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

Pais, filhos e outras relações familiares no DIP

Fathers, sons and other family relations in DIP

Cristina Mota
INESC-ID & Linguateca

Diana Santos
Linguateca & ILOS, UiO

Resumo

Neste artigo é descrita em pormenor a tarefa de identificação de relações familiares no Desafio de Identificação de Personagens (DIP), uma avaliação conjunta para identificar personagens em textos literários em português. Explicamos a motivação para esta subtarefa, e quais as dificuldades em criar uma coleção dourada com os valores corretos. Depois de referir em abstrato como se processa a avaliação desta sub-tarefa, relatamos os resultados do sistema participante, o PALAVRAS-DIP, e comentamos alguns problemas na sua avaliação. Além disso, descrevemos aquilo que aprendemos sobre a literatura lusófona com esta tarefa, assim como sugerimos outras pesquisas possíveis com este material.

Palavras chave

relações familiares, literatura lusófona, redes de personagens, identificação automática de relações

Abstract

In this paper we detail the relation identification extraction task of DIP, the character identification challenge in Portuguese. We describe the task and the choices made, explain how to evaluate it, and evaluate the results of the only participant, PALAVRAS-DIP. We also describe what we learned about lusophone literature with this task. Finally, we discuss further research with the compiled material.

Keywords

family relations, lusophone literature, character networks, automatic relation identification

1. Introdução

O desafio de identificação de personagens (DIP) foi uma avaliação conjunta organizada pela Linguateca, NuPILL-UFSC e Universidades do Maranhão e Oslo para estimular o desenvolvimento de sistemas que, dada uma obra completa, obtivessem as personagens, as formas de as identificar, características como o seu género e as suas profissões, ocupações e estatutos sociais, e — o que nos interessa no presente artigo — as relações familiares entre elas.

A própria avaliação conjunta foi descrita inicialmente em Santos et al. (2022) e mais especificamente em Santos et al. (2023). Aqui concentramo-nos na identificação de relações e nas várias formas de as caracterizar.

De qualquer maneira, é importante deixar claro que os sistemas teriam de entregar dois ficheiros: um que indicasse as personagens, a sua forma de serem chamadas, o seu género e a sua profissão, para cada obra. Nesse ficheiro uma identificação numérica era atribuída à personagem. E noutro ficheiro teriam de indicar as possíveis relações familiares entre essas personagens, usando os identificadores definidos antes.

A avaliação dos resultados dos sistemas seria feita comparando-os com uma coleção dourada (CD), ou seja, com as respostas certas previamente criadas pela organização, num subconjunto dos textos postos à disposição dos participantes. A organização do DIP criou uma coleção dourada para 40 obras, mais quatro usadas como exemplo, que é o que chamamos a coleção dourada total. Contudo, o PALAVRAS-DIP (Bick, 2023), o único sistema participante, apenas tratou as 100 obras em texto (e não as 100 obras em pdf), o que significa que apenas pôde ser avaliado sobre 21 obras, o que chamamos a CD de texto.

A motivação para incluir esta informação foi a nossa convicção de que as relações familiares eram importantes para o enredo, e frequentemente mencionadas, na literatura, e que por isso valeria a pena vê-las em conjunto, através de uma leitura distante.

2. Que relações?

As relações familiares que os sistemas participantes deveriam extrair de uma dada obra literária são as que existissem entre personagens com nome (identificadas, portanto, numa fase anterior), entre as seguintes: *mãe, pai, filho, filha,*

DOI: 10.21814/lm.15.1.402

neto, neta, avó, avô, irmã, irmão, cunhado, cunhada, primo, prima, tio, tia, sobrinho, sobrinha, bisavó, bisavô, bisneto, bisneta, nora, genro, sogro, sogra, mulher, marido, padrinho, madrinha, compadre, comadre, afilhado, afilhada.

Devido a variadas grafias (*mãi* e *mãe*, *pae* e *pai*, etc.) nós normalizámos o nome das relações, não só graficamente, mas também lexicalmente, visto que existem muitas e variadas formas (sobretudo) de identificar um casal, como mencionado em Santos et al. (2023).

A escolha de adotar um vocabulário controlado das relações familiares teve o objetivo de facilitar o processo de avaliação dos sistemas participantes no DIP. Outra alternativa seria solicitar a identificação da forma exata pela qual estas relações são expressas na obra, mas isso exigiria muito provavelmente ma nova etapa de processamento (possivelmente manual), que mapeasse as relações identificadas pelos sistemas com as disponíveis na coleção dourada.

Como já referido em Santos et al. (2023), o vocabulário controlado de relações familiares adotado pelo DIP não cobriu todas as relações familiares existentes, faltando, por exemplo, relações como *padastro*, *madastra*, *enteado* e *enteada*. Além disso, e embora não estivessem na lista inicial, foram contempladas, e discutidas no ensaio, as relações *noivo* e *noiva* e *viúvo* e *viúva*.

3. Dificuldades na construção da coleção dourada

Embora a identificação das relações escolhidas pareça uma tarefa relativamente fácil, surgiram várias questões ao criar a coleção dourada, que gostaríamos de documentar aqui.

Em primeiro lugar, deveriam os anotadores também explicitar outras relações (biologicamente indiscutíveis, mas não expressas), tal como `X filho de Y` e `Y filho de Z` implica necessariamente `X neto de Z`?

Em segundo lugar, e essa questão já menos biologicamente determinada, mas muito mais frequente, se `X é marido de Y`, e `Y é mãe de Z`, deveriam os anotadores assumir, se nada fosse dito em contrário, que `X é pai de Z`?

A nossa resposta foi negativa em ambos os casos, mas é claramente uma opção discutível, à qual voltaremos mais tarde.

Em terceiro lugar, ao observar como as palavras do campo da família eram usadas nas obras, descobrimos que as formas de tratamento nem sempre são uma prova indiscutível duma relação familiar. É possível, por exemplo, tratar uma madrasta, e mesmo uma sogra, por *mãe*, assim como muitas personagens tratam pessoas mais novas sem qualquer vínculo familiar por *filhos* ou *filhas*.

Além disso, em alguns casos há informação incorreta (ou inconsistente) nos livros, ou por lapso do autor ou para indicar que a personagem em questão está equivocada — ou é ignorante. Veja-se o seguinte exemplo, em *Pero da Covilhã: Episódio Romântico do Século XV* de Zeferino Norberto Gonçalves Brandão: Catarina de Áustria exprime a seguinte queixa a seu cunhado D. Luiz, falando em relação ao seu marido, D. João III:

> Amor do povo e da patria como o nutriam em seus heroicos seios seu pai e avô Dom Manoel e Dom João!

Ora o pai deles era de facto Dom Manuel, mas D. João II, o rei anterior, era primo e não pai de Dom Manuel, e portanto não era avô de D. João III e de D. Luiz.[1]

O criador da CD tem de decidir (e eventualmente corrigir) a informação dada pela rainha, o que não é necessariamente fácil, até porque não seria de esperar erros destes num romance histórico.

Este é um tema que convém realçar: não é tão fácil quanto seria de esperar determinar as relações familiares entre as personagens de uma obra literária.

Finalmente, considerámos que não valia a pena escrever em duplicado relações que apareçam duas vezes, como em `X mulher de Y` e `Y marido de X`[2] ou `X irmão de Y` e `Y irmão de X`,[3] deixando ao sistema de avaliação a expansão automática de todos estes casos. Mas uma suposição, não tornada explícita na altura, era a de que o anotador da CD colocaria a primeira (muitas vezes única) vez que a relação era mencionada na obra. É aliás por isso que apresentaremos a panorâmica das relações na CD antes e depois da expansão.

[1] Este é um caso de romance histórico, em que portanto se podem confirmar as relações familiares corretas das personagens históricas, mas essas relações também se encontram corretamente especificadas noutras partes da mesma obra. Se o lapso foi propositado ou acidental, o que é certo é que não prejudica o enredo, mas prejudica certamente a extração automática das relações.

[2] O que significa que `mulher` e `marido` são relações inversas.

[3] Estas relações são simétricas, ou seja, a relação é igual à sua inversa — no caso das personagens terem o mesmo género, claro.

4. Relações identificadas no DIP

Depois destas explicações sobre a forma como encarámos a anotação das relações familiares, vejamos agora o panorama das relações encontradas através do DIP.

4.1. Encontradas na coleção dourada

O primeiro, e trivial, resultado, é que de facto nas 44 obras lidas atentamente, apenas uma não apresentava quaisquer relações familiares entre as personagens, nomeadamente *O Bom-Crioulo*, de Adolfo Caminha, o que valida a nossa intuição inicial de que este era um campo semântico recorrente na literatura.

Vemos na Figura 1 que as formas mais frequentes de descrever parentesco nas obras da coleção dourada eram *filho* e *filha*. Ao expandir (Figura 2), a relação familiar mais frequente tornou-se *pai*, o que deu origem ao título do presente artigo.

Duas conclusões podem ser tiradas: por um lado, é inegável a importância que a paternidade e a estrutura patriarcal têm (ou tinham) na literatura lusófona.[4] Por outro lado, dado que há muito mais personagens masculinas do que femininas, a posição de *mãe* é afinal mais significativa do que *pai*. Ou seja, a percentagem das mulheres que são mães é maior do que a dos homens que são pais.[5] Conforme comentado por Marcia Langfeldt, esta presença da mãe pode ser interpretada como mais uma prova da estrutura patriarcal, dado que na época da maior parte dos romances tratados no DIP ser mãe era uma profissão, senão a profissão da mulher.

Observando mais uma vez as relações familiares mais frequentes antes de expandir, pode também observar-se ser mais comum apresentar alguém (mulher) como mulher de outra pessoa, do que alguém (homem) como marido de outra.[6] Naturalmente, após a expansão, o número de maridos e de mulheres é o mesmo.

Uma coisa que, contudo, não podemos ainda concluir é qual a percentagem de personagens que está relacionada familiarmente com outras. Na próxima subsecção tratamos disso.

[4]Com o DIP, só nos podemos pronunciar sobre a literatura lusófona, mas é bem provável que isto seja uma constante da literatura ocidental.

[5]Visto que há 106 pais e 64 mães em 2813 homens e 830 mulheres, na coleção dourada total.

[6]Embora não possamos ter certeza absoluta de que não foram os próprios anotadores que introduziram este viés.

4.2. Uma visão de género através das relações familiares

Se observarmos todas as personagens identificadas na coleção dourada, 1075, e as classificarmos pelo número de relações familiares que apresentam, depois da expansão, vemos que existe uma diferença significativa entre personagens femininas e masculinas.

Nas Tabelas 1 e 2 vemos quantas relações por personagem foram encontradas, por personagem feminina e masculina, respetivamente.

Enquanto 80,3% dos homens não têm qualquer relação familiar com outras personagens, apenas 61,9% das mulheres estão na mesma posição "independente".

Além disso, o número médio de relações familiares de um homem, 0,29, é muito menor do que o de uma mulher, 0,73.

Num. de relações	casos
0	155
1	50
2	24
3	9
4	10
5	4

Tabela 1: Número de relações de personagens femininas

As personagens femininas com 5 relações familiares são Rosalina, de *O Cabeleira*, que tem uma irmã, um marido, e três enteadas;[7] D. Laura de *Amar, verbo intransitivo*, que tem marido e quatro filhos; Etelvina de *O Doutor Luís de Sandoval* com dois maridos, um pai, uma mãe e um filho, e finalmente Dona Glória, do *Dom Casmurro*, que tem um filho, um marido do qual fica viúva,[8] um irmão e um primo.

As personagens masculinas com 5 relações familiares são Gabriel, também de *O Cabeleira*, que tem mulher e irmão, dois filhos e um genro, e D. Affonso V, em *Pero da Covilhã: Episódio Romântico do Século XV*, que tem duas mulheres, uma irmã, uma sobrinha e um filho.

[7]Embora não estivesse na lista das relações a marcar, o/a anotador/a marcou na CD a relação de *madrasta*. Donde se conclui que deveríamos também ter feito um programa que verificava a lista de relações na CD, e a lista de relações na resposta dos participantes, para garantir que apenas as relações "oficiais" seriam avaliadas. Diga-se de passagem que o PALAVRAS-DIP também marca *amigo* e *amiga*, que retirámos automaticamente.

[8]E que portanto foi marcado tanto como mulher dele, como viúva dele.

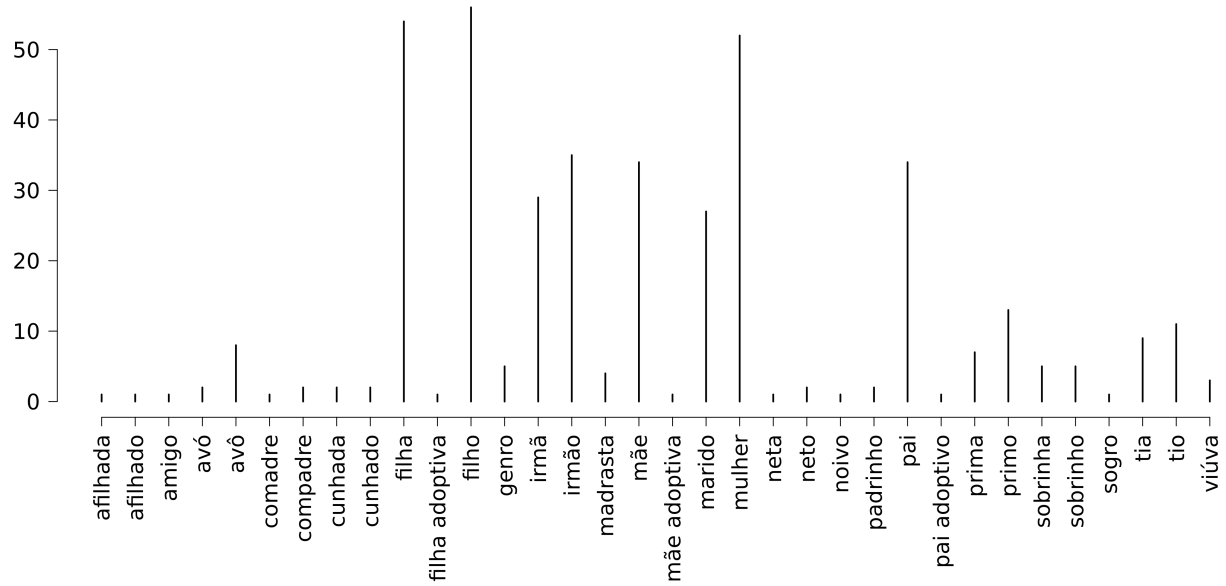

Figura 1: Distribuição das 414 relações em 43 obras

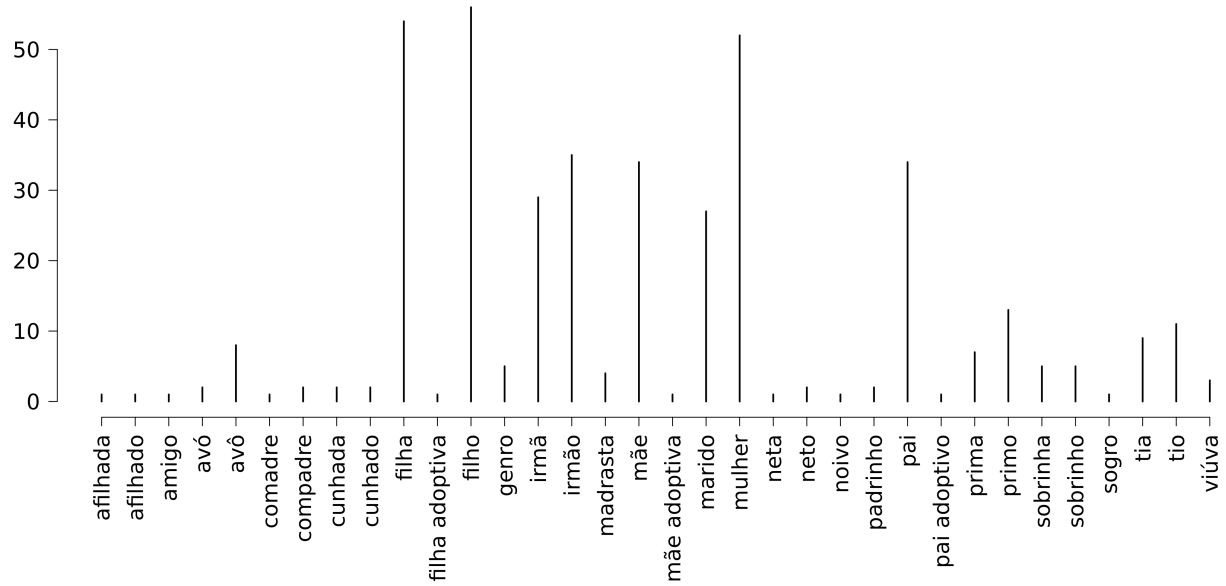

Figura 2: Distribuição das 811 relações após expansão

Num. de relações	casos
0	661
1	111
2	32
3	15
4	2
5	2

Tabela 2: Número de relações de personagens masculinas

4.3. Personagens principais e relações familiares

Vimos que é mais frequente que as personagens femininas sejam relacionadas familiarmente com outras personagens. Mas e se considerarmos simplesmente as personagens principais?

Para fazermos este estudo, e não tendo — como explicado em Santos et al. (2023) — feito diferenciação entre tipos de personagens, tivemos de usar uma operacionalização simples: A personagem com maior número de menções é a (ou uma das) personagens principais. Para isso

usámos a CD associada à coleção de texto (mais os três textos de exemplo), 24 textos, portanto, que pudemos investigar através do AC/DC (Santos, 2014).

Uma observação superficial permitiu confirmar que esta operacionalização parecia produzir resultados consonantes com a nossa impressão subjetiva das obras. No apêndice A apresentamos a personagem principal de cada obra obtida pelo método anterior, em que marcamos os poucos casos em que nos parece incorreta.

Obtivemos 5 obras com personagens principais femininas, e 19 com personagens principais masculinas. O número de relações familiares é muito maior nestes casos. De facto, apenas 4 personagens principais masculinas não tinham relações familiares com outras.

Em média, uma personagem principal feminina tem 2,20 relações com outras personagens, contra 1,35 relações de uma personagem principal masculina.

4.4. Relações extraídas pelo PALAVRAS-DIP

O panorama das relações extraídas das obras pelo PALAVRAS-DIP, apresentado na Figura 3, também foi submetido a um processo de expansão, embora o programa calcule automaticamente algumas relações inversas.

A situação ainda é mais flagrante em relação a *pai*, que é duas vezes mais frequente do que *mãe*. *Filho* é a segunda relação mais frequente identificada.

4.5. Obtidas pelo PALAVRAS-DIP na coleção extra

Como explicado em Santos et al. (2023), pedimos ao PALAVRAS-DIP para identificar as personagens noutra coleção maior, chamada coleção extra e descrita no texto referido, para podermos ter uma visão mais global da literatura lusófona. Os resultados apresentam-se na Figura 4.

Embora a avaliação tenha sido feita alguns meses após o DIP, e portanto com uma versão diferente do PALAVRAS-DIP, os resultados não são significativamente diferentes: *pai* continua a ser a relação mais frequente, *filho* continua a ser mais frequente do que *filha*, e *irmão* mais do que *irmã*, o que não admira sabendo que há mais personagens masculinas que femininas na coleção.

Seja como for, e visto que — ao contrário de outras características — os resultados do PALAVRAS-DIP não foram especialmente bons em relação à identificação das relações, não tiramos muitas conclusões destes dados.

096,4,mulher,1	Dona Laura mulher de Sousa Costa
096,5,irmão,6	Carlos irmão de Maria Luísa
096,4,mãe,5	D. Laura mãe de Carlos
096,4,mãe,6	**D. Laura mãe de Maria Luísa**
096,7,irmã,5	**Aldinha irmã de Carlos**
096,7,irmã,6	Aldinha irmã de Maria Luísa
096,7,filha,4	Aldinha filha de D. Laura
096,8,filha,4	**Laurita filha de D. Laura**
096,8,irmã,7	Laurita irmã de Aldinha
096,8,irmã,6	Laurita irmã de Maria Luisa
096,8,irmã,5	Laurita irmã de Carlos

Tabela 3: O conteúdo da CD: a negrito as relações que o sistema identificou, as outras estão em falta

096,1,filha,9	**Aldinha filha de D. Laura**
096,1,filha,7	Aldinha filha de Carlos
096,7,pai,1	Carlos pai de Aldinha (=ant)
096,1,irmã,7	**Aldinha irmã de Carlos**
096,7,irmão,1	Carlos irmão de ALdinha (=ant)
096,4,pai,9	não há 4
096,9,filha,4	não há 4 (=ant)
096,7,marido,9	Carlos marido de Dona Laura
096,24,filha,7	Laurita filha de Carlos
096,24,filha,9	**Laurita filha de D. Laura**

Tabela 4: O resultado de um sistema: a negrito as relações que o sistema identificou, as outras são espúrias

5. Avaliação da identificação das relações

Como explicado no artigo Willrich & Santos (2023), usámos como medida de avaliação a medida F, contando as relações certas, as em falta e as espúrias.

Contudo, a simplicidade da medida esconde vários pormenores complicados e não necessariamente satisfatórios — ou cabalmente resolvidos –, que tentaremos ilustrar aqui, com este exemplo concreto, relativo ao romance *Amar, verbo intransitivo*, de Mário de Andrade.

As relações anotadas na CD para esta obra são apresentadas na Tabela 3. A primeira coluna da tabela mostra o conteúdo da CD enquanto a segunda é a nossa tradução, tendo identificado as personagens a que se referem os identificadores.

Imaginemos agora que um sistema tinha produzdo o seguinte resultado, na Tabela 4.

Em ambas as tabelas colocámos já o resultado da avaliação. Na Tabela 3, vemos que há

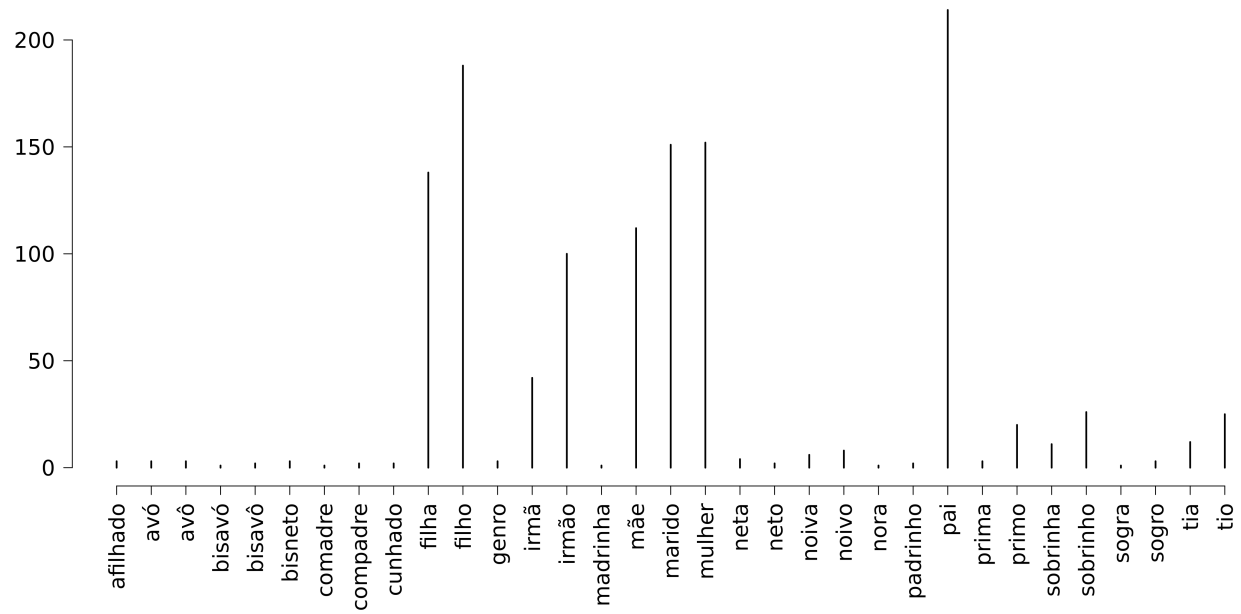

Figura 3: Distribuição das 1245 relações encontradas pelo PALAVRAS-DIP (expandidas) em 100 obras processadas

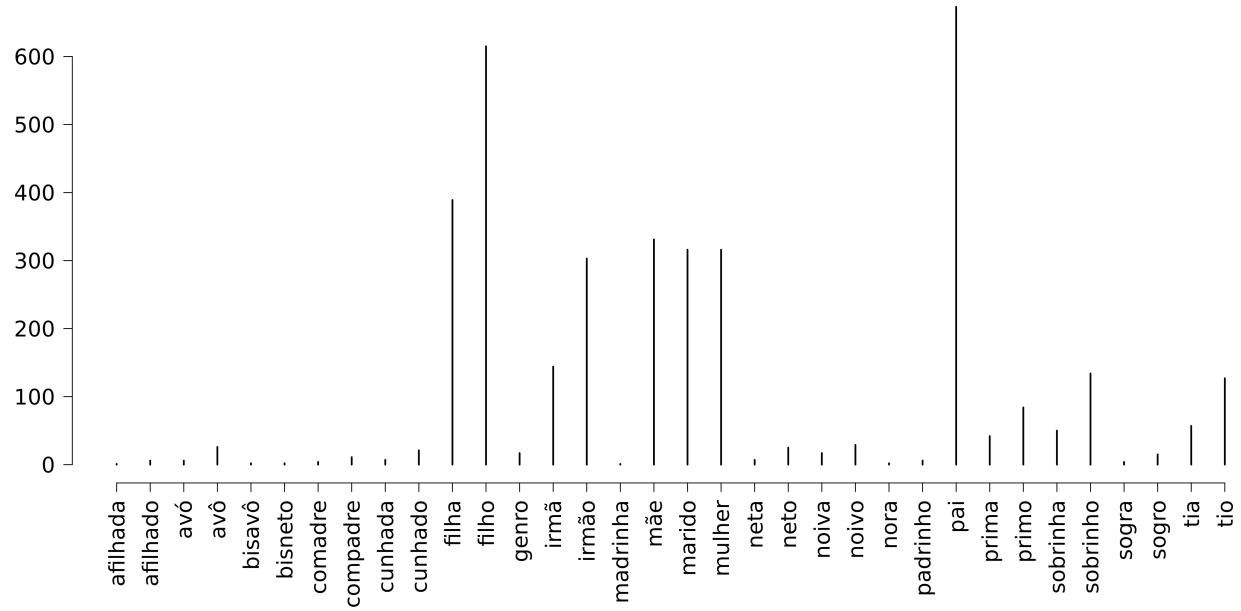

Figura 4: Distribuição das 3791 relações encontradas pelo PALAVRAS-DIP (expandidas) em 213 obras processadas

11 relações. O sistema identificou 3, e há portanto 8 relações que faltam. Ao expandir, ficamos com 16 que faltam e 6 certas, donde a abrangência é de 6/22=0,2727.

Na Tabela 4 — em que, note-se só contámos uma vez quando as relações e as suas inversas foram apresentadas pelo sistema –, há 7 relações: 3 certas (identificadas corretamente pelo sistema) e 4 espúrias, donde a precisão, depois de expandir, é de 6/14=0,4286.

A medida-F é, então, de 0,333.

O problema desta análise, que não é visível olhando apenas para estas tabelas, é o facto de o sistema ter amalgamado numa mesma personagem duas que além disso estavam relacionadas familiarmente, nomeadamente Carlos e o seu pai, como se pode ver na Figura seguinte, relativa às personagens:

```
096,7,Carlos Alberto Sousa Costa|Carlos|
Senhor Costa|Senhor Sousa Costa|Sousa
Costa|senhor Sousa Costa,M,palhaço|
filósofo|artista
```

Na CD, há uma personagem, Carlos, com os nomes `Carlos Alberto Sousa Costa|Carlos` e outra (o pai), com os nomes `Senhor Costa|Senhor Sousa Costa|Sousa Costa|senhor Sousa Costa`.

Para avaliar, os programas de avaliação tentam alinhar as personagens "reais" com as propostas pelo sistema, mas têm de escolher uma interpretação. E por isso, embora o sistema tivesse identificado corretamente que Sousa Costa era pai de Aldinha, e que Carlos era irmão de Aldinha, não pudemos considerar ambas as relações corretas, visto que Carlos e o pai tinham sido considerados a mesma personagem. Pior ainda, temos de considerar algumas certas e outras erradas, penalizando duramente o sistema.

Outra questão que este exemplo suscita é a arbitrariedade do conteúdo da coleção dourada. A mesma família (ou melhor, os mesmos graus de parentesco) podiam ter sido exprimidos mais completamente. Afinal se `A é irmã de B` e `B é irmão de C`, A e C também são irmãos (repare-se que não considerámos no DIP os meios-irmãos, que complicariam certamente estas deduções).

Mas se tivéssemos colocado mais uma relação de irmãos, e o sistema não, estaríamos a piorar a sua abrangência. Pior ainda, se o sistema tivesse colocado essa relação (certíssima) e não estivesse na CD, seria penalizado com a suposição de que a relação era espúria.

Isso poderia levar a crer que deveríamos sempre ter calculado a expansão total (o fecho transitivo) das relações, tanto na coleção dourada como no resultado do sistema a avaliar, para evitar estes problemas.

Contudo, estaríamos a penalizar muitíssimo um sistema que não tivesse encontrado, por exemplo, uma relação de irmão, numa família de cinco irmãos. Em vez de faltar essa (e a sua inversa) faltariam 8 relações, diminuindo radicalmente a abrangência. E a medida de avaliação seria mais ou menos rigorosa dependendo da estrutura familiar do romance, algo que também não parece muito justificado.

Por isso mesmo, acabámos por deixar que a falta de identificação de uma dada relação apenas contasse 2, embora conscientes de que poderíamos em alguns casos penalizar o sistema.

A nossa conclusão é que a única maneira de produzir uma avaliação justa obriga a realizar o processo de forma semi-automática, em que os casos espúrios são analisados por uma pessoa, a fim de, se for caso disso, adicionar mais relações à CD.

6. A identificação de famílias

Até agora temos apenas falado e refletido sobre relações entre duas personagens (e a sua relação inversa). Mas um dos interesses óbvios da extração de relações familiares é estudar a estrutura das famílias, que é o que tentaremos fazer nesta secção, juntando a informação das relações obtidas.

Nas Figuras 5 a 8, apresentamos as ligações familiares no mar de personagens de cada obra, para quatro obras com características diferentes.

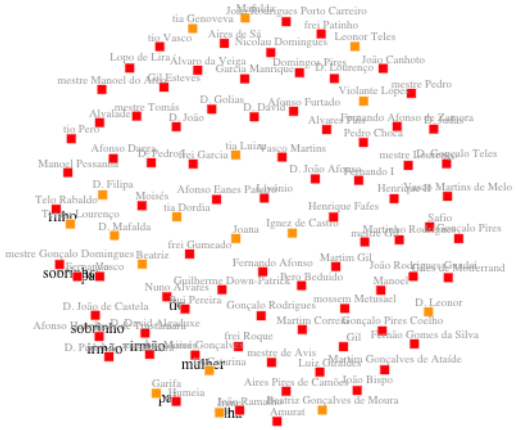

Figura 5: Relações entre as personagens de *Os tripeiros*

Os tripeiros, de António José Coelho Lousada, como romance histórico que é, tem muitas personagens, mas poucas relacionadas por parentesco, veja-se a Figura 5.

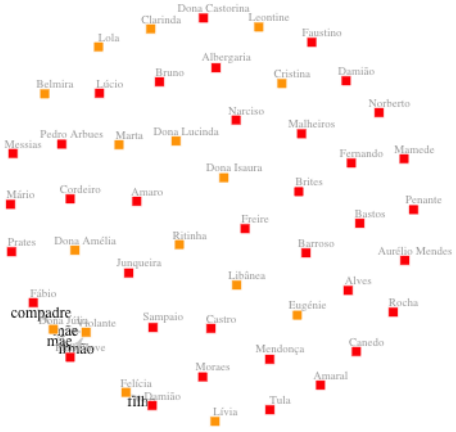

Figura 6: Relações entre as personagens de *Turbilhão*

Turbilhão, de Coelho Neto, por outro lado, tem um menor número de personagens, mas também poucas têm relações familiares entre si, cf. a Figura 6.

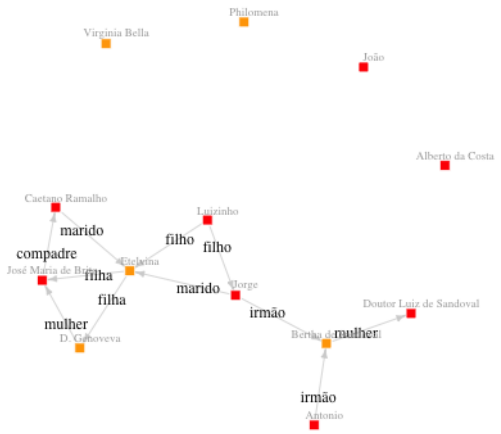

Figura 7: Relações entre as personagens de *O Dr. Luis Sandoval*

O Dr. Luis Sandoval, de Alice Pestana, pelo contrário, é um caso de pouquíssimas personagens, praticamente todas relacionadas entre si, como a Figura 7 ilustra.

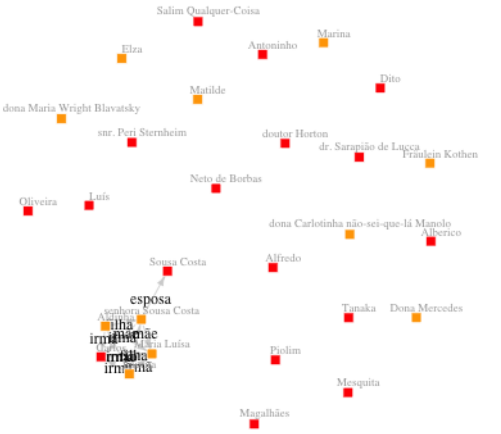

Figura 8: Relações entre as personagens de *Amar, verbo intransitivo*

Finalmente, *Amor, verbo intransitivo*, de Mário de Andrade, que é passado praticamente dentro de uma família apenas, é interessante notar na Figura 8 que existe um núcleo muito denso de relações, mas que se assemelha mais aos dois romances iniciais.

Se, pelo contrário, fizermos uma visualização em que apenas apresentamos as personagens relacionadas, podemos apreciar melhor a questão das diferentes famílias, e o tamanho destas. Vejam-se os seguintes casos:

Em *A ermida de Castromino*, de Teixeira de Vasconcelos, identificam-se três famílias, veja-se a Figura 9.

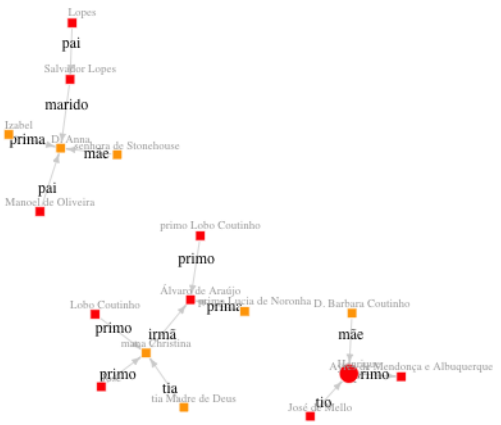

Figura 9: Relações entre as personagens de *A ermida de Castromino*

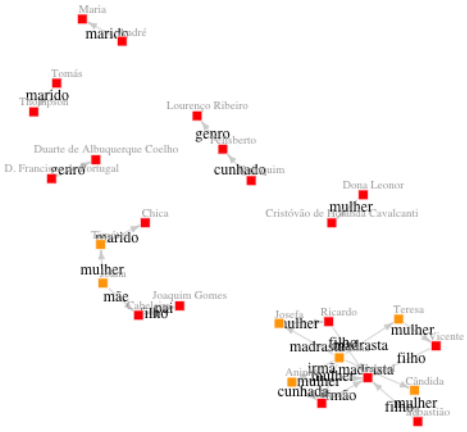

Figura 10: Relações entre as personagens de *Cabeleira*

Enquanto em *O Cabeleira*, de Franklin Távora, ilustrado na Figura 10, se descortinam não menos do que sete famílias diferentes, uma delas com dez elementos.

Em contraponto, apresenta-se novamente a única família de *Amar, verbo intransitivo*, na Figura 11, com seis elementos.

E para finalizar, na Figura 12 apresentamos as quatro famílias presentes em *Uma família inglesa*, de Júlio Dinis. Conhecido como é o enredo ser a relação entre as duas famílias Whitestone e Quintino, observa-se que outros dois relacionamentos, pouco relevantes para a obra, aparecem na figura como igualmente válidos, o que leva a concluir que seria definitivamente interessante refazer as figuras com alguma informação sobre a importância (e/ou frequência) que as diferentes personagens têm nas obras em questão.

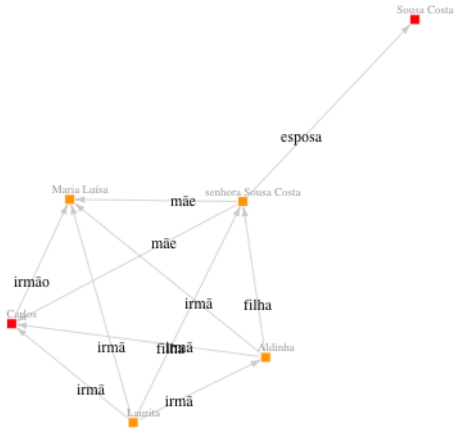

Figura 11: Relações entre as personagens de *Amar, verbo intransitivo*

Figura 12: Relações entre as personagens de *Uma família inglesa*

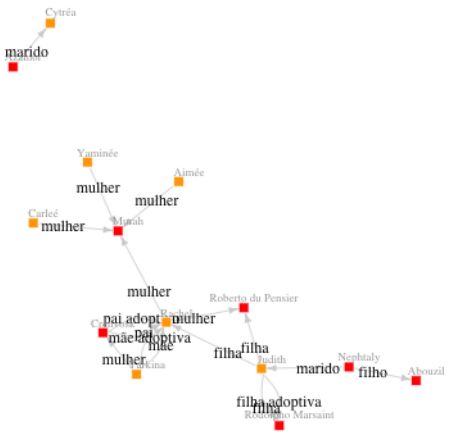

Figura 13: Relações entre as personagens de *A Judia Raquel*

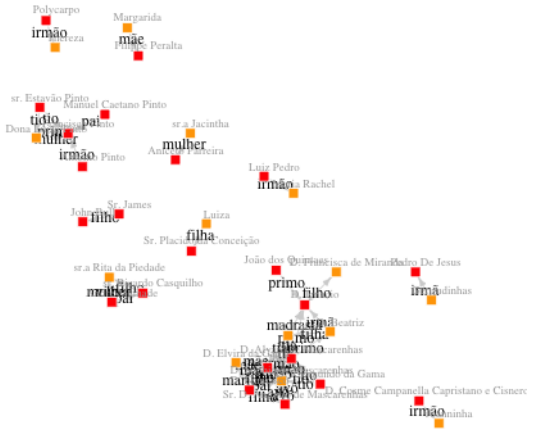

Figura 14: Relações entre as personagens de *Os cavaleiros da cruz vermelha*

Uma última distinção pode ser ilustrada com as famílias presentes em *A Judia Raquel*, de Francisca Senhorinha de Motta Dinis: uma família extremamente complexa e outra muito simples é o que podemos apreciar na Figura 13.

Compare-se com as 11 famílias dos 4 volumes de *Os homens da cruz vermelha*, de Carlos Pinto de Almeida, em que apenas uma é significativamente maior ou mais complexa do que as outras, na Figura 14.

Para mais figuras veja-se também a apresentação no Encontro do DIP (Mota, 2022).

7. Observações finais

Do aqui apresentado, podemos concluir o seguinte:

- A maioria das personagens não tem relações de parentesco com outras personagens, mas isso já não se verifica quando são personagens principais.
- A relação pai (e consequentemente a sua inversa, filho ou filha) é a que ocorre mais frequentemente entre personagens, sobrepondo-se à de mãe-filho/a ou mulher/marido.
- O número de famílias varia entre 1 a 11 (de acordo com a coleção dourada)
- A solução encontrada para avaliar as relações na primeira edição do DIP não é perfeita, e poderia ser preferível ter uma avaliação humana de casos que deveriam ser adicionados à coleção dourada das relações.

Um assunto extremamente importante não foi, contudo, ainda aqui abordado, nomeadamente: As relações familiares podem variar ao longo do tempo (em obras diferentes) ou no espaço temporal da obra — por exemplo, levando à união de duas famílias pelo casamento.

O que nos leva à seguinte autocrítica: No DIP quisemos apenas aceitar informação relativa à obra e não ao desenrolar do enredo, mas falhámos claramente nisso quando sugerimos identificar relações matrimoniais, visto que numa grande maioria dos casos as ditas fazem parte do enredo. Muitas vezes os protagonistas são solteiros e acabam por casar, ou casam no princípio e depois têm filhos, etc.

Seja como for, o que queremos sublinhar aqui é que deveríamos ter tentado distinguir esse estabelecimento de relações, parte integrante do enredo, dos casos em que essas relações permanecem as mesmas durante a obra toda.

Idealmente, deveríamos fazer uma nova revisão da coleção dourada para encontrar esses casos e marcá-los separadamente, por exemplo `X mulherENREDO Y` se o casamento fizer parte da história que se desenrola na obra.

O mesmo poderá fazer sentido nos casos — embora consideravelmente mais raros — em que há uma descoberta de relações familiares como parte da intriga,[9] que seria então marcado, por exemplo, `X filhoENREDO Y`.

8. Próximos passos

Nesta secção damos uma panorâmica de estudos ou projetos interessantes que poderiam ser feitos como continuação do trabalho descrito aqui.

Em primeiro lugar, a continuação (ou melhoria) do cruzamento da identificação das personagens principais com as relações familiares.[10] Fizemos isso apenas para 24 obras, mas poderíamos tentar fazê-lo para as 213+80 analisadas pelo PALAVRAS-DIP.

Por outro lado, até agora, nas redes que representámos, não existe qualquer informação sobre a importância das personagens, e assim não podemos saber realmente o que significa que a relação *pai* é a mais frequente: é porque são os pais os protagonistas, ou porque os protagonistas filhos são definidos/apresentados através da sua filiação?

A construção de redes que representassem não só as ligações familiares entre personagens mas também a sua importância relativa (medida através do número de vezes que eram mencionadas na obra) permitiria uma maior compreensão da importância ou não das relações. Isto é aliás prática corrente na construção de redes, como feito por exemplo em Santos & Freitas (2019).

De facto, outro trabalho (relacionado) que traria mais algum conhecimento sobre as relações familiares seria saber quantas vezes os familiares estão em presença uns dos outros — ou seja, fazer redes interacionais como as apresentadas em Santos & Freitas (2019) ou em Bick (2023).

Investigar as formas de tratamento entre personagens seria algo também extremamente interessante para compreender códigos culturais, identificando como é que as personagens se dirigem aos pais, aos filhos, e aos esposos.

Para isso, no entanto, seria preciso identificar os trechos em discurso direto, algo que também permite caracterizar melhor as obras. Por exemplo, obras com muito discurso direto também permitiriam investigar o protagonismo das diferentes personagens. Quem fala, e quem cala. Quem é abordado/mandado, e quem manda. Estas são chamadas redes conversacionais, e são das primeiras usadas na leitura distante em análise literária, propostas em Moretti (2011).

Finalmente, algo que pretendíamos fazer mas que ficou para trabalho futuro foi o uso de conceitos de redes (centralidade, conetividade, etc.) para caraterizar os diferentes romances e novelas.

9. Outros trabalhos de identificação de relações familiares

Terminamos este artigo com uma breve panorâmica de estudos relacionados. Não em relação a todas as possíveis redes de personagens possíveis de extrair de obras literárias, mas apenas daqueles (poucos) trabalhos que se dedicaram aos laços de parentesco. Veja-se Willrich & Santos (2023) para uma panorâmica de formas de avaliar a identificação de relações entre personagens e Abreu et al. (2013) para uma revisão da identificação de relações entre entidades em geral com um foco específico na aplicação dessa tarefa ao português.

No ReReLEM (Freitas et al., 2009), a tarefa de identificação de relações entre entidades mencionadas na avaliação do Segundo HAREM [11], algumas das relações identificadas eram de família, mas não foram tratadas de forma especial, sendo apenas um dos tipos de relação entre entidades, não sendo pedido para marcar especificamente o tipo de relação familiar.[12] Dos 10 sistemas par-

[9] Por exemplo, descobre-se o verdadeiro pai de uma protagonista.

[10] De facto, esse cruzamento seria interessante para todas as facetas estudadas no DIP, mas aqui limitamo-nos a considerar as relações de parentesco.

[11] https://www.linguateca.pt/HAREM/

[12] De facto, durante a avaliação, a relação familiar estava englobada na categoria OUTRA e só após a avaliação foi marcada especificamente juntamente com outras subcategorias num total de 22. Dessas, a relação familiar é a mais frequente a seguir à do vínculo institucional.

ticipantes no Segundo HAREM, apenas 3 participaram na pista do ReRelEM.

Em Higuchi et al. (2019) foi incluída a anotação das relações familiares no AC/DC para estudá-las no contexto da política brasileira, mas não limitámos a identificação dos laços de parentesco a políticos com nome, nem estudámos — até agora — a conetividade das redes familiares.

Em 2010, Santos et al. (2010) propuseram um sistema baseado em regras para identificar relações familiares não apenas entre entidade mencionadas mas também entre outras entidades mesmo que não referidas pelo seu nome próprio. Este sistema foi avaliado em dois corpos: um contendo as biografias de todos os reis portugueses encontradas na Wikipedia e outro composto por frases extraídas do CETEMPúblico (Rocha & Santos, 2000) que incluem um nome de relação. Numa primeira avaliação, a relação só era considerada correcta se os argumentos fossem exactamente iguais, mas numa segunda avaliação, os argumentos podiam ser parcialmente coincidentes. Este dois corpos foram anotados manualmente depois do sistema ter sido desenvolvido. Não é claro quais as relações familiares tratadas, mas foram identificadas pelos anotadores 105 relações no primeiro corpo e 21 relações explícitas em 110 frases no segundo corpo. Neste último caso, 89 frases incluem um nome de relação mas a relação familiar não está definida explicitamente entre duas personagens e como tal foram excluídas da avaliação.

Para o inglês, Azab et al. (2019) apresentam um novo modelo de palavras pulverizadas ("word embeddings") para representar personagens de filmes e as suas interações em diálogos, entrando em conta com os interlocutores. Esse modelo é utilizado para avaliar duas tarefas: identificação do grau de relação entre personagens ("character relatedness") e classificação da relação entre personagens ("character relation") — as relações familiares são classificadas de forma fina (e.g., pai/filho/tio/inimigo), grosseira (e.g., familar/social/profissional) e também quanto a sentimento (positivo/negativo/neutro). Este modelo foi avaliado em 31 obras de Shakespeare que fazem parte de um corpo literário com 109 peças anotado manualmente com recurso ao Mechanical Turk da Amazon (Massey et al., 2015) e que inclui 18 classes finas de relações, 4 classes grosseiras de relações e 3 classes de sentimento. Neste corpo foram anotadas 2170 relações das quais mais de 800 são relações de parentesco.

He et al. (2013) também identificam relações familiares e sociais (como de amizade) com o objectivo de construir redes de relações entre personagens cujos arcos entre personagens representam relações mútuas existentes entre essas personagens. Este modelo foi treinado na obra *Pride and Prejudice* de Jane Austen e testado adicionalmente nas obras *Emma* da mesma autora and *The Steppe* de Chekov (em inglês).

Agradecimentos

Agradecemos a Eckhard Bick muitas perguntas e críticas pertinentes, a Roberto Willrich várias referências relevantes e muitos comentários de melhoria ao próprio texto, e ao resto da organização do DIP pelo trabalho de equipa.

Agradecemos a Marcia Langfeldt vários comentários e sugestões para tornar os resultados mais relevantes para um público literário.

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

Abreu, Sandra Collovini de, Tiago Luis Bonamigo & Renata Vieira. 2013. A review on relation extraction with an eye on Portuguese. *Journal of the Brazilian Computer Society* 19. 553–571. 10.1007/s13173-013-0116-8.

Azab, Mahmoud, Noriyuki Kojima, Jia Deng & Rada Mihalcea. 2019. Representing movie characters in dialogues. Em *23rd Conference on Computational Natural Language Learning (CoNLL)*, 99–109. 10.18653/v1/K19-1010.

Bick, Eckhard. 2023. Extraction of Literary Character Information in Portuguese. *Linguamática* 15(1). 31–40. 10.21814/lm.15.1.397.

Freitas, Cláudia, Diana Santos, Cristina Mota, Hugo Gonçalo Oliveira & Paula Carvalho. 2009. Detection of relations between named entities: report of a shared task. Em *NAACL-HLT Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, 129–137.

He, Hua, Denilson Barbosa & Grzegorz Kondrak. 2013. Identification of speakers in novels. Em *51st Annual Meeting of the Association for Computational Linguistics*, 1312–1320.

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.

Massey, Philip, Patrick Xia, David Bamman & Noah A. Smith. 2015. Annotating character relationships in literary texts. *CoRR* abs/1512.00728. http://arxiv.org/abs/1512.00728.

Moretti, Franco. 2011. Network theory, plot analysis. *New Left review* 68. 80–102.

Mota, Cristina. 2022. Pais, filhos e outras relações no Desafio de Identificação de Personagens (DIP). Apresentação. https://www.linguateca.pt/aval_conjunta/dip/apr_encontro/Encontro_DIP_relacoes_Dec2022.pdf.

Rocha, Paulo Alexandre & Diana Santos. 2000. CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. Em *V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR)*, 131–140.

Santos, Daniel, Nuno Mamede & Jorge Baptista. 2010. Extraction of family relations between entities. Em *II Simpósio de Informática (INForum)*, 549–560.

Santos, Diana. 2014. Corpora at Linguateca: Vision and roads taken. Em Tony Berber Sardinha & Telma de Lurdes São Bento Ferreira (eds.), *Working with Portuguese Corpora*, 219–236. Bloomsbury.

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, 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.

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

A. Personagens principais na CD de texto

Só há três casos, sinalizados por ponto de interrogação, em que a nossa interpretação não concorda com o resultado quantitativo, e num deles, o caso da obra *O Ateneu*, a personagem principal é o narrador.

001-1	472	M	0	Agrippino Simões
002-1	91	M	2	António
004-32	148	M	0	João Bispo?
005-2	345	M	1	Diogo
006-2	340	M	2	Paulo
025-3	295	F	1	Helena
026-1	208	F	2	Simá
030-2	735	M	2	Amâncio
032-3	220	F	4	Etelvina
033-1	1134	M	3	Henrique
037-21	379	M	0	Fernão?
043-1	335	M	2	Cabeleira
047-3	395	F	2	Isaura
051-1	23	M	1	Luis
054-15	193	M	1	Eugénio
055-4	232	M	0	Amaro
064-1	150	M	4	Aristarco?
072-5	78	M	3	Dom Luiz
075-8	300	M	1	Pero da Covilhã
096-5	326	M	4	Carlos
099-3	1044	M	1	Carlos
201-1	698	M	1	Rubião
203-2	598	M	3	Daniel
204-13	344	F	2	Capitu

B. Famílias na CD total

ID	num. de famílias	tam. das famílias
001	1	2
002	1	3
004	8	3 3 2 2 2 2 2
005	3	3 2 2
006	2	4 2
025	2	2 2
026	4	3 3 3 2
030	3	3 3 3
032	1	9
033	3	7 6 4
037	2	7 3
043	7	10 5 3 2 2 2 2
047	2	4 3
051	1	2
054	6	5 3 3 2 2 2
055	0	
064	1	5
072	1	8
075	8	13 4 4 2 2 2 2 2
096	1	6
099	4	3 3 2 2
103	5	4 2 2 2 2
107	5	5 3 3 2 2
109	1	5
110	6	6 3 3 2 2 2
111	4	5 3 2 2
113	3	3 3 2
116	1	3
121	2	2 2
126	2	12 2
129	4	4 3 3 2
130	2	5 2
149	11	12 5 3 2 2 2 2 2 2 2 2 2
151	4	8 7 2 2
157	5	3 3 3 2 2
159	4	5 4 2 2
177	2	10 3
180	2	4 3
197	2	2 2
201	4	8 5 4 3
203	3	4 3 2
204	2	9 4