Estudando personagens na literatura lusófona

Diana Santos[1], **Cláudia Freitas**[2]

[1]Linguateca & ILOS, Universidade de Oslo
Pb 1003 Blindern, 0315 Oslo, Noruega

[2]Linguateca & PPGEL, PUC-Rio
Rua Marquês de São Vicente, 225 Rio de Janeiro, Brasil

d.s.m.santos@ilos.uio.no, claudiafreitas@puc-rio.br

Abstract. *In this paper we describe some studies and tools to deal with literary characters in Portuguese. After briefly describing the framework and the tools employed, we present character networks for some novels, and some preliminary attempts to characterize them.*

Resumo. *Neste artigo descrevemos alguns estudos feitos sobre personagens literárias em português. Após referir a infra-estrutura usada e as ferramentas utilizadas, apresentamos redes de personagens de algumas obras, e algumas tentativas de caracterização das mesmas.*

1. Introdução

Vários autores têm estudado as personagens de obras literárias no âmbito da leitura a distância. Assim, [Moretti 2011] criou redes de personagens do *Hamlet*, e [Grayson et al. 2016] fizeram o mesmo a romances chave da língua inglesa, de Charles Dickens e Jane Austen. [Klinger et al. 2016] estudaram por seu lado as personagens de duas obras de Kafka em termos de assinaturas emocionais para sete diferentes emoções, e [Bonch-Osmolovskaya and Skorinkin 2017] caracterizaram as personagens de Tolstoi em termos de papéis semânticos como agente, paciente, experienciador, possuidor, etc.

Pensamos que este é o primeiro artigo que relata a construção de redes de personagens de obras em português, embora já nos tenhamos debruçado sobre a caracterização de pessoas em textos literários, nomeadamente adjetivos associados a género em [Santos et al. 2018b].

2. O contexto

Este trabalho é feito no âmbito da Literateca [Santos 2019], um ambiente para estudar obras literárias em português usando a infraestrutura de corpos e anotação da Linguateca, mais especificamente o projeto AC/DC, que usa o PALAVRAS [Bick 2000] e outros mecanismos de anotação, e que contém, na sua versão 1.2 de 10 de julho de 2019, 730 obras completas de 167 autores diferentes.[1].

[1]https://www.linguateca.pt/acesso/corpus.php?corpus=LITERATECA

3. Pré-processamento: identificação das personagens

Uma personagem é geralmente referida de formas muito variadas numa obra. Por exemplo, a personagem Clara n'*As Pupilas do Senhor Reitor* de Júlio Dinis é também denominada pelos seguintes nomes próprios (ou entidades mencionadas): *Clarinha* e *Clarita*. E a personagem que dá o nome ao romance epónimo de Machado de Assis, *Helena*, também é referida por *Nhanhã Helena*, *D. Helena do Vale* ou apenas *D. Helena*. Um caso ainda mais difícil de obter automaticamente é a personagem de *Dom Casmurro* que aparece sob os nomes de *João*, *Pádua*, *Joãozinho*, *Sr. Pádua* e *Tartaruga*...

Por isso, uma das primeiras tarefas num estudo sistemático de personagens em obras literárias é identificar o conjunto dos nomes próprios que correspondem a uma mesma personagem, e definir uma maneira única de referi-las. [Grayson et al. 2016], por exemplo, criaram um dicionário de personagens, com uma única entrada para cada personagem e todos os seus nomes alternativos associados. Mas, além disso, é preciso garantir que um dado nome se refere à personagem certa.

Por outro lado, nem sempre é usado o nome próprio para referir uma personagem. Para terem uma identificação mais fiável das personagens em alemão, [Krug et al. 2018] marcaram todas as referências, que podem ser um nome comum, um pronome, e, no caso do português mesmo nada, visto que a língua portuguesa é uma língua de sujeito nulo: Numa investigação sobre a omissão do sujeito em português, considerando-se todos os romances, contos e crônicas de Machado de Assis, verificamos que quase 30% das frases têm o chamado *sujeito oculto* no verbo da oração principal – os números sobem para 40% considerando verbetes biográficos de uma enciclopédia e, num corpus jornalístico, 16% das frases têm sujeito oculto [Freitas et al. 2019]. Deste modo, em português, fica claro que há muito a perder quando não entramos em conta com essa questão.

4. População de um romance

Apesar de os trabalhos sobre personagens literárias tematizarem aquelas *principais*, *secundárias* ou *tipos*, é raro o texto que não inclua outros nomes próprios, referentes a entidades ficcionais partilhadas pela cultura, entidades históricas, e entes religiosos. A perspectiva das Humanidades Digitais é especialmente relevante para iluminar estes outros tipos de personagens, normalmente diluídos em uma história, mas que ganham corpo quando considerados parte de um acervo amplo. Veja-se a tabela 1 com essas quantidades (revistas) para obras diferentes:

Obra	tamanho em palavras	históricas	literárias	religiosas
Dom Casmurro	78606	36	13	72
Helena	66241	11	4	25
As Pupilas do Senhor Reitor	114343	9	16	176
Úrsula	54292	4	3	114

Tabela 1. Número de outras personagens mencionadas em quatro romances

Deus é de longe a entidade religiosa mais mencionada nos textos em português. Na tabela 2, comparamo-lo com menções ao diabo em romances, novelas e contos.

Deus	11.175	Diabo	3.308
Jesus	1.054	Satanás	135
Cristo	677	demo	113
Nosso Senhor	290	Lúcifer	49
Santo Deus	266	Belzebu	14
Nosso Senhor Jesus Cristo	260	Satã	13
Total	13.722	Total	3.632

Tabela 2. Distribuição de termos referentes a deus e ao diabo em 359 romances, novelas e contos de língua portuguesa

Deus (13.722) é assim invocado mais de três vezes mais do que o diabo (3.632), mas a variação entre autores também é relevante: Maria Peregrina de Sousa é quem faz (relativamente) mais alusões à divindade, seguida de perto por Maria Firmina dos Reis, enquanto que Olavo Bilac é quem mais se refere ao diabo, seguido de Artur Azevedo.

5. Redes de personagens

Com base na referência por nomes próprios, já unificados, criamos um programa que calcula a coocorrência numa janela deslizante de 3000 palavras (com sobreposição de 500 palavras), e que conta as vezes que duas personagens coocorrem nessa janela. Esses valores permitem-nos desenhar uma rede não dirigida, veja-se a figura 1, referente aos romances *As Pupilas do Senhor Reitor* e *Dom Casmurro*.

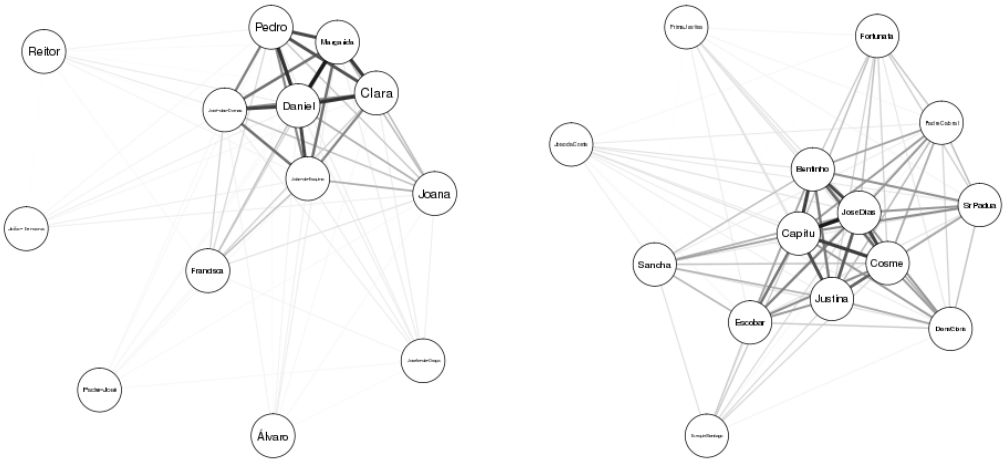

Figura 1. Redes de personagens

A simples visualização das redes relativas aos dois romances permite ver uma maior relação entre todas as personagens de Dom Casmurro comparada com as Pupilas do Senhor Reitor, que apresenta (aparentemente) cinco personagens pouco ligadas com o resto da trama.

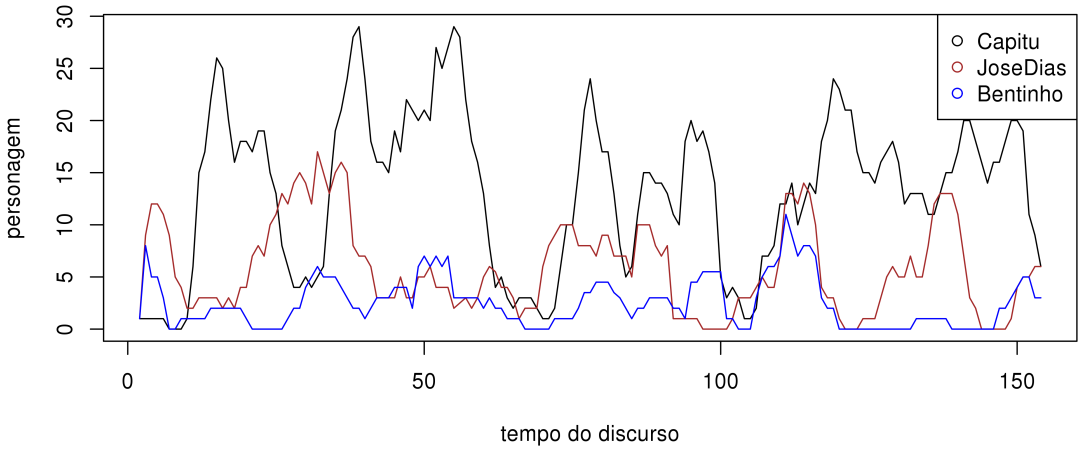

Figura 2. Algumas personagens de *Dom Casmurro* ao longo do tempo

6. Presença das personagens ao longo do enredo

Outro tipo de visualização que podemos fazer é a importância das personagens ao longo do livro, como a figura 2 mostra.

Quanto às ações desempenhadas pelas personagens, na esteira de [Bonch-Osmolovskaya and Skorinkin 2017, Archer and Jockers 2016], pesquisamos quais as ações das personagens (descritas por nomes próprios) nos quatro romances já mencionados, e pudemos constatar que a maior parte das ações eram comuns às personagens femininas e masculinas, embora as mulheres sorrissem mais e os homens respondessem mais.

7. Trabalho futuro

Pretendemos em breve associar sentimentos às personagens, como [Klinger et al. 2016], e polaridade à história, seguindo o exemplo de [Archer and Jockers 2016]. Para isso estamos a investigar as emoções em português e em texto literário, veja-se [Ramos and Freitas 2019, Santos and Simões 2019].

Ao termos revisado completamente a atribuição do género dos nomes próprios em todo o corpo OBras [Santos et al. 2018a], veja-se [Rocha et al. 2019], podemos fazer uma leitura mais distante das propriedades do género em toda a literatura a que temos acesso. Do mesmo modo, a explicitação dos sujeitos nos textos literários é um ponto que precisa ser tratado, a fim de evitar limitações das análises.

Pretendemos também obter redes sem revisão para estudar qual o mínimo de intervenção humana necessário para poder comparar centenas de obras, no âmbito da nossa filosofia de colaboração humana-máquina.

No âmbito da comparação entre várias línguas, poderemos, além do género das personagens olhar para as profissões mais mencionadas, como proposto em [Stanković et al. 2019].

Referências

Archer, J. and Jockers, M. L. (2016). *The Bestseller Code: Anatomy of the Blockbuster Novel*. Sr. Martins's Press.

Bick, E. (2000). *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. PhD thesis, Aarhus University, Aarhus, Denmark.

Bonch-Osmolovskaya, A. and Skorinkin, D. (2017). Text mining War and peace: Automatic extraction of character traits from literary pieces. *Digital Scholarship in the Humanities*, 32. Supplement 1.

Freitas, C., de Souza, E., and Rocha, L. (2019). Quantificando (e qualificando) o sujeito oculto em português. In *VI Jornada de Descrição do Português, STIL 2019*.

Grayson, S., Wade, K., Meaney, G., Rothwell, J., Mulvany, M., and Greene, D. (2016). Discovering structure in social networks of 19th century fiction. In *Proceedings of the 8th ACM Conference on Web Science*, pages 325–326. ACM.

Klinger, R., Suliya, S. S., and Reiter, N. (2016). Automatic emotion detection for quantitative literary studies. In *DH*.

Krug, M., Weimer, L., Reger, I., Macharowsky, L., Feldhaus, S., Puppe, F., and Jannidis, F. (2018). Description of a Corpus of Character References in German Novels - DROC [Deutsches ROman Corpus]. Technical report. DARIAH-DE Working Papers, Nr. 27.

Moretti, F. (2011). Network theory, plot analysis. *New Left review*, 68:80–102.

Ramos, B. and Freitas, C. (2019). "sentimento de quê?": uma lista de sentimentos para a análise de sentimentos. In *STIL 2019 - The 12th Brazilian Symposium in Information and Human Language Technology, Salvador, BA, Brazil, October, 15-18, 2019*.

Rocha, L., Freitas, C., and Santos, D. (2019). Recursos para leitura distante em português: diálogos entre pln e humanidades digitais. In *TILic 2019*.

Santos, D. (2019). Literature studies in Literateca: between digital humanities and corpus linguistics. In Doerr, M., Øyvind Eide, Grønvik, O., and Kjelsvik, B., editors, *Humanists and the digital toolbox: In honour of Christian-Emil Smith Ore*, pages 89–109. Novus forlag.

Santos, D., Freitas, C., and Bick, E. (2018a). Obras: a fully annotated and partially human-revised corpus of brazilian literary works in the public domain. OpenCor.

Santos, D., Freitas, C., and Lopes, J. M. (2018b). Ler e estudar a literatura lusófona como parte da literatura mundial: recursos para leitura distante em português. In Higuchi, S. and Ribeiro, C. J. S., editors, *I Congresso Internacional em Humanidades Digitais no Rio de Janeiro*, pages 375–383.

Santos, D. and Simões, A. (2019). Towards a computational environment for studying literature in Portuguese. In *DH Budapest 2019, Digital Humanities Conference*.

Stanković, R., Santos, D., Frontini, F., Erjavec, T., and Brando, C. (2019). Named entity recognition for distant reading in several european literatures. In *DH Budapest 2019, Digital Humanities Conference*.