o Desafio de Identificação de Personagens (DIP) em profundidade. Documenta a sua motivação, as escolhas feitas, o desenrolar do processo de organização, a avaliação conjunta, e os resultados que podemos mostrar, assim como os recursos compilados e que são públicos. Relatamos o que aprendemos com a organização do DIP e o que aprendemos sobre a literatura em português. Por exemplo, nas obras do DIP, (1) o número de personagens femininas é muito inferior ao das personagens masculinas, (2) existem sempre algumas personagens referidas com nomes diferentes na mesma obra, (3) a profissão mais mencionada é a de padre, (4) há mais referência a pais do que a mães, e (5) os diminutivos são bastante frequentes.

Palavras chave

avaliação conjunta, literatura lusófona, identificação de personagens

Abstract

This paper presents in-depth DIP, the character identification challenge in Portuguese. It aims to fully document its motivation, the choices taken, the organization process, the evaluation contest proper, and the results achieved. It also presents the public resources created by DIP. We report on what we have learned with DIP's organization, and what we learned about lusophone literature. For example, in the works analysed by DIP (1) the number of feminine characters is way less than masculine characters, (2) every work has some character with more than a name, (3) the most frequent profession is priest, (4) the works refer more to fathers than to mothers, and (5) diminutives are pretty frequent as character names.

Keywords

evaluation contest, lusophone literature, character identification

1. Introdução

1.1. Motivação

A ideia de organizar uma avaliação conjunta no âmbito da leitura distante surgiu na senda da organização do Primeiro Encontro de Leitura Distante em Português, que teve lugar em Oslo em outubro-novembro de 2019, relatado em Santos et al. (2020a). Nessa altura, estabeleceu-se uma relação informal entre o Núcleo de Pesquisas em Informática, Literatura e Linguística (NUPILL)[1] e a Linguateca[2], veja-se Santos (2022b), em que uma das vertentes de colaboração futura seria a organização dessa mesma avaliação conjunta, dada a experiência que a Linguateca tinha na organização de avaliações conjuntas para o português.

De facto, existia já uma longa história de avaliações conjuntas organizadas pela Linguateca, iniciadas no EPAV (Encontro Preparatório de Avaliação Conjunta em Processamento Computacional do Português) em 2002[3], e documentadas em Santos (2022a). O NUPILL, por seu lado, é um dos centros dedicados à literatura computacional lusófona mais antigos no mundo, com quase 30 anos de atividades na área da literatura e da computação.

[1] https://nupill.ufsc.br/
[2] https://www.linguateca.pt/
[3] https://www.linguateca.pt/aval_conjunta/Faro2002/

DOI: 10.21814/lm.15.1.399

Além dos corpos literários da Linguateca (Vercial[4] e OBras (Santos et al., 2018)), a participação da primeira autora na ação COST "Distant reading for European literature history"[5] levou à criação de mais recursos para a leitura distante em português, nomeadamente o corpo NOBRE[6] e a coleção ELTeC-por Schöch et al. (2021). Concomitantemente, houve um aumento significativo do número de obras brasileiras digitalizadas por ocasião da bolsa de pós-doutorado de Emanoel Pires na Universidade de Oslo no período 2020–2022.

Foi, portanto, considerado possível iniciar um trabalho na leitura distante em português, incitando grupos a desenvolver sistemas que pudessem contribuir para esse objetivo. A escolha recaiu sobre a identificação de personagens, por nos parecer, de todas as possíveis demandas em leitura distante, a mais simples de concretizar e também de tornar visível a um público não especializado. Além disso, pensávamos poder contar com a existência de vários sistemas de reconhecimento de entidades mencionadas que talvez pudessem ser adaptados a esta nova tarefa.

1.2. A noção de personagem

Houve um rápido consenso em relação à escolha da identificação e caracterização de personagens como a tarefa mais atraente e exequível. Aliás, já tinham sido manualmente anotadas personagens no projeto AC/DC em relação a alguns livros, para criar redes de personagens (Santos & Freitas, 2019) e sabíamos que seria muito prático se pudéssemos executar essa tarefa automaticamente para muitas obras.

Contudo, a noção de personagem veio a mostrar-se mais complexa do que imaginávamos à partida, por várias razões:

- Em primeiro lugar, embora os estudiosos de literatura concordem geralmente na atribuição das etiquetas personagem principal, secundária e figurantes em relação a obras estudadas, não conhecemos uma metodologia geral, operacionalizável, e consensual que dada uma obra qualquer, ainda não estudada, produza essas decisões sem ruído. Por isso, decidimos identificar todas as personagens, e deixar para depois, se necessário, fazer esse tipo de distinção.
- Em segundo lugar, não parece sequer haver um consenso sobre a diferença entre pessoas mencionadas numa obra, e personagens dessa obra. Isso é tanto mais complicado no caso de, por exemplo, romances históricos que romanceiam a vida e as ações de figuras históricas, ou que as mencionam de passagem para situar a época em que o romance se passa. Alguns teóricos da literatura consideram mesmo que qualquer menção a uma pessoa dentro de uma obra a torna personagem dessa obra.

Nós seguimos a seguinte definição no DIP: as personagens em que estamos interessados são fictícias, ou são pessoas históricas que participam/fazem avançar o enredo numa dada obra. Referências a outras pessoas fictícias através de intertextualidade, ou a pessoas históricas que não participam no enredo, não devem ser consideradas como personagens da obra em questão.

Adotamos a distinção entre personagens e referências a outras pessoas (fictícias ou não) por duas razões. Uma, por nos parecer ser concetualmente distinto o estatuto dos dois tipos de entidades, e nos interessar sobretudo as entidades próprias de uma obra, por contraposição àquelas mencionadas por muitas.[7] E a segunda razão foi para diferenciar esta tarefa do simples reconhecimento de pessoas em texto, que é um subconjunto do problema do reconhecimento de entidades mencionadas.[8]

Depois de fixarmos o que era uma personagem para o DIP e de explicarmos o interesse de as analisar para os estudos literários em Langfeldt et al. (2021), tivemos de definir o conjunto de características que pretendíamos que os sistemas identificassem sobre essas personagens.

Enquanto a questão da correferência sempre esteve decidida, visto que sabíamos que uma personagem pode e costuma ter vários nomes e/ou formas pela qual é tratada — e esta questão dos diferentes nomes já tem sido objeto de investigação em vários outros trabalhos, veja-se Santos & Freitas (2019) ou Krug et al. (2018) —, foi preciso tomar uma decisão em relação a que outras características de uma personagem nós gostaríamos que os sistemas nos facultassem.

Uma característica razoavelmente consensual (no sentido de ser estudada por muitos inves-

[4] https://www.linguateca.pt/acesso/corpus.php?corpus=VERCIAL

[5] https://www.distant-reading.net/

[6] https://www.linguateca.pt/acesso/corpus.php?corpus=NOBRE

[7] Vimos mais tarde que uma distinção semelhante, entre entidades *plot-internal* e *plot-external*, também tinha sido feita no projeto Namescape (de Does et al., 2017).

[8] Estamos a afirmar que a tarefa de reconhecimento de personagens é mais complexa do que a do reconhecimento de pessoas, porque além de ter de separar os casos de lugares e organizações ainda precisa de distinguir, e rejeitar, pessoas que não sejam personagens.

tigadores de literatura) é o género da personagem. Mas convém esclarecer que o género de uma personagem literária e o género ou o sexo de uma pessoa no mundo real são conceitos distintos. O DIP limitou-se a identificar a representação textual do género, ou seja, o modo como a personagem é representada numa obra literária, mas não os traços de personalidade, os comportamentos, as ações e os estereótipos associados ao género. De facto, o género de uma personagem literária é uma construção determinada pela cultura e o período histórico no qual a obra está inserida, bem como pela intenção do autor. No DIP, o género de uma personagem pode ser masculino, feminino, ambos, ou desconhecido, mas veja-se mais a este respeito em Pires et al. (2023).

Outra tarefa que nos pareceu interessante para estudos históricos e do romance, sugerida por um estudo preliminar feito no âmbito da ação COST já citada, Santos et al. (2020b), foi determinar quais as profissões e ocupações mencionadas nos livros. Mas à medida que fomos operacionalizando a tarefa, como está também descrito em Pires et al. (2023), a definição foi-se revelando mais complicada. Para algumas personagens é o seu título nobiliárquico que as define, como *conde*, para outras é uma ocupação não remunerada, como *dona de casa* ou mesmo forçada, como *escravo*. E em português não há um termo que represente estas três formas de descrever a posição social ou ocupacional de uma pessoa, por isso optámos por usar a expressão "profissão, ocupação ou estatuto social", abreviada por POES.

Ao contrário do género, que em geral se mantém constante ao longo da obra — embora tenhamos trabalhado com um modelo em que pode mudar — a POES pode ser múltipla, e variada, ou seja, uma pessoa pode ser ao mesmo tempo médico e duque, ou passar de pastora de cabras para professora, e — o caso claramente mais frequente — transformar-se de estudante em profissional de uma dada área.

Finalmente, pensámos que seria interessante detetar relações familiares entre as personagens, talvez inspirados pelo trabalho feito sobre relações familiares no Dicionário Histórico-Biográfico Brasileiro (DHBB) (Higuchi et al., 2019). Mas também a operacionalização desta escolha teve várias consequências, descritas em Mota & Santos (2023). Foi sobretudo muito discutido que relações entre as personagens faria sentido tentar identificar. Durante a fase inicial, algumas pessoas criticaram que nos dedicássemos simplesmente a relações "oficiais"[9] e não a outras como amigo, amante, concubina, namorado ou admirador. A principal razão da não incorporação destas (importantes) relações foi a de que não eram estáticas: muitas vezes o próprio enredo é dedicado a um namoro, os amigos podem deixar de o ser, assim como os admiradores.

Essa foi, aliás, também a razão por que decidimos, nesta primeira edição pelo menos, não pedir os lugares onde a ação decorria, nem a estrutura temporal da obra, ambos temas que alguns interessados no DIP queriam tentar obter.

A escolha do que pedimos aos sistemas para tentar identificar automaticamente nos textos literários foi uma tentativa de equilíbrio entre algo suficientemente interessante mas não demasiado difícil. Não é garantido que o tenhamos conseguido, mas essa foi uma preocupação que nos norteou.[10]

2. Organização do DIP

Para organizar uma avaliação conjunta é preciso, além de escolher inicialmente uma tarefa, divulgá-la pelo maior número de pessoas e grupos, para que seja uma avaliação verdadeiramente conjunta. Em seguida, é preciso obter algum consenso sobre o calendário e sobre os recursos a serem desenvolvidos, assim como qual o formato exato da avaliação. E é preciso documentar todas as escolhas e ter um lugar na rede que os interessados possam consultar sempre que precisem.[11]

Começamos por apresentar o calendário final, na Tabela 1. Os principais eventos são descritos nas seções que se seguem.

2.1. Especificação da tarefa

A primeira atividade da organização do DIP foi especificar a tarefa a ser realizada no desafio. Foi definido que seriam tornados públicos 100 textos em formato de texto, na codificação UTF-8, e 100 textos em pdf. Uma vez disponibilizadas as obras, os sistemas participantes do DIP teriam 48 horas, o período do desafio propriamente dito, para produzir toda a informação sobre cada obra.

Como parte da definição da tarefa, a organização definiu como as obras iriam ser dispo-

[9] No sentido de legais, verificáveis num cartório.

[10] Contudo, não podemos deixar de concordar com os comentários de Luísa Coheur, Alexandre Rademaker e Roberlei Alves Bertucci de que a justificação de que as relações familiares não de sangue não é fixa, e que aceitamos várias profissões — porque não aceitar lugares também? Ou seja, a justificação das nossas escolhas não é muito coerente.

[11] DIP criámos pois `https://www.linguateca.pt/DIP`.

Data	Atividade
10/2021	Início da organização
05/11/2021	Anúncio público
29/11/2021	Encontro virtual
29/11/2021 a 15/03/2022	Ensaio: participantes e interessados anotam dois novos textos
16/03/2022	Encontro virtual sobre o ensaio
15-17/09/2022	Desafio
01/10/2022	Resultados publicados
21/11/2022	Encontro do DIP

Tabela 1: Calendário do DIP

nibilizadas e qual a sintaxe de representação dos dados extraídos das obras. As obras foram distribuídas com o nome obra$_i$.txt ou obra$_j$.pdf, onde i e j são números inteiros, variando respetivamente de 0 a 99 e de 100 a 199. O resultado da análise da obra deveria ser representado na forma de dois ficheiros CSV: personagens.csv e relacoes.csv. O primeiro deveria indicar as personagens, uma por linha, seguindo a seguinte sintaxe: $\{i, k, correferencias, genero, POES\}$, onde i é o identificador da obra, k é o identificador da personagem na obra, $correferencias$ é a relação de menções no texto referenciando a personagem, $genero$ indica seu género (M, F ou A para ambos), e $POES$ indica o conjunto de profissão/ocupação/estatuto social da personagem. Dois exemplos deste formato encontram-se na Tabela 2.

O arquivo relacoes.csv deve incluir todas as relações familiares entre personagens utilizando a seguinte sintaxe: $\{i, s, relacao, o\}$, onde i é o identificador da obra, e s e o identificam as personagens (mesmos identificadores em personagens.csv) que têm a relação $relacao$. Exemplos deste formato encontram-se na Tabela 3.

Os participantes deveriam usar o sistema *EasyChair*[12] para enviar o resultado dos seus sistemas, num ficheiro comprimido `zip`. Para evitar problemas, deveriam testar todo este processo (formato e EasyChair) no ensaio.

2.2. Ensaio

Para familiarizar os participantes com o que lhes era pedido, e também com as várias escolhas que teriam de fazer, pedimos a todos os interessados para anotar manualmente, após leitura próxima, mais dois textos, e discutirmos os resultados em conjunto, o que deu origem a várias precisões e melhores diretivas, assim como a mais dois ficheiros de exemplo.

[12] https://easychair.org/

Como já mencionado, esses dois ficheiros também teriam de ser enviados pelo EasyChair para testar o envio.

O processo do ensaio e a discussão no encontro (remoto) foram também muito importantes para fixar a tarefa de construção da coleção dourada, que foi a atividade mais trabalhosa que a organização levou a cabo: ler mais trinta e oito obras de fio a pavio e produzir toda a informação exigida pelo DIP.

2.3. Avaliação

Outra incumbência da organização foi sugerir medidas de avaliação para a tarefa do DIP, e implementar e testar os programas que as calculassem, antes da própria avaliação conjunta, para todos os participantes saberem como iam ser avaliados.

O resultado deste trabalho, e as medidas a que chegámos, estão descritos pormenorizadamente em Willrich & Santos (2023).

2.4. A compilação da coleção do DIP

Finalmente, outra tarefa extremamente importante foi fixar quais as obras que seriam distribuídas no DIP (e, dentre estas, quais as que fariam parte da coleção dourada). Isso será o tema da próxima secção.

3. Recursos criados

Talvez o mais importante resultado de uma avaliação conjunta sejam os recursos criados no seu âmbito, além da especificação da tarefa e da medição do seu sucesso.

Além da descrição objetiva dos dados que pusemos à disposição da comunidade, parece-nos importante documentar a sua construção e as várias decisões que tivemos de tomar.

i,k,correferencias,genero,POES
021,0,Margarida|Guida|Guida dos Meadas|Margaridinha,F,cabreira|professora
021,1,Clara|Clarinha|Clarita|Clarita dos Meadas,F,
021,2,Daniel|Sr. Daniel|Danielzinho|Daniel do Dornas|Danielzinho do Dornas,M,estudante|médico
021,3,Francisca|Chica|Chica da Esquina,F,
021,4,Joana|Sra. Joana,F,criada

Tabela 2: Exemplo de como a informação deveria estar codificada no ficheiro personagens.csv

i,s,relacao,o
021,0,irmã,1
021,2,irmão,6
021,5,pai,2
021,9,marido,10
021,9,pai,3

Tabela 3: Exemplo de como a informação deveria estar codificada no ficheiro relacoes.csv

3.1. A que textos/obras podíamos recorrer

Em primeiro lugar, e como já descrito em várias outras ocasiões (Schöch et al., 2021), não existe infelizmente um manancial de obras em texto em português que possa ser imediatamente usado para o seu processamento, ao contrário de outras línguas. Essa foi, aliás, uma das razões que nos levou a pensar na vertente "tratamento do pdf" no DIP, visto que existem, ou imaginamos que existam, muito mais textos simplesmente digitalizados em PDF como imagem, como é o caso dos acessíveis através do Google Books[13] ou do Internet Archive.[14]

Para sermos mais explícitos: o reconhecimento ótico de carateres associado à maior parte das obras em domínio público digitalizadas em língua portuguesa, por exemplo as existentes no Internet Archive, produzidas por iniciativas de digitalização nos Estados Unidos, é de qualidade tão má que se pode considerar que a simples digitação manual do livro levaria o mesmo tempo que a revisão do que foi reconhecido automaticamente.[15] Digitalizações mais modernas, e/ou feitas por instituições com conhecimento (e ferramentas) mais adequadas para a língua portuguesa, como as das bibliotecas nacionais de Portugal e do Brasil, por exemplo, produzem objetos digitais muito mais fiáveis, mas mesmo assim (ainda) não perfeitos. Seja como for, se o objetivo último é fazer leitura distante sobre milhões de obras, não podemos esperar que estas sejam revistas e, por isso, era importante levar sistemas a trabalhar com PDF.

3.2. Critérios para a escolha da coleção do DIP

Para escolher os 100 textos que fariam parte da coleção em formato de texto, socorremo-nos da lista de romances e novelas acessíveis na Literateca (Santos, 2019), das quais escolhemos uma obra de cada autor. Havia exatamente 50 autores brasileiros, e um pouco mais portugueses, mas não fizemos grandes reflexões sobre o assunto, exceto que esgotámos as autoras (visto que sabíamos que havia poucas).

A maior parte dos 100 textos em formato PDF foram selecionados dentre os 460 arquivos com romances ou novelas em domínio público disponibilizados na Biblioteca Digital de Literaturas em Língua Portuguesa (BDLP[16]) do NUPILL. Neste caso havia muito mais obras brasileiras em PDF do que portuguesas — provavelmente porque o trabalho da BDLP, em andamento, é feito no Brasil —, e tivemos dificuldade em arranjar 50 obras em PDF de autores portugueses que ainda não constassem da coleção de texto. Para resolver esse problema, adicionámos obras diferentes de autores que já estavam na coleção de texto.

No caso dos autores brasileiros, como havia muito mais do que cinquenta novos autores (em relação aos incluídos na coleção de texto), além de escolher todas as autoras, utilizámos como critério o tentar maximizar a variação do estilo e da data, embora não conhecêssemos a maioria das obras em questão.

A lista de obras contidas na coleção DIP, em txt e em pdf, está no apêndice A. Na Figura 1 está a distribuição temporal destas obras. De referir que incluímos as obras usadas como exemplo e estudadas no ensaio.

[13] https://books.google.com/

[14] https://archive.org/

[15] Não temos uma demonstração desta afirmação, que corresponde simplesmente à nossa experiência com alguns textos.

[16] https://literaturabrasileira.ufsc.br/

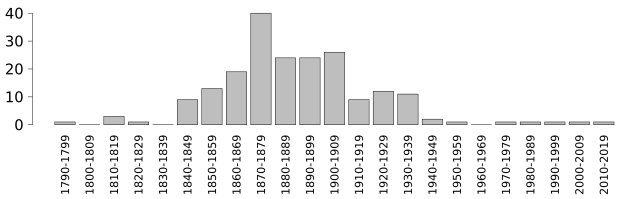

Figura 1: A distribuição das obras da coleção do DIP por década

Podemos constatar que existem 25 obras escritas por autoras e 178 por autores. O número de autores diferentes portugueses é de 88[17] e o número de autores diferentes brasileiros é de 96.[18]

3.3. Critérios para a escolha da coleção dourada

A coleção dourada é o subconjunto da coleção DIP para a qual compilámos os valores que pretendíamos que os sistemas obtivessem, 21 obras em texto e 17 obras em pdf, das 100 de cada.

A escolha destas obras não seguiu critérios pensados do início. Pelo contrário, no caso do texto aconteceu ao mesmo tempo que escolhemos as obras para a coleção do DIP e, no caso do pdf, foi eminentemente escolhida por questões práticas, nomeadamente já termos a obra em formato PDF em nosso poder, e escolhemos a maioria das outras obras em PDF mais tarde.

O único critério que tentámos seguir, para maximizar a variação da coleção, foi o de não termos mais do que uma obra por autor – critério esse que foi infelizmente desobedecido, por lapso, no caso de Carlos Pinto de Almeida, com duas obras na coleção dourada relativa aos pdf.[19]

Embora possa parecer que demos pouca importância à escolha da coleção dourada, é preciso salientar duas coisas: não fazíamos ideia da distribuição de personagens na literatura em geral e, portanto, não tínhamos muitas características sobre as quais diversificar; e esperávamos que, mais tarde ou mais cedo, obteríamos, a partir do resultado dos sistemas, informação para todas as 100 obras, por isso não era assim muito importante por quais começar.[20]

[17] Os autores portugueses com duas obras são Alberto Pimentel, Alice Pestana, Ana de Castro Osório, Ana Plácido, António Francisco Barata, Arnaldo Gama, Camilo Castelo Branco, Carlos Malheiro Dias, Carlos Pinto de Almeida, Eça de Queirós, Francisco Gomes de Amorim e Virgínia de Castro e Almeida

[18] Os autores brasileiros com duas obras são Aluísio Azevedo, Bruno Seabra, José da Rocha Leão e Júlio Ribeiro

[19] As obras são *A filha do emir* e *Os homens da cruz vermelha*.

[20] De facto, neste momento — junho de 2023 — encontra-se em curso a leitura próxima de mais obras da coleção de texto, alargando assim os recursos produzidos pelo DIP.

3.4. Decisões na anotação da coleção dourada

Muito mais trabalho e discussão (no seio da organização, e com os participantes no ensaio) levou a própria criação do conteúdo da coleção dourada, indicando que haveria muitas questões mais finas sobre as quais teríamos de chegar a um consenso de forma a poder documentar o que os sistemas deveriam fazer. Tal como no HAREM, em que tivemos de escrever páginas e páginas de diretivas (veja-se Cardoso & Santos (2007)), na questão das personagens houve muitas decisões que precisámos de tomar.

Embora não consigamos trazer aqui todas, a seleção que apresentamos dará uma ideia de que a operacionalização de uma tarefa computacional (ou, seja como for, exaustiva) requer a consideração de muitas questões, muitas delas não necessariamente intuitivas.

3.4.1. Narrador como personagem

Uma das primeiras coisas sobre as quais nos tivemos de pronunciar foi sobre personagens sem nome. Embora importantes para a análise literária, não nos pareceu possível arranjar uma forma natural de as incorporar no DIP.

Esse é o caso de narradores autodiegéticos ou homodiegéticos, na primeira pessoa, que não sejam nunca tratados pelo nome. Nesse caso, não aparecem nas listas das personagens do DIP.

3.4.2. Formas de indicar um parentesco

Há muitíssimas formas diferentes de indicar um parentesco, sobretudo no caso de um casamento (por exemplo: *mulher*, *esposa*, *a pessoa com quem casei*, *a minha patroa*, *a sua cara-metade*, ou *casaram-se*, *deram o nó*, *uniram os trapinhos*). Decidimos que o trabalho de as identificar e normalizar ficaria do lado dos sistemas participantes, e fixámos os nomes das relações de parentesco, nomeadamente: *mãe, pai, filho/a, neto/a, avó, avô, irmã/o, cunhado/a, primo/a, tio/a, sobrinho/a, bisavó, bisavô, bisneto/a, nora, genro, sogro/a, mulher, marido, padrinho, madrinha, compadre, comadre, afilhado, afilhada.*

Alguns casos são o que poderíamos chamar de "parentesco social", e não laços de sangue. São os casos de *madrinha/padrinho*, *comadre/compadre* e *afilhado/afilhada*, e os casos de *madrasta/padrasto* e *enteado/enteada*. Por lapso, não colocámos estes últimos na lista.[21]

[21] Viemos a observar mais tarde que a relação de *madrasta* ou *padrasto* aparecia em sete das obras da coleção

Considerámos que as palavras *noivo* e *noiva* eram suficientemente próximas de formalização de um casamento para serem identificadas, ao contrário de *namorado*, *conversado*[22], etc.

Também considerámos que era relevante a "relação" de *viúvo* ou *viúva*, e que esta implicava uma situação marital diferente de *casado*.

Estabelecemos ainda que todas as relações que ocorressem entre as mesmas duas personagens durante a obra deviam ser encontradas, por isso `A noivo B` , `A marido B` e `A viúvo B` podiam ser a resposta certa, se a vida toda de A fosse contada na obra.

Mas é importante indicar que não requeríamos que uma relação e a sua inversa fossem indicadas, nem pelos sistemas participantes, nem na coleção dourada. O cálculo das relações inversas é feito durante a avaliação.

3.4.3. Que caracterização profissional escolher

Ao contrário das relações familiares, considerámos impossível normalizar e/ou prever de antemão tudo o que seria encontrado nas obras do DIP, e decidimos que a profissão, estatuto social e ocupação retornada deveria ser a encontrada na obra.

Mesmo assim, sugerimos que alguma reformulação teria de ser feita em casos como "despediram todos os cozinheiros, excepto a Maria". Num caso como esse, os sistemas deveriam colocar *cozinheira* na profissão da Maria.

Tal como em relação ao parentesco, todas as ocupações mencionadas na obra deveriam ser indicadas, por isso uma personagem poderia ter mais do que uma POES.

Se a personagem era descrita como ex-profissional, seria isso que constaria. Em princípio, e se descrito por ambas em épocas diferentes da obra, uma personagem podia ser por exemplo *professor* e *ex-professor*.

No caso de *aposentado* ou *reformado*, estipulámos que, se fosse antecedido pela profissão, ambas deviam ocorrer. Ou seja, *juiz aposentado*, *cozinheiro reformado*.

Quando a palavra *herdeiro* ou *herdeira* não aparecesse relacionada com outros de quem herda, e significasse uma pessoa rica porque herdou, também deveria ser considerada uma ocupação (ou falta dela).

3.4.4. Animais com nome

Se existissem animais com nome nas obras, eles deveriam ser considerados como personagens. No local do POES, deveria ser indicado o tipo de animal: cão, cavalo, etc.

3.4.5. Personagens que não chegam a existir

Este parece um caso estranho, mas não é tão raro como se poderia pensar. Por exemplo, considere-se a frase *imaginou que mais tarde teriam um filho a que chamariam Álvaro*. Nesse caso, determinámos que essa personagem deveria ser marcada.

3.4.6. Personagens provindas da loucura ou alucinação

Mais um caso que talvez não se imaginasse, sem ter lido de fio a pavio várias obras, é aquele em que as personagens deliram e pensam que eles e os outros são outras pessoas, como é o caso no *Quincas Borba* de Machado de Assis, em que a personagem principal enlouquece. Decidimos que, se tiverem nome, devem ser considerados como outros nomes (co-identificação) dessas mesmas personagens.

Também quando as profissões se referem a jogos de crianças, como *Laurita cozinheira* em *Amar, verbo intransitivo* de Mário de Andrade, consideramos que devem ser marcadas.

3.4.7. Nomes com partes em minúsculas

Em geral os nomes próprios em português são em maiúsculas, mas há um caso especial que é o das alcunhas (em português de Portugal) ou apelidos (em português do Brasil) em que só uma parte é em maiúscula, como em *João das pantorrinhas*. Nesse caso, e embora isso corresponda a uma exigência muito mais elevada, decidimos que seria necessário que o sistema identificasse o nome todo e não só a parte em maiúsculas.

3.4.8. Títulos de nobreza ou cargos

Finalmente, uma decisão muito importante e da qual mais tarde nos arrependemos, mas já não podíamos voltar atrás, sob pena de ter de refazer a leitura de várias obras já prontas, foi a de considerar que uma personagem mencionada apenas pelo seu título não era para identificar. E, da mesma forma, não era para identificar o título se fosse chamado por ele.

A justificação para esta decisão era de que poderia haver várias pessoas diferentes todas *conde*

dourada de texto, o que mostra que teria sido importante também identificar estes casos, já presentes, aliás, numa das obras de exemplo.

[22] Forma antiga de dizer namorado.

de Oeiras, ou *marquês da Palma*, e que o título não seria suficiente para as distinguir. Mas o que é certo é que verificámos mais tarde que, em muitas obras, sobretudo romances históricos, personagens extremamente importantes para o enredo eram assim descritas ao longo de toda a obra.

3.4.9. Outras microdecisões

Outras decisões referem-se a situações pontuais, mas que também não eram evidentes. Por exemplo, decidimos não marcar quando um nome próprio é chamado como um insulto, como no passo seguinte:

> O irrequieto arcebispo foi pôr cerco a Simancas; mas do alto das muralhas da velha cidade os sitiados escarneceram-n'o, chamando-lhe D. Opas; – o que significava compará-lo com o typo mais repugnante dos homens conhecidos por traidores (em *Pero da Covilhã*, de Zeferino Norberto Gonçalves Brandão)

Sobretudo em relação a profissões ou ocupações, muitas microdecisões tiveram de ser tomadas. Foi especialmente difícil decidir em relação a POES que têm um significado negativo ou usadas em contextos não tradicionais. Por exemplo, não considerámos *bohemio*, *fradalhão* ou *capataz de uma turma de vadios* como ocupações, mas marcámos *agiota* e *prostituta* em casos que poderiam ser interpretados como subjetivos.

Descrições como *dono de barco* ou *hospedeiro* (dono de hospedaria), em que não considerámos o primeiro como POES, mas sim o segundo, mostram como a fronteira é ténue entre aquilo que se pode ou não considerar uma atividade profissional, ocupação ou estatuto social. A explicação é que o dono de um barco — pelo menos no romance em questão — não necessita de estar associado às viagens marítimas, enquanto se pressupõe que o dono da hospedaria está lá a receber os hóspedes, e ocupa a maioria do seu tempo nessa atividade.

Finalmente, personagens no plural, tal como *os Pereiras*, ou *as manas Madureira*, não foram consideradas.

3.5. Caracterização da coleção dourada

Após a marcação das personagens em 43 obras, as 40 da coleção dourada e as 3 de exemplo, podemos dar uma primeira aproximação do assunto na literatura lusófona.[23]

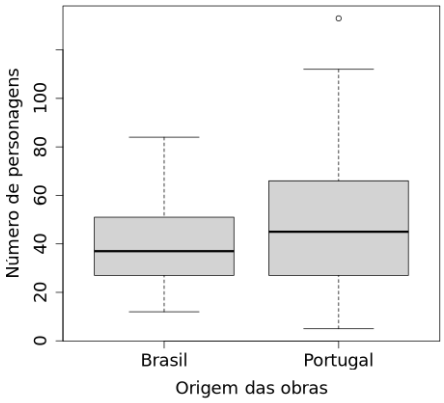

Figura 2: O número de personagens por obra, nas 43 obras brasileiras e portuguesas a que atribuímos uma solução

Na Figura 2, vemos o número de personagens nas 43 obras, por literatura. Este número variava entre 5 para a novela *A vinha* de Ana de Castro e Osório e 112 para o romance histórico *Pero da Covilhã* de Zeferino Norberto Gonçalves Brandão.

Observa-se que existe mais variação nas obras portuguesas, que também têm em média um número um pouco mais elevado de personagens.

Apenas para as 24 obras em texto para as quais temos a solução, averiguamos a relação entre o tamanho da obra e o número de personagens.

Na Figura 3 apresentamos o panorama relacionado com o tamanho em número de palavras das obras. Vemos que em geral existem mais personagens em obras mais longas, mas que existem algumas obras não muito longas com muitas personagens, e são todas romances históricos.

A Figura 4 apresenta a mesma questão de outra forma, indicando a densidade relativa de personagens nas 24 obras das quais temos o tamanho em palavras.

3.5.1. O género na coleção dourada

Se observarmos a panorâmica do género na coleção dourada, apresentada na Figura 5, observamos que em todas as obras — exceto uma, *A vinha*, uma novela de Ana de Castro Osório, com apenas 5 personagens, 3 mulheres — a maioria das personagens em cada romance são masculinas.

É preciso lembrar que estamos a medir a existência de todas as pessoas mencionadas na

[23] Convém referir, como discutiremos na Secção 4.2, que a marcação das personagens nas 21 obras da coleção de texto beneficiou do cotejo com os resultados do sistema participante.

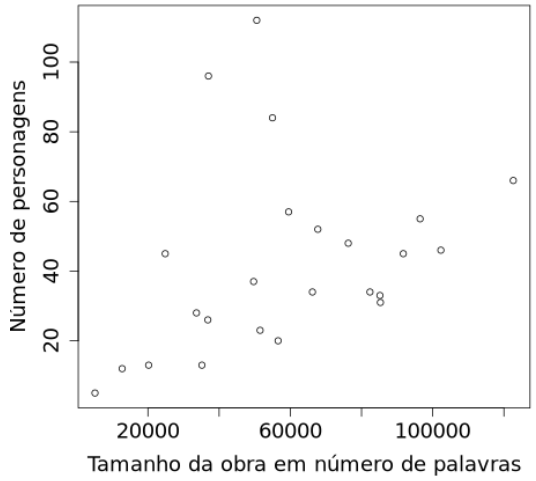

Figura 3: As obras em termos de tamanho e de número de personagens

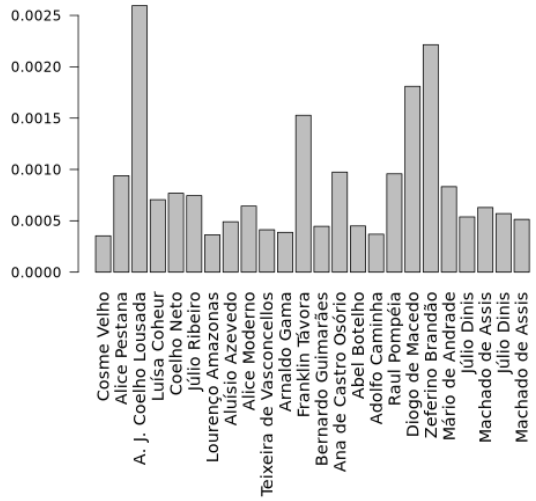

Figura 4: A densidade relativa do número de personagens por obra

obra por nome, não apenas as personagens principais. De facto, na novela que acabámos de mencionar, a personagem principal é um homem. Por isso a maioria de homens pode simplesmente significar que há mais homens na esfera pública na sociedade descrita nas obras, que as mulheres têm menos cargos, estão presentes sobretudo no âmbito privado.

Ao analisar o gênero dos personagens na literatura lusófona e sua profissão, é possível identificar desigualdades de género que refletem a realidade histórica e social das sociedades em que as obras foram escritas. Na maioria das obras, as personagens masculinas ocupam mais cargos públicos e têm mais visibilidade social do que as femininas, que muitas vezes são retratadas em papéis secundários e na esfera privada.

Esta conclusão é corroborada pela proporção maior de personagens femininas sem características de profissão, ocupação ou estatuto social em comparação com as personagens masculinas, como veremos a seguir.

No que se refere às profissões, estatuto social e ocupações, na coleção dourada total (incluindo os textos de exemplo), há 1944 personagens, das quais 942 (48,5%) não têm este tipo de caracterização.

No caso das personagens masculinas, 633 em 1504 não têm profissão, ocupação ou estatuto social, ou seja, 42,1%. No caso das personagens femininas, o mesmo ocorre para 309 em 440, ou seja, 70,2%.

Seja como for, para estudar a importância das mulheres nas obras em si, teríamos de primeiro identificar as personagens principais, e refazer as contas baseadas nestas.

3.5.2. O estatuto profissional na CD

Quanto às 409 diferentes POES identificadas (convém relembrar que muitas delas são apenas variações ortográficas da mesma, como *abadessa*, *abbadeça* ou morfológicas, *abade*), as mais frequentes foram *padre* (55), *general* (42), *escravo* (38) e *estudante* (37). Se adicionarmos a variação *escrava* (12) — *estudante* refere-se aos dois géneros, e não existe uma profissão equivalente a *padre* para mulheres[24] —, obteríamos 50 casos, portanto o segundo lugar.

Separando entre as duas literaturas, o panorama é surpreendentemente semelhante: Há 266 POES diferentes na literatura portuguesa, e 258 na literatura brasileira. E os casos mais frequentes são, nas obras brasileiras, os mesmos que em geral: *padre* (27), *escravo* (18), *general* (17) e *estudante* (16). Adicionando *escrava* (6), teríamos exatamente a mesma ordem da totalidade das obras. Nas obras portuguesas, a situação é a mesma: *padre* (28), *general* (25), *estudante* (21) e *escravo* (20), mas adicionando *escrava* (6) este estatuto social passa para segundo lugar.

A importância das profissões religiosas na literatura lusófona já tinha sido discutida por Santos (2022c) em relação ao número de vezes que as profissões eram mencionadas no texto literário, mas note-se que o estudo do DIP é diferente no sentido de contar as profissões das personagens — independentemente de serem mencionadas muitas vezes, serem personagens principais

[24] Não conhecemos casos de generais femininos nem sabemos qual a forma de mencionar, por isso podemos assumir, o que aliás se verificou nas obras lidas, que *general* se refere apenas a homens.

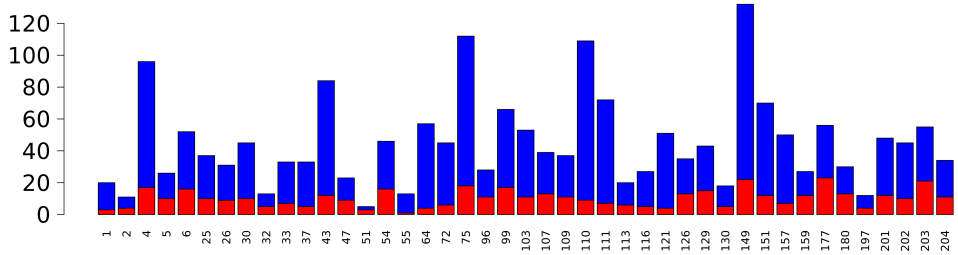

Figura 5: A distribuição de personagens por género na coleção dourada total: a vermelho, as personagens femininas; a azul, as masculinas

ou simples figurantes. Poder-se-ia imaginar que por exemplo padres seriam muitas vezes mencionados sem nome, e que portanto não necessariamente os dois estudos conduzissem aos mesmos resultados.

Por outro lado, quando uma personagem é padre e tem nome, será sempre descrita e mencionada por *padre X*, e geralmente tratada por *senhor padre X*, o que implica um grande acréscimo da palavra *padre* nos textos. Enquanto qualquer outra profissão, por exemplo médico, não seria usada em português para referir uma personagem que o fosse. (Seria tratada por *doutor Y*, não por *médico Y*).

Para uma análise mais detalhada das profissões no DIP, veja-se Pires et al. (2023).

3.5.3. As relações familiares na CD

Quanto às relações familiares nas 42 obras (visto que uma obra, como já mencionámos, não apresentava quaisquer relações entre as personagens), expandimos todas as relações passíveis de expansão, e obtivemos 810 relações familiares, apresentadas na Figura 6. Essas relações referiam-se a 777 personagens distintas.

Várias outras medidas e análises podem ser encontradas em Mota & Santos (2023), aqui apenas apontamos para a importância da relação pai, 106 casos (significativamente maior do que mãe, 64 casos) nas obras consideradas.[25]

3.5.4. Nomes diferentes para uma mesma personagem na CD

Quanto aos diferentes nomes pelos quais as personagens eram identificadas, que era um dos pressupostos do DIP, nomeadamente que seria um problema se não se identificasse a co-referência, as 43 obras que inspecionámos confirmaram indubitavelmente a necessidade de unir diferentes nomes. De facto, em todas as obras houve mais do que um nome para pelo menos uma personagem, como a Figura 7 ilustra.

Convém referir que a Figura 7 foi construída depois de termos juntado todas as possíveis formas de indicar a mesma coisa, apenas grafada diferentemente, ou seja, termos convertido por exemplo todos os casos de *sr.*, *snr.*, *Sr.*, e *Snr.* numa mesma forma, e corrigido o problema de acentos devido a reconhecimento ótico de carateres, como em *Álvaro*, *Alvaro* e *Àlvaro* referindo a mesma personagem numa dada obra. Se mantivéssemos as diferentes grafias da mesma forma de tratamento ao longo de uma obra, os números refletiriam ainda maior diversidade. Na Figura 8 mostramos a distribuição das personagens pelo número de formas diferentes por que são mencionadas, antes e depois da normalização.

3.5.5. As formas de tratamento na CD

Uma das características da língua portuguesa que também nos interessava explorar no DIP é a diversidade das formas de tratamento, tradicionalmente consideradas de grande complexidade na nossa língua. Aqui no DIP apenas poderíamos naturalmente identificar aquelas usadas em conjunção com um nome próprio, mas mesmo assim pudemos compilar alguns dados interessantes.

As formas de tratamento mais frequentes aparecem na Figura 9, em que as formas de senhor são claramente as mais usadas (com ou sem outras formas, como em *sr. dr.*). "redsenhor" corresponde a reduções de senhor, como *sôr*, ou *sinhô*, que não fazem parte da norma padrão mas que pretendem transmitir um certo dialeto ou socioleto. É interessante também a grande quantidade de *D.* (que corresponde a Dom ou Dona, por extenso). A palavra *sinhá* é apenas usada no Brasil, e não aparece frequentemente associada a um nome próprio nas obras que analisámos.

[25] Por outro lado pode-se também argumentar que, havendo mais personagens masculinas do que femininas, a relação de mãe é mais frequente para uma mulher nas obras (7,7%) do que a relação de pai para um homem (3,7%).

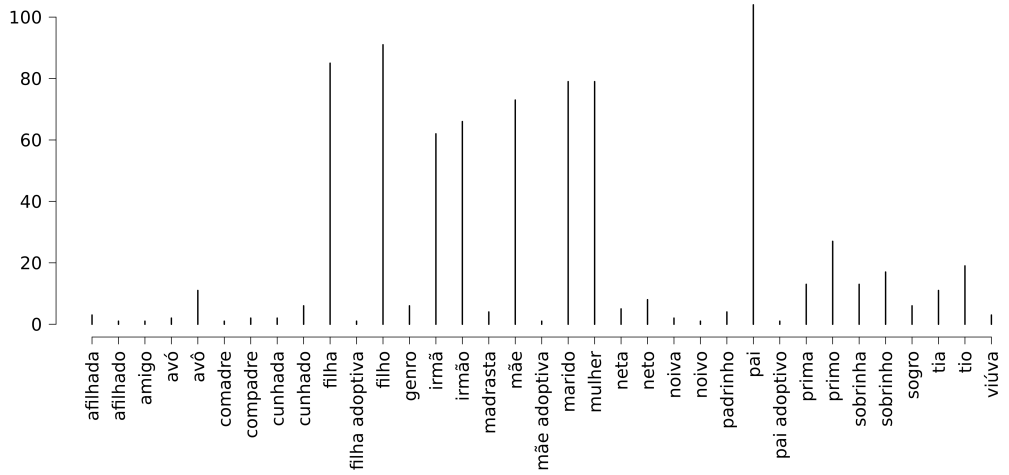

Figura 6: As relações familiares na coleção dourada total

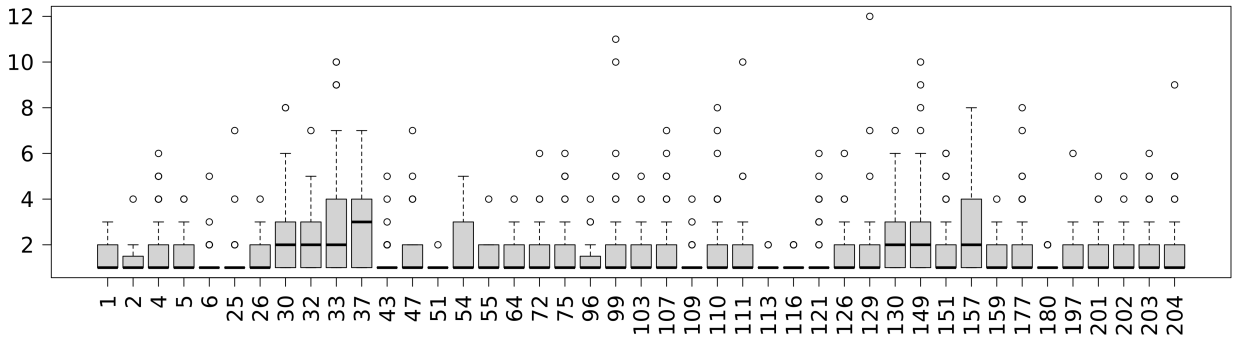

Figura 7: Número de nomes diferentes por personagem, depois da normalização, por obra

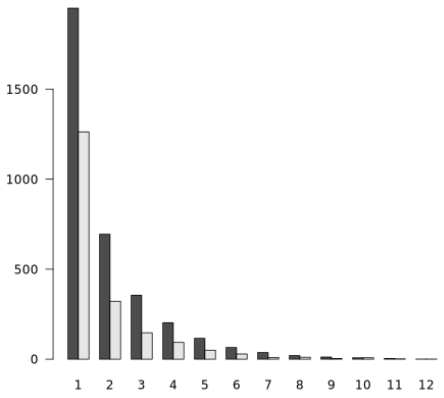

Figura 8: Número de nomes diferentes por personagem, antes e depois da normalização

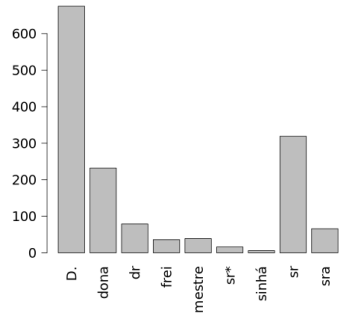

Figura 9: Formas de tratamento, depois da normalização. "sr*" significa formas não padrão do tratamento por senhor, como *seu*, *sinhô*, etc.

Outra vertente associada tanto a nomes diferentes como a formas de tratamento é o uso de diminutivos relativos a personagens. No artigo de apresentação do DIP à comunidade do processamento computacional do português (Santos et al., 2022), chegámos a apresentar a hipótese, baseada nas duas obras analisadas até aí, que o uso de diminutivos poderia diferir entre as duas literaturas, visto que em *As Pupilas do senhor Reitor* era usado para mostrar familiaridade e ternura para com as personagens, enquanto em *Dom Casmurro* era usado para distinguir mãe e filha com o mesmo nome.

Não nos debruçámos ainda sobre isso, por isso podemos apenas apresentar os valores quantitativos da existência de diminutivos nas obras da coleção dourada total, que apresenta 80 diminutivos, 44 masculinos e 36 femininos. 36 provêm da literatura de Portugal, e 44 da do Brasil. Ao todo são 57 diferentes.

Uma observação imediata é que em vários casos o diminutivo ocorre com outras formas de tratamento que em princípio indicariam mais distância e menos familiaridade, como *sr*, *sinhô*, *capitão*, *doutô*, etc. Em alguns casos, o diminutivo parece fazer parte da alcunha/apelido, como é o caso de *Miguel Mulatinho* ou *Mata Corcundinha*, ou ser aplicado ao apelido/sobrenome em vez de ao primeiro nome, como em *Mendonçazinho* ou *Pereirinha*. Parece-nos pois que a riqueza desta forma de tratamento merece ser mais explorada, para identificar o que está contido nestas formas de mencionar a personagem.

4. Resultados

4.1. Participação

Embora tivéssemos tido pelo menos quatro expressões de interesse, e tivéssemos mesmo adiado a data da avaliação conjunta para o semestre seguinte para ver se conseguíamos mais participantes, no fim apenas um participante enviou os resultados do seu sistema, o PALAVRAS-DIP, cuja descrição pode ser encontrada em Bick (2023).

O PALAVRAS-DIP participou apenas para as obras em formato de texto, o que resultou em que apenas pudéssemos calcular os resultados baseados em 21 obras da coleção dourada.

Pensamos que a falta de participação dos outros interessados, e da comunidade de estudos literários computacionais em geral, se pode dever a diferentes factores:

- Não haver suficientes grupos que trabalhem em estudos literários computacionais em língua portuguesa
- Não haver grande incentivo para esses grupos se dedicarem a uma tarefa que não estava nas prioridades deles
- A tarefa ser razoavelmente difícil, e geralmente feita por linguistas computacionais, ou investigadores de PLN (Processamento de Linguagem Natural): de facto, todos os interessados provinham desta área
- A tarefa, sendo um misto de extração de informação e de recolha de informação, também não se enquadrar naturalmente na maior parte das tarefas dos grupos de PLN
- Não podermos fornecer materiais de treino que permitissem usar técnicas de aprendizagem automática/aprendizado de máquina, que é o paradigma mais utilizado atualmente: o único sistema que concorreu é baseado em regras.

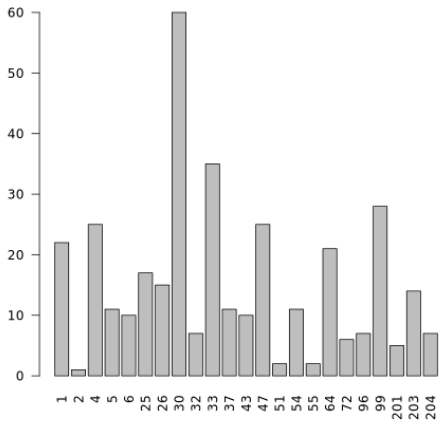

Figura 10: Melhoria das 24 obras da coleção dourada de texto e exemplos, graças ao cotejo com as respostas do PALAVRAS-DIP

4.2. A colaboração entre o resultado automático e a inspeção humana

Um dos resultados mais interessantes e inesperados que surgiu, quando principíamos a avaliação do PALAVRAS-DIP, foi o reconhecimento de que as soluções criadas por seres humanos tinham muitas faltas, que podiam ser facilmente colmatadas por um processo automático. Sobretudo em tarefas como identificar todas as grafias diferentes, os seres humanos não se comparam com uma máquina.

Por isso, fizemos uma nova ronda de "saneamento da coleção dourada", de forma a enriquecer a informação correta, e também para não penalizar a participação do sistema por ser avaliado com base em recursos deficientes.

Na Figura 10 mostramos a mais-valia contribuída pelo sistema participante, para a análise dos 21 textos da coleção dourada em texto e dos 3 textos exemplo em txt.

Em todas as obras foi possível melhorar o número de nomes das personagens, e em muitas delas também identificar mais casos de profissões, ocupações ou estatutos sociais.

4.3. Avaliação

Os resultados da participação do PALAVRAS-DIP foram tornados públicos, e são discutidos em pormenor em Willrich & Santos (2023).

Contudo, parece-nos que, mais do que os números obtidos, a existência de um sistema que conseguiu, embora não perfeitamente, fazer a tarefa que propusemos foi um dos resultados mais importantes do DIP.

Assim, temos uma forma de fazer leitura distante da literatura lusófona, mesmo que o sistema

não a faça perfeitamente. Vamos portanto apresentar de seguida o que aprendemos sobre as 100 obras usando o PALAVRAS-DIP como oráculo. E, depois, aquilo que podemos dizer sobre mais 213 obras, classificadas com uma nova versão do PALAVRAS-DIP em março de 2023, listadas no apêndice A. Essas 213 obras são todos os romances e novelas que faziam parte do corpo Literateca em março de 2013 e não tinham sido selecionadas para a coleção DIP de texto. Chamamos a esta coleção a "coleção extra".

4.4. O género no romance lusófono

Não houve diferenças significativas entre o que observámos com as 42 obras da coleção dourada, com as 100 da coleção de texto do DIP (que incluem 21 das primeiras), e as 213 obras que constituem a coleção extra.

Em praticamente todos os casos houve mais personagens masculinas. Nos dois casos de obras na coleção DIP em que se encontraram mais personagens femininas, eram escritos por mulheres e a personagem principal era um homem: *A vinha* de Ana de Castro Osório e *Jovens interessantes* de Paulina Filadélfia.

A Figura 11 mostra a distribuição de género na análise do PALAVRAS-DIP da coleção DIP, durante a avaliação conjunta.

Na coleção extra, houve onze obras em que o PALAVRAS-DIP identificou mais personagens femininas do que masculinas: *A feiticeira*, *Diário de uma criança* e *Sacrificada*, três novelas de Ana de Castro e Osório, *Um Homem de Brios* de Camilo Castelo Branco, *Os novelos da Tia Filomela*, uma novela de Júlio Dinis, *Herança de lágrimas*, de Ana Plácido, *Statira e Zoroastes*, de Lucas José de Alvarenga, *A Marquesa de Vale Negro*, de Maria O'Neill, *Astúcias de namorada*, de Manuel Pinheiro Chagas, e *Húmus* e *O pobre de pedir* de Raul Brandão.

A Figura 12 mostra a distribuição de género na análise do PALAVRAS-DIP da coleção extra em março de 2023, depois da melhoria do sistema baseada na avaliação do DIP.

4.5. Profissões

Nos resultados do PALAVRAS-DIP relativos à coleção DIP, das 6027 personagens, 4315 não têm profissão, ocupação ou estatuto social (71,6%). Se desagregarmos por género, 1275 mulheres em 1490 não têm este atributo, ou seja, 85,6%. Para os homens, 3040 em 4536 também não têm, ou seja, 67,0%.

Usando os resultados do PALAVRAS-DIP, obtemos exatamente 500 profissões distintas (99 femininas distintas e 428 masculinas distintas). As POES masculinas mais frequentes são *padre*, *general*, *escravo*, *estudante*, *rei* e *capitão*, enquanto que as femininas são *criada*, *rainha*, *escrava* e *princesa*.

No caso da coleção extra, o PALAVRAS-DIP identifica, em março de 2023, 896 profissões diferentes (com a ressalva de que muitas profissões são apenas variantes (orto)gráficas), e as mais frequentes são *padre* (365 casos!), *conde*, *rei* e *capitão*. Para mulheres, 180 profissões distintas foram identificadas pelo PALAVRAS-DIP (algumas erradamente, como por exemplo *abade*) e as profissões, ocupações ou estatutos sociais mais frequentes foram *criada* (72), *rainha*, *condessa* e *soror*.

Reconheceu além disso 22 personagens femininas que eram *escravas* e 3 *mucamas*, e 32 personagens masculinas que eram *escravos*.

Comparando superficialmente a literatura brasileira e a portuguesa através das POES mais frequentes, e embora ambas tenham como profissão mais frequente *padre*, as profissões que se seguem na literatura brasileira são *capitão*, *coronel*, *chefe* e *médico*, enquanto que na literatura portuguesa são *rei*, *conde*, *príncipe* e *mestre*, denunciando claramente o peso dos romances históricos. *Imperador*, pelo contrário, que aparentemente descreveria melhor a realidade brasileira, apenas aparece 13 vezes nesta, contra 12 de *rei*. (Para comparação, e relembrando que a coleção extra tem uma maioria de obras portuguesas, aparecem 14 *imperadores* e 110 *reis* na subcoleção portuguesa.)

4.6. Relações familiares

As relações familiares que o PALAVRAS-DIP identificou durante o DIP, e as relativas à coleção extra encontram-se nas Figuras 13 e 14.

É interessante ver que deixa de ser pai a relação mais frequente para ser filho em ambas as coleções, mas que pai continua a ser mais frequente que mãe.

4.7. Nomes e formas de tratamento

Apresentamos aqui os nomes mais comuns nas coleções analisadas, segundo o PALAVRAS-DIP, tendo sido manualmente removidos casos de erro.

A Figura 15 apresenta os nomes com frequência superior a 20 na coleção do DIP, e a Figura 16 os nomes com mais de 20 casos na coleção extra.

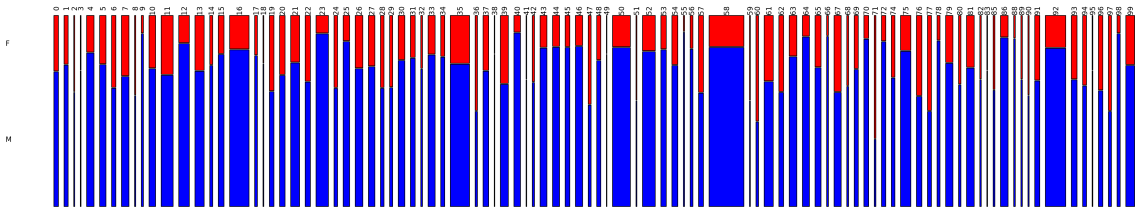

Figura 11: Género na coleção DIP de texto, de acordo com o PALAVRAS-DIP: cada coluna do mosaico representa uma obra, e a área vermelha marca as personagens femininas, e a azul as masculinas

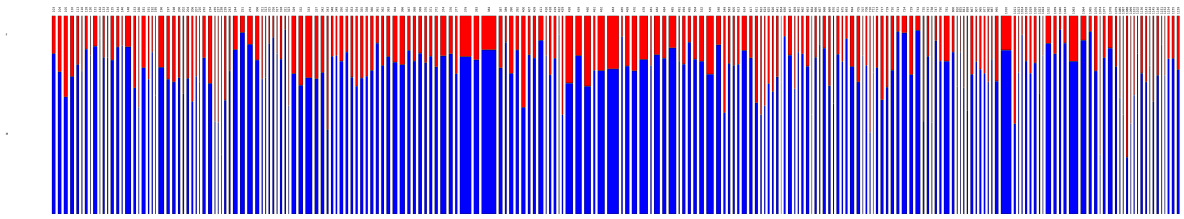

Figura 12: Género na coleção extra, de acordo com o PALAVRAS-DIP em março de 2023

Se quisermos apenas os nomes femininos, veja-se as Figuras 17 e 18.

Não é um assunto que seja provavelmente muito interessante do ponto de vista literário ou linguístico, mas é de notar que os nomes próprios mais frequentes não apresentam quase diferença nenhuma entre as literaturas portuguesa e brasileira, algo que muito provavelmente se deve à predominância de obras do século XIX.

Quanto às formas de tratamento, a situação também é semelhante nas várias coleções, veja-se as Figuras 19 e 20.

A maior diferença em relação à coleção dourada é o aparecimento da forma de tratamento *mestre*.

Debrucemo-nos agora sobre os diminutivos: Na coleção do DIP, o PALAVRAS-DIP identificou 299 diminutivos, 137 masculinos e 162 femininos. Destes, 157 eram portugueses, e 142 brasileiros.

Na coleção extra, o PALAVRAS-DIP identificou 497 diminutivos: 244 diminutivos masculinos e 253 femininos. Isto significa, visto que o PALAVRAS-DIP identificou 3281 personagens femininas e 10144 masculinas, que há muito mais diminutivos femininos: 7,7% contra 2,4%. Os 497 casos correspondem a 284 diminutivos diferentes.

4.8. Em resumo

Em resumo, embora certamente o PALAVRAS-DIP não consiga obter exatamente todas as personagens e só as personagens, os resultados acumulados nas duas coleções vão na mesma direção que a informação que tínhamos coligido manualmente na coleção dourada, o que nos dá esperança de que a visão — em leitura distante — da literatura lusófona que conseguimos obter, usando este sistema, seja relativamente correta.

5. Comentários finais

O DIP foi uma avaliação conjunta destinada a desenvolver sistemas que, dada uma obra literária, obtivessem as suas personagens e algumas características e relacionamentos destas. A ideia era conseguir olhar para a literatura lusófona como um todo e produzir algumas generalizações, assim como distinguir obras ou grupos de obras que se destacassem.

Oferecemos estes dados — que, aliás, se encontram públicos para permitir que outros investigadores os manipulem, corrijam e estudem — como um princípio para esse objetivo.

É importante salientar que o grosso da literatura lusófona ainda não se encontra em forma de texto de alguma qualidade, e que por isso todas as obras ainda não digitalizadas ou com um Reconhecimento óptico de Caracteres (ROC) muito deficiente não podem ainda ser tomadas em conta. Urge desenvolver sistemas de ROC fiáveis para a literatura não contemporânea em português.

Este artigo pretendeu antes do mais dar uma panorâmica geral sobre a organização do DIP e sobre os recursos compilados. Os outros artigos

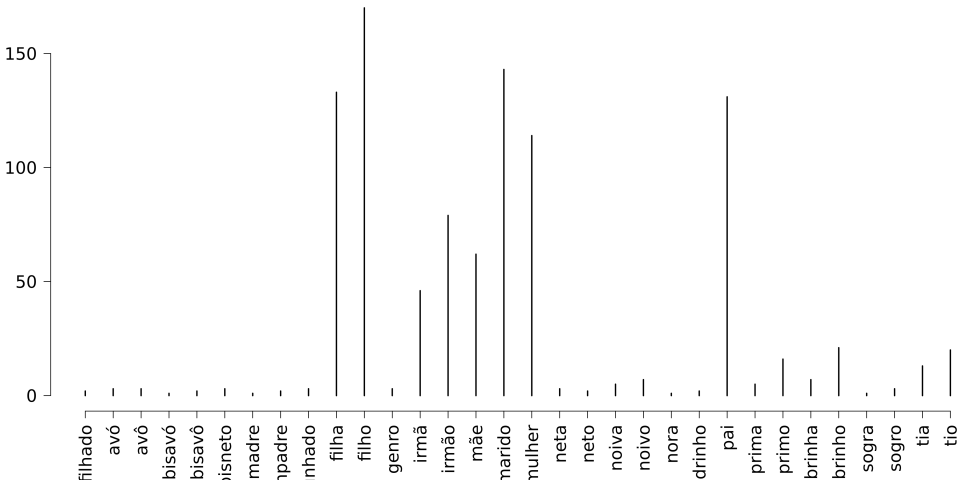

Figura 13: Relações familiares na coleção DIP de texto, de acordo com o PALAVRAS-DIP

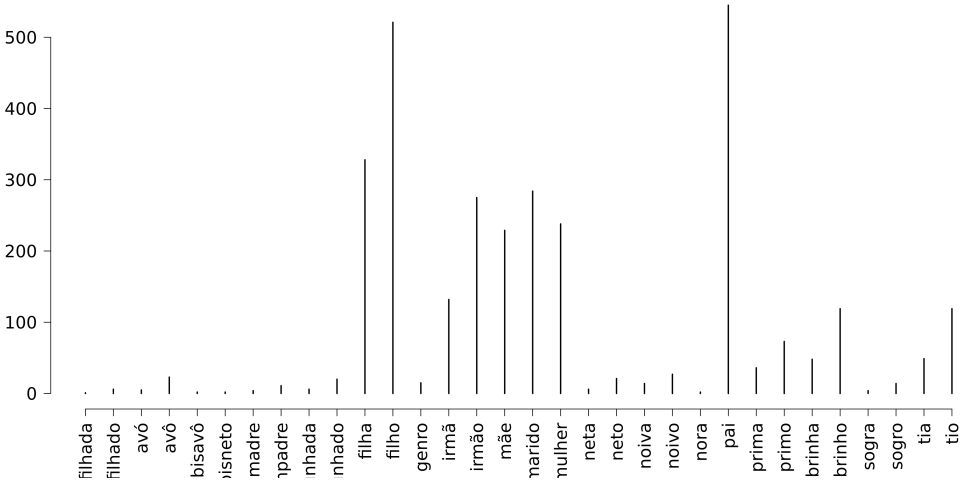

Figura 14: Relações familiares na coleção extra, de acordo com o PALAVRAS-DIP em março de 2023

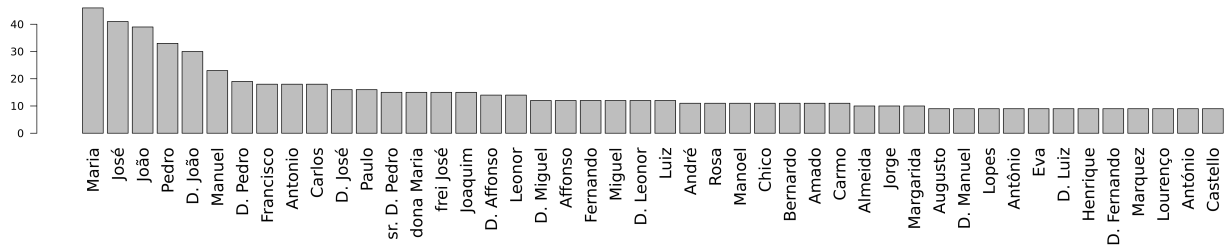

Figura 15: Nomes mais frequentes na coleção DIP de texto, de acordo com o PALAVRAS-DIP

deste volume descrevem com mais profundidade várias vertentes do DIP, como a caracterização do género e do estatuto profissional de um ponto de vista dos estudos literários (Pires et al., 2023), o estudo das relações familiares na literatura recorrendo a conceitos da teoria de redes (Mota & Santos, 2023), a forma de avaliação (Willrich & Santos, 2023) e, por último, mas provavelmente o mais importante, como o sistema participante resolveu a tarefa do DIP e como continua a evoluir (Bick, 2023).

Pensamos que, antes de organizar nova edição, é importante olharmos com muito cuidado para a informação sobre as obras que conseguimos obter, eventualmente melhorando-a e enriquecendo-a, de forma a propor outras maneiras de prosseguir na caracterização das personagens e das obras. É preciso que a comunidade se debruce sobre os dados já obtidos, os problemas encontrados, e os desejos expressos, para que todos possamos saber qual a contribuição que o DIP terá dado aos estudos literários lusófonos.

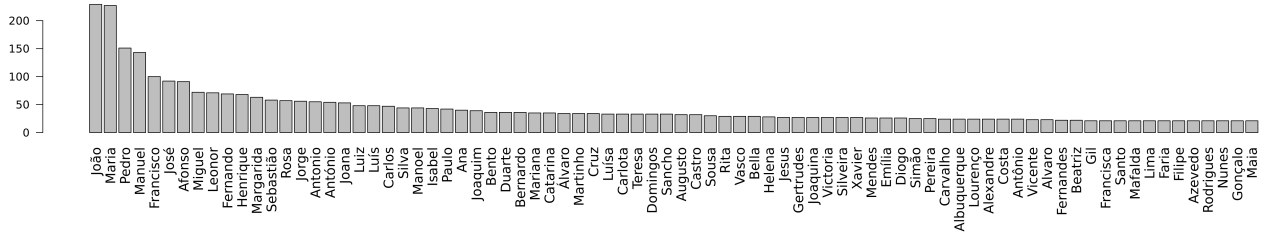

Figura 16: Nomes mais frequentes na coleção extra, de acordo com o PALAVRAS-DIP

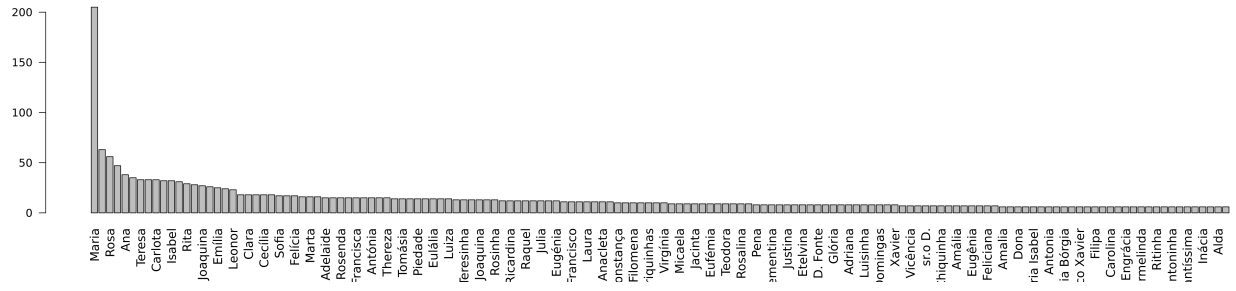

Figura 17: Nomes femininos mais frequentes na coleção DIP, de acordo com o PALAVRAS-DIP

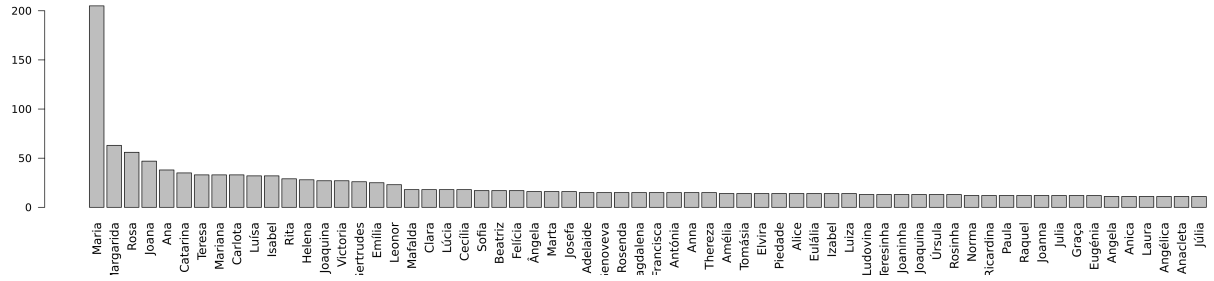

Figura 18: Nomes femininos mais frequentes na coleção extra, de acordo com o PALAVRAS-DIP

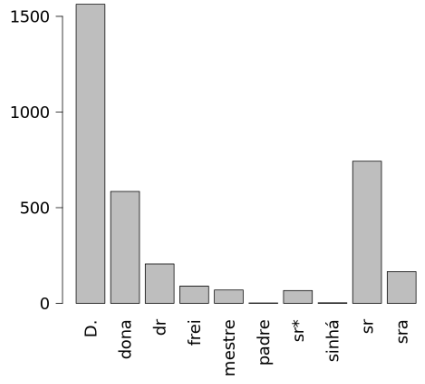

Figura 19: Formas de tratamento na coleção do DIP, de acordo com o PALAVRAS-DIP

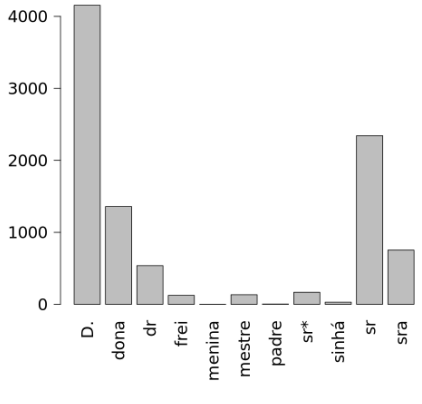

Figura 20: Formas de tratamento na coleção extra, de acordo com o PALAVRAS-DIP

Apresentamos pois o nosso trabalho apenas como um primeiro passo, cuja relevância depende de ter ou não estudiosos que se debrucem sobre os dados produzidos.

Agradecimentos

A organização do DIP tem de agradecer a muitas pessoas nele envolvidas. Em primeiro lugar, tivemos os compiladores da coleção dourada, que tiveram de ler as obras de fio a pavio e obter os resultados certos. Além da própria organização, agradecemos, por ordem alfabética, a Jonas Albuquerque, Luíla Lima, Luísa Lima, Marcus Vinicius Correa, Oriana Pereira, Patrícia Magalhães, Ruth Hoff e Sara Botelho.

Agradecemos calorosamente a Eckhard Bick, o único participante que conseguiu desenvolver um sistema para participar no DIP, sem o qual não poderíamos apresentar resultados.

Agradecemos aos participantes no Encontro do DIP, sobretudo aqueles que deram o seu contributo na hora da discussão, nomeadamente e mais uma vez por ordem alfabética, Cláudia Freitas, Luísa Coheur, Maria José Finatto, Raquel Amaro e Roberlei Alves Bertucci.

Agradecemos aos observadores do DIP pelo interesse e pelos comentários críticos, especialmente a Dionéia Motta, Alexandre Rademaker e Sílvia Araújo.

Agradecemos a Luisa Coheur, Alexandre Rademaker e Roberlei Alves Bertucci os muitos e pertinentes comentários a uma versão anterior deste texto, que contribuíram decisivamente para a sua melhoria.

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

Agradecemos também ao ILOS a contratação de dois assistentes de investigação para ajudar a organização do DIP, assim como o financiamento da viagem do participante a Oslo para o Encontro do DIP.

E finalmente, agradecemos à FCCN–Fundação para a Computação Científica Nacional (Portugal) o alojamento da Linguateca nos seus servidores, e ao UNINETT Sigma2 - the National Infrastructure for High Performance Computing and Data Storage in Norway pelos recursos computacionais.

Referências

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

Cardoso, Nuno & Diana Santos. 2007. Directivas para a identificação e classificação semântica na colecção dourada do HAREM. Em Diana Santos & Nuno Cardoso (eds.), *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*, 211–238. https://www.linguateca.pt/aval_conjunta/LivroHAREM/Cap16-SantosCardoso2007-CardosoSantos.pdf.

de Does, Jesse, Katrien Depuydta, Karina van Dalen-Oskamb & Maarten Marx. 2017. Namescape: Named entity recognition from a literary perspective. Em Jan Odijk & Arjan van Hessen (eds.), *CLARIN in the Low Countries*, 361–370.

Higuchi, Suemi, Diana Santos, Cláudia Freitas & Alexandre Rademaker. 2019. Distant reading Brazilian politics. Em *4th Conference of The Association Digital Humanities in the Nordic Countries*, 190–200.

Krug, Markus, Lukas Weimer, Isabella Reger, Luisa Macharowsky, Stephan Feldhaus, Frank Puppe & Fotis Jannidis. 2018. Description of a corpus of character references in german novels: DROC. Relatório técnico. Georg-August-Universität. http://webdoc.sub.gwdg.de/pub/mon/dariah-de/dwp-2018-27.pdf.

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.

Mota, Cristina & Diana Santos. 2023. Pais, filhos, e outras relações familiares no DIP. *Linguamática* 15(1). 41–53. 10.21814/lm.15.1.402.

Pires, Emanoel, Marcia Caetano Langfeldt & Rebeca Schumacher Fuão. 2023. Desafios e vantagens do processo de identificação automática do gênero e das profissões das personagens no dip. *Linguamática* 15(1). 55–67. 10.21814/lm.15.1.401.

Santos, Diana. 2019. Literature studies in literateca: between digital humanities and corpus linguistics. Em Martin Doerr, Øyvind Eide, Oddrun Grønvik & Bjørghild Kjelsvik (eds.), *Humanists and the digital toolbox: In honour of Christian-Emil Smith Ore*, 89–109. Novus Forlag.

Santos, Diana. 2022a. Evaluation contests in Portuguese: Linguateca. *submitted* https://www.linguateca.pt/Diana/download/AvalConjLRE.pdf.

Santos, Diana. 2022b. Futuro risonho: prolegómenos para uma colaboração entre a Linguateca e o NuPILL. Em Isabela Melim Borges & Paulo Henrique Pergher (eds.), *Literatura e seus híbridos III: 25 anos do NuPiLL*, 285–308. UFSC.

Santos, Diana. 2022c. A gramateca e a literateca como macroscópios linguísticos. *Domínios de Lingu@gem* 16(4). 1242–1265. 10.14393/DL52-v16n4a2022-2.

Santos, Diana, Daniel Alves, Raquel Amaro, Isabel Araújo Branco, Olivia Fialho, Cláudia Freitas, Suemi Higuchi, Marcia Langfeldt, João Marques Lopes, Alckmar Luiz dos Santos, Emanoel Pires, Barbara Ramos, Danielle Sanches, Rebeca Schumacher Fuão, Paulo Silva Pereira & Paula Terra. 2020a. Leitura distante em Português: resumo do primeiro encontro. *Materialidades da Literatura* 8(1). 279–298. 10.14195/2182-8830_8-1_16.

Santos, Diana, Eckhard Bick & Marcin Wlodek. 2020b. Avaliando entidades mencionadas na coleção ELTeC-por. *Linguamática* 12(2). 29–49. 10.21814/lm.12.2.336.

Santos, Diana & Cláudia Freitas. 2019. Estudando personagens na literatura lusófona. Em *XII Symposium in Information and Human Language Technology and Collocates Events (STIL)*, 48–52.

Santos, Diana, Cláudia Freitas & Eckhard Bick. 2018. OBras: a fully annotated and partially human-revised corpus of Brazilian literary works in public domain. Em *CorLex*, s.p.

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. 10.1007/978-3-030-98305-5_39.

Schöch, Christof, Tomaz Erjavec, Roxana Patras & Diana Santos. 2021. Creating the European Literary Text Collection (ELTeC): Challenges and perspectives. *Modern Languages Open* 1. 1–19. 10.3828/mlo.v0i0.364.

Willrich, Roberto & Diana Santos. 2023. Avaliação no desafio de identificação de personagens. *Linguamática* 15(1). 69–87. 10.21814/lm.15.1.398.

A. A lista das obras da coleção do DIP

Id	Autor	Título	Ano
0	Miguel Vale de Almeida	*Euronovela*	1998
1	Cosme Velho	*Miss Kate*	1909
2	Alice Pestana	*A vida por um prejuízo*	1908
3	José da Rocha Leão	*A cruz de fogo*	1862
4	António José Coelho Lousada	*Os tripeiros*	1857
5	Luísa Marques da Silva	*mISTério@Tagus*	2021
6	Coelho Neto	*Turbilhão*	1904
7	Lindolfo Rocha	*Maria Dusá*	1910
8	José Joaquim Rodrigues de Bastos	*O médico do deserto*	1864
9	Apolinário Porto-Alegre	*O vaqueano*	1872
10	Inglês de Sousa	*O cacaolista: scenas da vida do Amazonas*	1899
11	Manuel de Oliveira Paiva	*Dona Guidinha do Poço*	1891
12	Artur Lobo de Ávila	*Os Caramurús: romance histórico da descoberta e independência do Brasil*	1900
13	Luís Guimarães Júnior	*Histórias para gente alegre: A família Agulha*	1870
14	Antônio Gonçalves Teixeira e Souza	*Maria ou a menina roubada*	1852
15	Conde de Ficalho	*Uma Eleição Perdida*	1888
16	António Campos Junior	*A Ala dos Namorados*	1905
17	António Pedro Lopes de Mendonça	*Memorias d'um doido: romance contemporâneo*	1849
18	Guiomar Torrezão	*Severina*	1890
19	Almiro Caldeira da Andrade	*Arca açoriana*	1984
20	Faustino da Fonseca	*Os bravos do Mindelo: Romance histórico*	1906
21	Alfredo de Mesquita	*A rua do oiro*	1905
22	Domingos Olímpio	*Luzia Homem*	1903
23	Luís Augusto Rebelo da Silva	*A Casa dos Fantasmas*	1908
24	Maria O'Neill	*Por bom caminho*	1914
25	Júlio Ribeiro	*A carne*	1888
26	Lourenço Amazonas	*Simá*	1899
27	Antonio Lobo	*A carteira de um neurastênico*	1903
28	Pedro Ivo	*O selo da roda*	1876
29	Manuel Antonio de Almeida	*Memórias de um sargento de milícias*	1852
30	Aluísio Azevedo	*Casa de pensão*	1884
31	Visconde de Taunay	*No declínio*	1889
32	Alice Moderno	*O Dr. Luís Sandoval*	1892
33	Antonio Augusto Teixeira de Vasconcellos	*A ermida de Castromino*	1870
34	Rodolpho Theophilo	*O paroara: scenas da vida cearense e amazonica*	1899
35	João da Câmara	*O Conde de Castel Melhor*	1903
36	Manuel Maria Rodrigues	*A Rosa do Adro*	1870
37	Arnaldo Gama	*El-rei dinheiro*	1876
38	Alfredo de Morais Pinto	*Aventuras do sr. Criptogamo : romance humorístico*	1899
39	João do Rio	*A Profissão de Jacques Pedreira*	1910
40	Teófilo Braga	*Viriato*	1904
41	Maria Amália Vaz de Carvalho	*Alice*	1877

(Continua)

Id	Autor	Título	Ano
42	Raul de Azevedo	*Tríplice Aliança*	1907
43	Franklin Távora	*O cabeleira: História pernambucana*	1876
44	Luiz Gonzaga Duque Estrada	*Mocidade morta*	1899
45	Urbano Loureiro	*A infâmia de Frei Quintino*	1878
46	António Francisco Barata	*Um duelo nas sombras, ou D. Francisco Manuel de Melo (1630)*	1875
47	Bernardo Guimarães	*A escrava Isaura*	1875
48	Francisco Gomes de Amorim	*Os selvagens*	1875
49	Lima Barreto	*Triste Fim de Policarpo Quaresma*	1911
50	Oliveira Mascarenhas	*O trovador da infanta*	1906
51	Ana de Castro Osório	*A vinha*	1908
52	Alberto Pimentel	*A guerrilha de Frei Simão*	1895
53	Ramalho Ortigão e Eça de Queirós	*O Mistério da Estrada de Sintra*	1870
54	Abel Botelho	*O Barão de Lavos*	1891
55	Adolfo Caminha	*O Bom-Crioulo*	1895
56	Souza Lima	*O Tupinambá*	1931
57	José Augusto Vieira	*A divorciada*	1881
58	Rocha Martins	*Gomes Freire*	1900
59	Viriato Bandeira Duarte	*Ida*	1865
60	Júlia Lopes de Almeida	*A viúva Simões*	1895
61	Teixeira de Queirós	*O Salústio Nogueira*	1909
62	Ana Plácido	*Adelina*	1863
63	Almeida Garrett	*Viagens na minha terra*	1846
64	Raul Pompéia	*O Ateneu*	1888
65	Graça Aranha	*A viagem maravilhosa*	1929
66	Machado de Assis	*O alienista*	1882
67	Carlos Malheiro Dias	*Filho das ervas*	1900
68	Pardal Mallet	*Hóspede*	1887
69	Aluísio Azevedo	*O coruja*	1889
70	Alexandre Herculano	*O bobo*	1843
71	Paulina Filadélfia	*Jovens interessantes*	1865
72	Diogo de Macedo	*O Cristão novo*	1876
73	José do Patrocínio	*Os retirantes*	1879
74	Francisco Luís Gomes	*Os Brâmanes*	1866
75	Zeferino Norberto Gonçalves Brandão	*Pero da Covilhã: Episódio Romântico do Século XV*	1897
76	M.M.S.A. e Vasconcelos	*O Cura de São Lourenço*	1855
77	Augusto Loureiro	*A bruxa: cenas açorianas*	1901
78	Amadeu Amaral	*Memorial de um passageiro de bonde*	1938
79	Valentim Magalhães	*Flor de sangue*	1897
80	Maria Peregrina de Sousa	*Henriqueta*	1867
81	Eça de Queirós	*A ilustre Casa de Ramires*	1900
82	Bruno Seabra	*Paulo*	1861
83	José de Alencar	*A viuvinha*	1857
84	Antônio de Alcântara Machado	*Brás, Bexiga e Barra Funda*	1927
85	Maria Firmina dos Reis	*Úrsula*	1859
86	Paulo Setúbal	*O sonho das esmeraldas*	1935
87	Antônio de Alcântara Machado	*Mana Maria*	1936
88	Manuel Pinheiro Chagas	*Um melodrama em Santo Tirso*	1873
89	Caetano Alves de Sousa Filgueiras	*Adelaide de Sargans*	1870

(Continua)

Id	Autor	Título	Ano
90	Raúl Brandão	*A Morte do Palhaço*	1926
91	Emília Bandeira de Melo	*A luta*	1911
92	Camilo Castelo Branco	*A Brasileira de Prazins*	1882
93	Virgílio Várzea	*George Marcial*	1901
94	Virgínia de Castro e Almeida	*Aventuras de Dona Redonda*	1943
95	Casimiro de Abreu	*Carolina*	1856
96	Mário de Andrade	*Amar, verbo intransitivo*	1927
97	Joaquim Manoel de Macedo	*A Moreninha*	1844
98	Moreira de Azevedo	*Os franceses no Rio de Janeiro*	1870
99	Júlio Dinis	*Uma família inglesa*	1867
100	Pinheiro Chagas	*O terremoto de Lisboa*	1874
101	António Inocêncio Barbuda	*O português generoso, ou Aventuras de J... e S... e seu ditoso fim: História verdadeira*	1820
102	A. Augusto de Pinho	*Remédio para matar paixões*	1879
103	Afonso Arinos de Melo Franco	*Os jagunços*	1897
104	Camilo Castelo Branco	*O judeu*	1866
105	Luiz Caetano de Campos	*Viagens de Altina, nas cidades mais cultas da Europa e nas principais povoações dos Balinos, povos desconhecidos de todo o mundo*	1790
106	Eduardo de Borja Reis	*O crime do beato Antônio: Romance original português*	1887
107	Carneiro Vilela	*Noêmia*	1894
108	Claudia de Campos	*Último amor*	1894
109	Germano Hasslocher	*A espelunca: Romance de atualidade*	1889
110	Pinheiro Chagas	*Os guerrilheiros da Morte*	1872
111	Domingos Jaguaribe	*Os herdeiros do caramuru*	1880
112	Érico Veríssimo	*Caminhos Cruzados*	1935
113	Araripe Júnior	*Luizinha: Romance de costumes cearenses*	1878
114	Alice Pestana	*A filha do João do Outeiro*	1933
115	Flávio de Carvalho	*Os ossos do mundo*	1936
116	Felício dos Santos	*Acayaca*	1866
117	Lúcio de Mendonça	*O marido da adúltera: Crônica fluminense*	1881
118	Visconti Coaracy	*Amor que mata*	1873
119	António Ernesto Tavares de Andrade	*Eugénio, ou o livre pensador*	1871
120	Bezerra de Menezes	*A casa mal-assombrada: Romance de costumes sertanejos*	1888
121	António Joaquim da Rosa	*A cruz de cedro*	1854
122	Teixeira e Sousa	*O filho do pescador: Romance brasileiro original*	1843
123	José Agostinho de Macedo	*O arrependimento premiado: História verdadeira*	1818
124	Francisco Coelho Duarte Badaró	*Fantina (Cenas da escravidão)*	1881
125	António Francisco Barata	*Os jesuítas na corte*	1877
126	Francisca Senhorinha da Motta Dinis	*A Judia Raquel*	1886
127	Menotti Del Picchia	*Laís*	1921
128	José Lins do Rego	*Doidinho*	1934
129	Carlos Pinto de Almeida	*A filha do emir*	1875
130	Alfredo Hogan	*A pedinte de Lisboa*	1859
131	Maria ONeill	*Lucta de sentimentos*	1912

(Continua)

Id	Autor	Título	Ano
132	Pedro Américo	*O foragido*	1899
133	Pedro Ribeiro Viana	*Elzira, a morta virgem*	1883
134	Tomás de Mello	*O Conde de S. Luiz*	1874
135	Reinaldo Ferreira	*O Presidente da República*	1923
136	Júlio Ribeiro	*Padre Belchior de Pontes*	1874
137	Cornélio Pena	*A menina morta*	1954
138	Júlio César Machado	*Cláudio. Romance*	1852
139	Alberto Pimentel	*As netas do Padre Eterno*	1895
140	Vicente Temudo Lessa	*O velho manuscrito*	1888
141	Luísa F. de Camargo Pacheco	*Alice*	1903
142	Lourenço Caiola	*Conversão*	1923
143	Ana Luísa de Azevedo Castro	*D. Narcisa de Vilar: Legenda do tempo colonial pela Indígena do Ipiranga*	1858
144	desc.	*O sapateiro de Azeitão*	1865
145	Elias António da Fonseca	*Doroteia, ou A lisbonense infeliz*	1816
146	Batista Cepelos	*O vil metal*	1910
147	Rosendo Moniz	*Favos e travos*	1872
148	José da Fonseca	*Aventuras de Telêmaco, filho de Ulisses, seguidas das de Aristonoo e de Ulisses: Compendiadas para uso dos meninos*	1854
149	Carlos Pinto de Almeida	*Os homens da cruz vermelha*	1879-1880
150	Augusto Emílio Zaluar	*O doutor Benignus*	1874
151	Teixeira de Vasconcellos	*O prato de arroz doce*	1862
152	Fortunato Correia Pinto	*O agitador*	1906
153	Júlio Lourenço Pinto	*O Bastardo: Cenas da vida contemporânea.*	1889
154	Leonel de Alencar, Barão de Alencar	*A sonâmbula da Itapuca*	1861
155	Bruno Seabra	*Memórias de um pobre diabo*	1868
156	Antônio Deodoro de Pascual	*Um episódio da história pátria: As quatro derradeiras noites dos inconfidentes (1792)*	1868
157	L.A. Rebello da Silva	*Ódio velho nao cança*	1848
158	Conde Afonso Celso	*Lupe*	1894
159	Virgínia de Castro e Almeida	*Trabalho bem-dito*	1908
160	José da Rocha Leão	*Os subterrâneos do Morro do Castelo: Seus mistérios e tradições*	1878
161	Eça de Queirós	*O mandarim*	1880
162	Dionísia Gonçalves Pinto	*Dedicação de uma amiga*	1850
163	Joaquim Norberto	*O martírio do Tiradentes, ou Frei José do Desterro. Lenda Brasileira*	1882
164	Antônio Manuel Policarpo da Silva	*Cadelinha*	1816
165	Antero de Figueiredo	*Leonor Teles: Flor de altura*	1916
166	Teixeira de Queiroz	*Morte de D. Agostinho*	1895
167	Emília Freitas	*A rainha do ignoto: Romance psicológico*	1899
168	António Joaquim de Mesquita e Melo	*D. Sancho II, quarto rei de Portugal*	1869
169	Medeiros e Albuquerque, Afrânio Peixoto, Coelho Neto, Viriato Correia	*O mistério*	1920
170	Lúcio Bruno	*A mão negra e a polícia: Sensacional romance dos crimes célebres, praticados pelo Dioguinho, o terror dos sertões paulistas*	1923
171	Mário de Sá Carneiro	*A confissão de Lúcio*	1913
172	João de Andrade Corvo	*O sentimentalismo*	1871

(Continua)

Id	Autor	Título	Ano
173	Soeiro Pereira Gomes	*Esteiros*	1941
174	D. Bruno da Silva	*A beata de Évora: Romance histórico 1764-1828*	1890
175	J. M. Pereira da Silva	*Jerônimo Corte Real*	1840
176	Carlos Malheiro Dias	*A mulata*	1975
177	Francisco Gomes de Amorim	*As duas fiandeiras*	1881
178	Carolina Michaëlis de Vasconcelos e Afonso Lopes Vieira	*O romance de Amadis: Composto sobre o Amadis de Gaula, de Lobeira*	1922
179	Luís da Silva Alves de Azambuja Suzano	*O Capitão Silvestre e Frei Veloso, ou A plantação de Café no Rio de Janeiro*	1847
180	Rachel de Queiroz	*O quinze*	1930
181	Luís Ratozi	*Amores pagãos*	1934
182	Cônego Ulisses de Penaforte	*Mandu (o eremícola): Romance indo-brasileno neontológico e nativista*	1901
183	Francisco Soares Franco Júnior	*Memórias da mocidade:*	1867
184	Ana de Castro Osório	*Mundo novo*	1927
185	Alfredo Campos	*A filha do cabinda*	1873
186	Xavier Marques	*O feiticeiro*	1922
187	Júlio César Leal	*Casamento e mortalha no céu se talha*	1876
188	Barão de Teffé	*A corveta Diana: Romance marítimo. Original brasileiro.*	1873
189	Ana Plácido	*Herança de lágrimas*	1871
190	Luís Ramos Figueira	*Dalmo ou Mistérios da noite*	1863
191	Salvador de Mendonça	*Marabá*	1875
192	José Antônio do Vale Caldre Fião	*A divina pastora: Novela rio-grandense*	1847
193	Alberto Osório de Castro	*Dramas da côrte*	1905
194	Alberto Braga	*Os confidentes*	1887
195	Dona Maria Benedita Câmara de Bormann	*Estátua de neve*	1890
196	Ana Maria Ribeiro de Sá	*Matilde*	1874
197	João Salomé Queiroga	*Maricota e o Padre Chico*	1871
198	Arnaldo Gama	*O satanás de Coura: Memórias do século XVII*	2002
199	Almada Negreiros	*Nome de guerra*	1938

B: A lista das obras da coleção extra

Id	Autor	Título	Ano
103	Abel Botelho	*Amanhã*	1901
104	Abel Botelho	*Amor crioulo*	1919
105	António da Costa Couto Sá de Albergaria	*Os filhos do padre Anselmo*	1904
110	Adolfo Caminha	*A Normalista*	1893
113	Adolfo Caminha	*Tentação*	1896
128	Alberto Osório de Castro	*Dramas da corte*	1905
130	Alberto Pimentel	*Cristo não volta*	1873
131	Alberto Pimentel	*O Anel Misterioso: Cenas da Guerra Peninsular*	1873
132	Alberto Pimentel	*A última ceia do Doutor Fausto*	1876
133	Alberto Pimentel	*As noites do asceta*	1876

(Continua)

Id	Autor	Título	Ano
134	Alberto Pimentel	*O Romance da Rainha Mercedes*	1879
135	Alberto Pimentel	*Noites de Sintra*	1892
144	Alexandre Herculano	*Eurico o Presbítero*	1844
146	Alexandre Herculano	*O Galego*	1846
148	Alexandre Herculano	*O Monge de Cister I*	1848
153	Alexandre Herculano	*O Pároco de Aldeia*	1851
160	Alfredo Campos	*A filha do Cabinda*	1873
181	J. B. da Silva L. de Almeida Garrett	*O Arco de Santana*	1845
191	J. B. da Silva L. de Almeida Garrett	*Helena*	1854
192	José de Almada Negreiros	*A engomadeira: novela vulgar lisboeta*	1917
195	Aluísio Azevedo	*Uma lágrima de mulher*	1879
196	Aluísio Azevedo	*O Mulato*	1881
197	Aluísio Azevedo	*A Condessa Vésper ou Memórias de um Condenado*	1882
198	Aluísio Azevedo	*Girândola de Amores ou Mistério da Tijuca*	1882
200	Aluísio Azevedo	*Filomena Borges*	1884
202	Aluísio Azevedo	*O Homem*	1887
204	Aluísio Azevedo	*O Cortiço*	1890
206	Aluísio Azevedo	*A Mortalha de Alzira*	1894
207	Aluísio Azevedo	*O Livro de uma Sogra*	1895
210	Álvaro do Carvalhal	*Os Canibais*	1868
232	A.M. da Cunha e Sá	*Da parte d'el-rei*	1873
234	Ana de Castro Osório	*Ambições*	1903
235	Ana de Castro Osório	*A feiticeira*	1908
237	Ana de Castro Osório	*Diário de uma criança*	1908
238	Ana de Castro Osório	*Sacrificada*	1908
239	Ana Plácido	*Adelina*	1863
243	Anna Maria Ribeiro de Sá	*Matilde*	1874
244	António de Albuquerque	*O Marquês da Bacalhoa*	1908
251	António Francisco Barata	*O Manuelinho de Évora*	1873
253	António Francisco Barata	*O último cartuxo da Scala Caeli de Évora: Romance histórico (1808-1865)*	1891
306	Augusto Sarmento	*Providência*	1863
311	Bernardo Guimarães	*O ermitão do Muquém*	1868
313	Bernardo Guimarães	*O seminarista*	1872
315	Bernardo Guimarães	*Maurício*	1877
316	Bernardo Guimarães	*O bandido do Rio das Mortes*	1905
317	Bernardo Guimarães	*O garimpeiro*	1972
318	Bernardino Pereira Pinheiro	*Arzila: Romance do Século XV*	1862
321	Brito Camacho	*Ao de leve*	1913
323	Bulhão Pato	*A pálida estrela*	1864
329	Camilo Castelo Branco	*Anátema*	1851
332	Camilo Castelo Branco	*A Filha do Arcedíago*	1854
333	Camilo Castelo Branco	*Mistérios de Lisboa*	1854
337	Camilo Castelo Branco	*Livro Negro de Padre Dinis I*	1855
342	Camilo Castelo Branco	*Onde Esta a Felicidade*	1856
343	Camilo Castelo Branco	*Um Homem de Brios*	1856
348	Camilo Castelo Branco	*A Vingança*	1858
349	Camilo Castelo Branco	*O Que Fazem Mulheres*	1858
350	Camilo Castelo Branco	*Cenas da Foz*	1860
352	Camilo Castelo Branco	*Romance dum Homem Rico*	1861
353	Camilo Castelo Branco	*Amor de Perdição*	1862

(Continua)

Id	Autor	Título	Ano
354	Camilo Castelo Branco	*Coisas Espantosas*	1862
355	Camilo Castelo Branco	*Coração Cabeça e Estômago*	1862
358	Camilo Castelo Branco	*Aventuras de Basílio Fernandes Enxertado*	1863
360	Camilo Castelo Branco	*O Bem e o Mal*	1863
361	Camilo Castelo Branco	*A Filha do Doutor Negro*	1864
362	Camilo Castelo Branco	*Amor de Salvação*	1864
363	Camilo Castelo Branco	*No Bom Jesus do Monte*	1864
364	Camilo Castelo Branco	*Vinte Horas de Liteira*	1864
366	Camilo Castelo Branco	*A Queda dum Anjo*	1866
367	Camilo Castelo Branco	*O olho de vidro: romance histórico*	1866
368	Camilo Castelo Branco	*A Doida do Candal*	1867
369	Camilo Castelo Branco	*O Retrato de Ricardina*	1868
370	Camilo Castelo Branco	*Os Brilhantes do Brasileiro*	1869
371	Camilo Castelo Branco	*A Infanta Capelista*	1872
372	Camilo Castelo Branco	*Livro de Consolação*	1872
374	Camilo Castelo Branco	*O Carrasco de Vitor Hugo*	1872
376	Camilo Castelo Branco	*A Filha do Regicida*	1875
377	Camilo Castelo Branco	*A Freira no Subterraneo*	1875
379	Camilo Castelo Branco	*A Caveira da Mártir*	1876
383	Camilo Castelo Branco	*Eusébio Macário*	1879
384	Camilo Castelo Branco	*A Corja*	1880
387	Camilo Castelo Branco	*Vulcões de Lama*	1886
389	Cândido de Figueiredo	*Lisboa no Ano Três Mil*	1892
390	Carlos Malheiro Dias	*A Mulata*	1896
392	Carlos Pinto de Almeida	*A conquista de Lisboa*	1866
400	Claudia de Campos	*Ele*	1899
407	Coelho Neto	*Miragem*	1895
409	Coelho Neto	*O morto*	1898
411	Coelho Neto	*A conquista*	1899
416	Coelho Neto	*Esfinge*	1908
418	Coelho Neto	*Rei negro*	1914
419	Coelho Neto	*A capital federal*	1915
422	Coelho Neto	*Mano*	1924
429	Conde de Ficalho	*Mais Uma*	1888
458	José Maria Eça de Queirós	*O Crime do Padre Amaro*	1875
459	José Maria Eça de Queirós	*A Tragédia da Rua das Flores*	1878
460	José Maria Eça de Queirós	*O Primo Basílio*	1878
461	José Maria Eça de Queirós	*O Mandarim*	1880
462	José Maria Eça de Queirós	*A Relíquia*	1887
463	José Maria Eça de Queirós	*Os Maias*	1888
465	José Maria Eça de Queirós	*As Minas de Salomão*	1891
469	José Maria Eça de Queirós	*Fradique Mendes*	1900
470	José Maria Eça de Queirós	*A Cidade e as Serras*	1901
478	José Maria Eça de Queirós	*A Capital*	1925
481	José Maria Eça de Queirós	*Alves e Companhia*	1925
482	José Maria Eça de Queirós	*O Conde d Abranhos*	1925
484	José Maria Eça de Queirós	*Cartas Inéditas de Fradique Mendes*	1929
485	Eduardo de Noronha	*O agonizar de uma dinastia*	1908
491	Francisco d'Athayde Machado de Faria e Maia	*Vencido*	1914
494	António Feliciano de Castilho	*A chave do enigma*	1861
495	José Maria Ferreira de Castro	*A selva*	1930
504	Faustino da Fonseca e Joaquim Leitão	*Os filhos de Inês de Castro*	1905

(Continua)

Id	Autor	Título	Ano
532	Francisco Barros Lobo	*O Tio João Gil*	1906
535	Francisco da Fonseca Benevides	*No tempo dos franceses*	1908
548	Franklin Távora	*O Matuto*	1878
549	Franklin Távora	*O Sacrifício*	1879
583	Graça Aranha	*Canaã*	1902
605	Harry Laus	*Os papéis do coronel*	1995
613	Inácio Pizarro de Morais Sarmento	*O Engeitado*	1846
614	Inglês de Sousa	*O missionário*	1891
617	Jayme de Magalhães Lima	*Transviado*	1899
621	Joaquim Manuel de Macedo	*O moço louro*	1845
622	Joaquim Manuel de Macedo	*Os Dois Amores*	1848
624	Joaquim Manuel de Macedo	*A luneta mágica*	1869
625	Joaquim Manuel de Macedo	*As Vítimas-Algozes*	1869
626	Joaquim Manuel de Macedo	*As Mulheres de Mantilha*	1870
641	João José Grave	*A morte vence*	1916
654	José de Alencar	*Cinco Minutos*	1856
656	José de Alencar	*O Guarani*	1857
657	José de Alencar	*As minas de prata*	1862
659	José de Alencar	*Diva*	1864
661	José de Alencar	*A pata da gazela*	1870
662	José de Alencar	*O gaúcho*	1870
663	José de Alencar	*Sonhos dOuro*	1872
664	José de Alencar	*A alma de Lázaro*	1873
666	José de Alencar	*O Garatuja*	1873
667	José de Alencar	*Ubijarara*	1874
668	José de Alencar	*O sertanejo*	1875
669	José de Alencar	*Senhora*	1875
670	José de Alencar	*Encarnação*	1877
672	José do Patrocínio	*Mota Coqueiro*	1877
675	José Régio	*O príncipe com orelhas de burro*	1942
693	José da Silva Mendes Leal	*Infaustas Aventuras de Mestre Marçal Estouro: Vítima duma paixão*	1862
694	Júlio César Machado	*A vida em Lisboa*	1858
705	Júlio Dinis	*A Morgadinha dos Canaviais*	1868
707	Júlio Dinis	*As apreensões de uma mãe*	1870
708	Júlio Dinis	*Justiça de Sua Majestade*	1870
710	Júlio Dinis	*Os novelos da tia Filomela*	1870
711	Júlio Dinis	*Uma flor de entre o gelo*	1870
713	Júlio Dinis	*Os Fidalgos da Casa Mourisca*	1871
717	Julia Lopes de Almeida	*A falência*	1901
719	Julia Lopes de Almeida	*A Intrusa*	1905
720	Júlio Lourenço Pinto	*Margarida*	1879
723	Lima Barreto	*O subterrâneo do morro do castelo*	1905
724	Lima Barreto	*Recordações do escrivão Isaías Caminha*	1909
729	Lima Barreto	*Clara dos anjos*	1948
733	Artur Lobo de Ávila	*A descoberta e conquista da Índia pelos portugueses: romance historico*	1898
735	Lopo de Sousa	*Herança de lágrimas*	1871
737	Luciano Cordeiro	*A senhora duquesa*	1889
738	Lucas José de Alvarenga	*Statira, e Zoroastes*	1826
747	Luís Filipe Silva	*O futuro à janela*	1991
750	Luís Magalhães	*O Brasileiro Soares*	1886

(Continua)

Id	Autor	Título	Ano
781	Joaquim Maria Machado de Assis	*Os trabalhadores do mar*	1866
800	Joaquim Maria Machado de Assis	*Oliver Twist*	1870
810	Joaquim Maria Machado de Assis	*Ressurreição*	1872
825	Joaquim Maria Machado de Assis	*A Mão e a Luva*	1874
841	Joaquim Maria Machado de Assis	*Helena*	1876
858	Joaquim Maria Machado de Assis	*Iaiá Garcia*	1878
867	Joaquim Maria Machado de Assis	*Memórias póstumas de Brás Cubas*	1881
907	Joaquim Maria Machado de Assis	*Casa velha*	1885
965	Joaquim Maria Machado de Assis	*Esaú e Jacó*	1904
977	Joaquim Maria Machado de Assis	*Memorial de Aires*	1908
982	S. de Magalhães Lima	*A senhora viscondessa*	1875
987	Manoel da Cruz Pereira Coutinho	*Elvenda, ou Conquista de Coimbra por Fernando Magno*	1858
995	Manuel de Oliveira Paiva	*A afilhada*	1899
1010	Marcelino Mesquita	*Os quatro reis impostores*	1908
1011	Maria O'Neill	*A Marquesa de Vale Negro*	1914
1013	Mário de Sá-Carneiro	*Loucura...*	1912
1014	Mário de Sá-Carneiro	*A Confissão de Lúcio*	1913
1018	Matilde Isabel de Santana e Vasconcelos Moniz Bettencourt	*O soldado de Aljubarrota*	1857
1019	João Baptista de Mattos Moreira	*Tempestades do Coração*	1867
1020	Maurícia C. de Figueiredo	*O exilado*	1900
1023	Maria Benedicta Mousinho de Albuquerque Pinho	*Marina: romance passional*	1912
1024	Melo de Matos	*Lisboa no ano 2000*	1906
1032	Miguel J. T. Mascarenhas	*Um conto português: episódio da guerra civil: a Maria da Fonte*	1873
1039	J.P. Oliveira Martins	*Febo Moniz*	1867
1040	Oliveira Mascarenhas	*O frade arrábido: romance histórico do século XVIII*	1881
1043	Othon Gama d'Eça	*Vindita braba*	1923
1063	Paulo Setúbal	*A Marquesa de Santos*	1925
1064	Paulo Setúbal	*O Príncipe de Nassau*	1925
1065	Paulo Setúbal	*Os irmãos Leme*	1933
1070	A.J. Pereira Varela	*Os miseráveis da aristocracia*	1864
1074	Manuel Pinheiro Chagas	*Astúcias de namorada*	1873
1077	Manuel Pinheiro Chagas	*A Lenda da Meia-Noite*	1906
1078	Pedro José Supico de Morais	*O mundo no ano 3000*	1895
1079	Policarpo da Silva	*O piolho viajante: Viagens em mil e uma carapuças*	1802
1085	Raúl Brandão	*A Farsa*	1903
1087	Raúl Brandão	*Os Pobres*	1906

(Continua)

Id	Autor	Título	Ano
1088	Raúl Brandão	*Húmus*	1919
1099	Raúl Brandão	*O Pobre de Pedir*	1931
1101	Raul Pompeia	*As jóias da Coroa*	1882
1102	Raul Pompeia	*Uma tragédia no Amazonas*	1882
1136	Tomaz de Melo	*O Conde de S. Luís*	1874
1143	Virgínia de Castro e Almeida	*Decameron*	1916
1144	Virgínia de Castro e Almeida	*Inocente*	1916
1145	Virgínia de Castro e Almeida	*O Solar dos Pavões*	1916
1146	Virgínia de Castro e Almeida	*A história de Dona Redonda e da sua gente*	1941
1151	Virgílio Várzea	*Rose-Castle*	1893
1152	Virgílio Várzea	*A noiva do paladino*	1901
1154	Virgílio Várzea	*O brigue flibusteiro*	1904
1155	Visconde de Taunay	*Inocência*	1872
1159	Visconde de Villa-Moura	*Nova Sapho: Tragedia Extranha*	1911