Marcação de correferência para a caracterização de personagens em obras literárias em português

Diana Santos
Linguateca & Univ. de Oslo
Postboks 1003 Blindern
N-0315 Oslo, Noruega
d.s.m.santos@ilos.uio.no

Luisa Lima e Emanoel Pires
Univ. Estadual do Maranhão
Praça Duque de Caxias, SN
Morro do Alecrim. CEP: 65600-000
Caxias-MA, Brasil
{luisamaralima01,emanoeluema}@gmail.com

Resumo

Neste artigo descrevemos a adição de correferência a corpos literários públicos, para a tarefa de caracterização de personagens na leitura distante. Começamos por motivar essa tarefa na área dos estudos literários computacionais, explicamos a forma como tornamos essa tarefa legível e revisável a pessoas da área dos Estudos Literários, transferindo para o BRAT, descrevemos os primeiros resultados e um pequeno corpo anotado público, assim como discutimos a criação de dois módulos de correferência.

1 Motivação

Uma característica importante da análise narratológica é o conceito de personagem, essencial para caracterizar o enredo e os temas de uma obra literária. Em humanidades digitais, o conceito de leitura distante (Moretti, 2013) tenta "ler"muitas obras em vez de se cingir a apenas uma obra ou um autor, exigindo mecanismos que permitam atribuir e investigar automaticamente um grande número de personagens.

Em português, uma primeira tentativa foi feita através do DIP ("Desafio de identificação de personagens) (Santos et al., 2023), que permitiu uma visão panorâmica de uma quantidade de obras em português relativamente ao número de personagens, seu género, suas profissões ou ocupações e suas relações familiares com outras personagens.

Contudo, vale deixar claro que a literatura não reflete necessariamente a realidade sociológica do período em que foi criada, mas apenas a ideologia dos autores. Levantamentos como os feitos por Dalcastagnè (2021, 2025) mostram que o romance brasileiro contemporâneo, por exemplo, é marcado por fortes assimetrias de representação e por desigualdades estruturais que impactam a maneira como as personagens são construídas e reconhecidas.

Seja como for, ficou claro durante o processo de estabelecer e avaliar o DIP que usar apenas nomes próprios para designar as personagens constituía uma limitação significativa, visto que a maior parte das vezes as personagens eram referidas não por um nome próprio, mas por outras formas: pronomes, descrições definidas, ou implicitamente – também chamado pronomes ocultos na tradição gramatical brasileira. Assim, para chegar às informações relativas ao parentesco ou à profissão das personagens, o sistema PALAVRAS-DIP (Bick, 2023b), participante no DIP, usou um detetor de correferência para chegar ao seu resultado.

Para estudar personagens de forma mais pormenorizada, desde o estabelecimento de redes de personagens (Santos e Freitas, 2019) até à caracterização da sua presença no tempo do discurso, é preciso identificar uma personagem ao longo do texto, sem depender de esta ser mencionada por um nome próprio.

Um desses estudos, proposto e demonstrado por exemplo por Bamman et al. (2013), é o agrupamento de personagens de diferentes livros através do tipo de ações em que participam, do tipo de coisas que possuem, e das formas como são caracterizadas. Para isso, é preciso naturalmente identificar todas as ocorrências de caracterização, ação ou posse relacionadas com as personagens.

Bamman et al. (2014) sugerem uma cadeia de montagem ("pipeline") bastante sofisticada, para chegar à caracterização das personagens, incluindo (a) reconhecimento de entidades mencionadas (pessoas), (b) deteção de co-referência, (c) deteção das ações em que as personagens participam, e (d) várias formas de agrupar as mesmas.

Já existem bastante sistemas de identificação de Pessoas para a língua portuguesa – veja-se, por exemplo, os que participaram no HAREM (Mota e Santos, 2008)) ou a panorâmica feita em Claro et al. (2024) – e também de coreferência entre nomes próprios (Freitas et al., 2009), ou seja, várias nomes indicando a mesma personagem – essa informação, aliás, está pública para acesso no corpo

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 572–580

Literateca (Santos, 2018). Decidimos por isso apostar num sistema que permitisse alargar a identificação dessa correferência a casos pronominais e de sujeito nulo. Para o desenvolver, precisamos de construir um corpo (ou dataset) com dados para inspeção e treino, procedimento esse seguido por Krug et al. (2018) e por Bamman et al. (2020), que disponibilizaram recursos de textos literários anotados, respetivamente para o alemão e para o inglês.

É exatamente a construção desse conjunto de dados que é o foco do presente artigo, e que nos permite quantificar a questão do sujeito nulo e da correferência pronominal no caso das personagens literárias.

Começamos por abordar o problema dos sujeitos nulos, ou ocultos, na secção 2, visto que é uma característica da língua portuguesa e que não foi portanto tratada nas obras mencionadas para o inglês e para o alemão.

Depois explicamos a forma de proceder para anotar os sujeitos nulos e as menções anafóricas a personagens num pequeno corpo literário, descrevendo a forma de prosseguir (subsecção 3.2) e os critérios de anotação (subseccção 3.3). A secção 4 apresenta os variados resultados obtidos.

Na secção 5, finalmente, indicamos como pretendemos implementar um sistema automático baseado em regras, em código aberto, que poderá agilizar significativamente a marcação de personagens em corpos literários, e explicitar algumas características da correferência entre personagens, ao mesmo tempo que sugerimos desenvolver um sistema mais tradicional, baseado em aprendizagem automática ("machine learning").

2 Sujeitos nulos

Um fenómeno interessante e inescapável da gramática portuguesa é a questão dos sujeitos nulos. Um estudo preliminar (Martins e Freitas, 2019) sobre o assunto, analisando manualmente seis verbetes do DHBB (Higuchi et al., 2019), indicou que de 18% a 45% dos verbos naquela enciclopédia tinham sujeito nulo.

O foco neste artigo é o desenvolvimento de um sistema de correferência para o português, que permita a construção de uma cadeia de montagem semelhante, mas entrando em conta com as características especiais da gramática do português, em particular sujeitos nulos. Freitas et al. (2019), expandindo a contagem a mais dois géneros textuais nomeadamente jornalístico e literário, e alargando consideravelmente o número de textos, usaram uma ferramenta automática sobre dependências para identificar o número de verbos sem sujeito expresso, chegando a uma estimativa de 39,5%, 28,4% e 16% de sujeitos nulos para os géneros enciclopédicos (DHBB), literário – apenas Machado de Assis – e jornalístico, respetivamente. Finalmente, Freitas e Souza (2021) executam um trabalho ainda mais abrangente, desta vez sobre os corpos completos OBras (Santos et al., 2018) e DHBB, anotados com o UDpipe[1], para obter uma estimativa ainda mais confiável, e chegam à conclusão de que 41% dos verbos nos textos literários brasileiros constantes do OBRAS têm sujeito oculto.

Isso demonstra indiscutivelmente a necessidade de resolver estes casos, atribuindo explicitamente o sujeito, de forma a poder ter uma panorâmica mais completa das várias ações e estados descritos nos textos. Note-se, além disso, que o sujeito nulo é mais comum na variante de Portugal do que na variante brasileira (Kato et al., 2023), o que leva a concluir que a estimativa apresentada pode pecar por defeito se contarmos com toda a literatura em português presente na Literateca.

Contudo, ainda falta obter uma estimativa precisa em relação à importância para a caracterização das personagens destas duas questões: correferência e explicitação do sujeito nulo, o que faremos na secção 4.

3 Anotação

3.1 Três obras literárias

Escolhemos três obras literárias brasileiras do cânone, já no domínio público, mas pertencendo a três escolas literárias diferentes:

- *Úrsula*, de Maria Firmina dos Reis, 1859, obra romântica, com 53.568 unidades
- *O Cortiço*, de Aluísio de Azevedo, 1890, obra naturalista, com 95.472 unidades
- *Canaã*, de Graça Aranha, 1902, obra pré-modernista, com 90.167 unidades

As três obras já tinham sido objeto de definição e publicação das suas redes de personagens, ou seja, as personagens – com os variados nomes por que são chamadas nas obras – já tinham sido

[1] https://ufal.mff.cuni.cz/udpipe

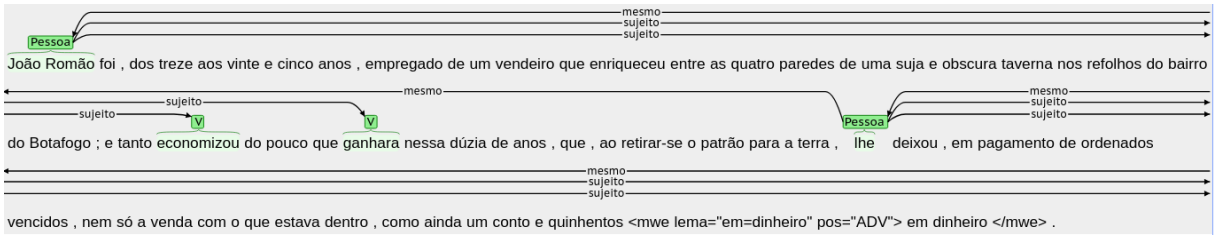

Figura 1: Anotação no BRAT

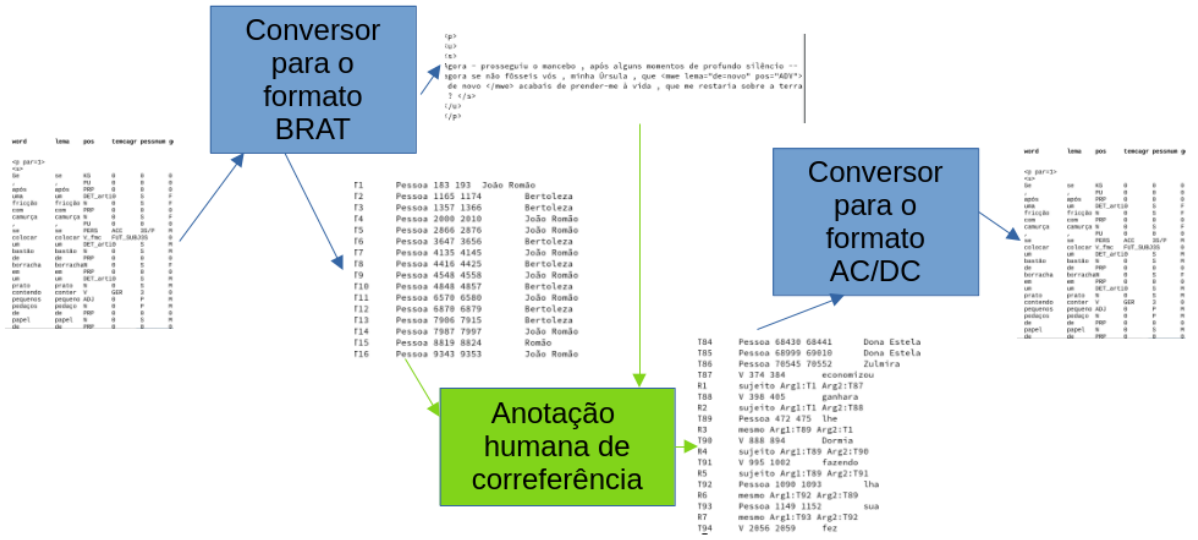

Figura 2: Conversão entre os vários formatos, antes e depois da anotação humana

organizadas[2], e as obras com a anotação são também públicas, em formato AC/DC, visto que estão incluídas no corpo OBras.

3.2 Processo de anotação

Nesta secção, apresentamos a tradução (bidirecional) do formato AC/DC (tabular) para um sistema pensado para simplificar a anotação e sua revisão para não-informáticos, nomeadamente o BRAT [3]. A Figura 1 apresenta uma tela deste sistema, com o início de *O Cortiço*.

O formato AC/DC (formato tabular, com colunas que marcam, por exemplo, uma dada personagem), foi traduzido para o formato BRAT, que conta o número de caracteres, e não o número de unidades (em inglês chamadas "tokens"). Esse formato BRAT – que são dois ficheiros: o texto simples, e as anotações – foi tornado remotamente acessível aos vários anotadores, que tinham autorização para anotar os ficheiros.

O resultado dessa anotação, que fica no servidor, é depois traduzido para uma nova coluna em formato AC/DC, e um novo corpo é criado com essas anotações. A Figura 2 ilustra o processo.

Isso permite, por um lado, que os anotadores usem a interface amigável do BRAT e, por outro, que os resultados da anotação possam ser procurados e manipulados através do AC/DC, com as potencialidades de um sistema de corpos, nomeadamente concordâncias e distribuição por valores de atributos.

Veja-se a Figura 3, em que pedimos para ver todos os casos em que a personagem `João Romão` é o sujeito de verbos de sujeito nulo em forma de concordância, e a lista desses mesmos verbos na Figura 4.

Será a partir destas listas (complementadas com os sujeitos explícitos, marcados automaticamente pelo PALAVRAS (Bick, 2000)) que tentaremos criar o perfil da personagem, e mais tarde compará-las com outras. Depois disso, passaremos ao passo em que classificaremos os verbos em categorias mais genéricas, para poder comparar e caracterizar as personagens, mas esse assunto não será detalhado aqui.

[2]Veja-se `https://www.linguateca.pt/Gramateca/Literateca/galeria.html`.

[3]`https://brat.nlplab.org/`

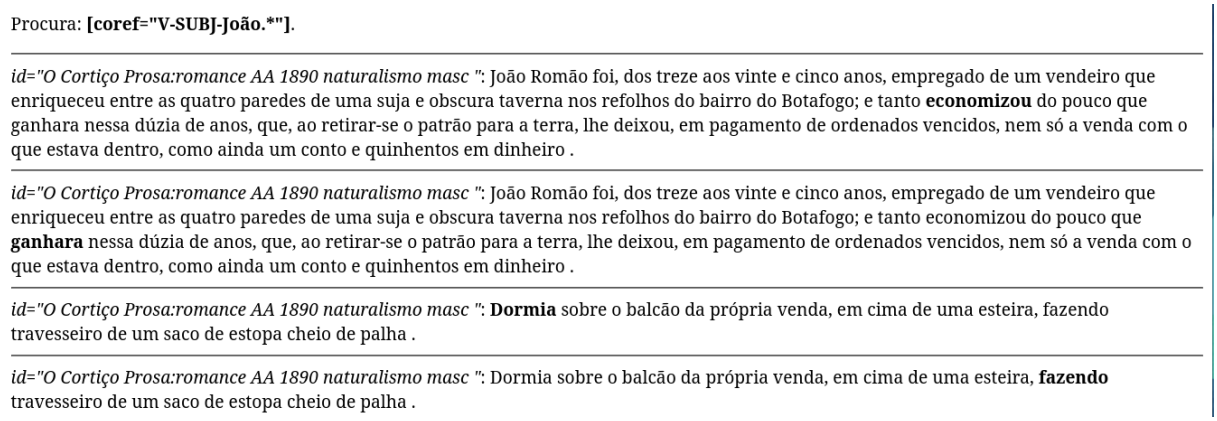

Figura 3: Procura, mostrando os primeiros casos em que a personagem João Romão corresponde ao sujeito (nulo)

Figura 4: Procura, mostrando os lemas dos verbos cujo sujeito nulo é João Romão

3.3 Critérios de anotação

Decidimos marcar manualmente os seguintes casos que correspondiam a anáfora:

- pronomes possessivos (p. ex. *seu*)
- pronomes pessoais (p. ex. *lhe*)
- formas de tratamento (p. ex. *sr. dr.*)

Não marcámos anáfora nominal, nem anáfora pronominal associada a nomes anafóricos. Por exemplo, vejam-se as três primeiras frases de *O Cortiço* de Aluísio Azevedo, em que os casos em negrito não são marcados, embora se refiram à personagem João Romão. Os casos em itálico, em contrapartida, são marcados:

> João Romão foi, dos treze aos vinte e cinco anos, empregado de um vendeiro que enriqueceu entre as quatro paredes de uma suja e obscura taverna nos refolhos do bairro do Botafogo; e tanto *economizou* do pouco que *ganhara* nessa dúzia de anos, que, ao retirar-se o patrão para a terra, *lhe* deixou , em pagamento de ordenados vencidos, nem só a venda com o que estava dentro, como ainda um conto e quinhentos em dinheiro .
>
> Proprietário e estabelecido por **sua** conta, o **rapaz** atirou-se à labutação ainda com mais ardor, **possuindo**-se de tal delírio de **enriquecer**, que **afrontava** resignado as mais duras privações.
>
> *Dormia* sobre o balcão da própria venda, em cima de uma esteira, *fazendo* travesseiro de um saco de estopa cheio de palha.

Não marcámos catáfora, ou seja, apenas marcámos os casos que se encontravam no texto a seguir à ocorrência do nome próprio. O que significa que em alguns livros, muitas das acções iniciais das personagens não são cobertas por esta identificação. Por exemplo a personagem Tancredo da obra *Úrsula* de Maria Firmina dos Reis só é identificada pelo nome no quarto capítulo.

Além disso, personagens cujo nome próprio nunca é fornecido, como o menino com quem Milkau fala no primeiro capítulo da obra *Canaã* de Graça Aranha, não podem ser estudadas com esta marcação.

Também não marcámos casos em que a personagem faz parte de um coletivo, veja-se por exemplo o seguinte trecho de *Canaã*, em que o sujeito plural se refere às duas personagens Milkau e Lentz, e em que se perdem portanto muitos sujeitos nulos (11) que caracterizariam as ações de Milkau.

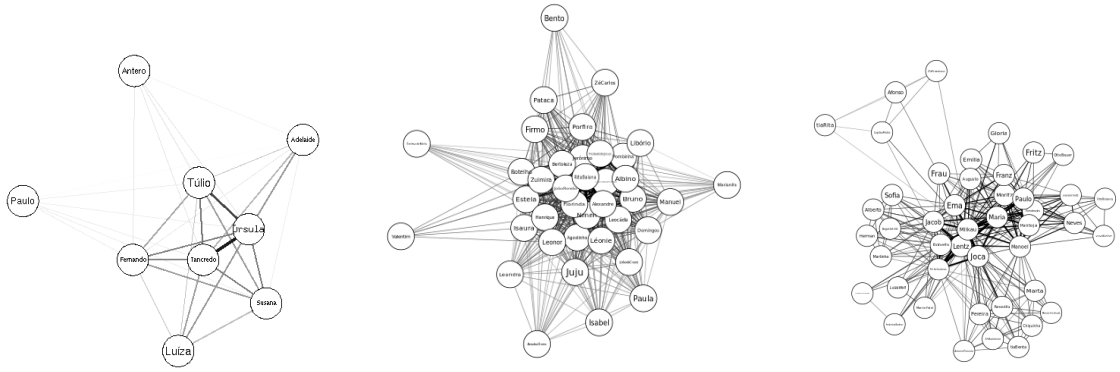

Figura 5: As redes de personagens das três obras, retiradas de https://www.linguateca.pt/Gramateca/Literateca/galeria.html

> Passearam ainda algum tempo, sentindo uma entranhada dificuldade em abandonar aquele lugar.
>
> Dirigiram os passos para os caminhos que abeiravam Santa Teresa.
>
> Procuravam as pequenas elevações, giravam abaixo e acima pelo parque, paravam à porta das casas, miravam atentos o serviço que nelas se fazia, sorriam às crianças, e, perseguindo com olhos de admiração as saudáveis raparigas, enrubesciam-nas.

Finalmente, marcámos sujeitos nulos quando a ação descrita tinha como sujeito a personagem analisada, mas não objetos nulos. É interessante que os objetos nulos são mencionados, como elipse, por Silva (2011, pag. 31), mas não os sujeitos nulos.

Todos estes casos podem vir a ser objeto de uma marcação posterior, se for considerado relevante, mas em primeiro lugar quisemos desenhar uma tarefa mais simples para ser feita e avaliada computacionalmente.

4 Resultados da anotação

Neste momento temos 6 personagens anotadas nas 3 obras (ver Tabela 1), anotação essa que disponibilizada publicamente, nos dois formatos: BRAT e tabular.[4] Para ver essas personagens nas redes de personagem das obras, criadas pela Linguateca, apenas usando os nomes próprios, veja-se a Figura 5.

Contudo, essa anotação já nos permite ver a importância que os dois fenómenos que estamos a anotar representam, para a identificação – e consequente caracterização – das personagens nas obras literárias. Os valores estão na Tabela 2.

Personagem	Próprio	Nulo	Pron	Total
João Romão	134	617	343	1094
Rita Baiana	115	262	207	584
Úrsula	228	322	673	1223
Tancredo	111	167	441	719
Milkau	349	220	294	863
Maria	145	129	242	516

Tabela 1: Menções das personagens: "Próprio"significa nome próprio, que pode ser múltiplo, "Nulo"significa que a personagem é o sujeito (nulo) de um dado verbo, e "Pron"significa que um elemento pronominal (pronome pessoal, possessivo, ou outro) se refere a essa mesma personagem.

PROP	1.082	21,6%
nulo	1.717	34,3%
PRON	2.200	44,1%

Tabela 2: Distribuição das instâncias de personagens nas obras em português

A Tabela 3 reproduz a tabela encontrada em Bamman et al. (2020) relativa às obras em inglês.

PRON	15.816	54,3%
NOM	9.737	33,5%
PROP	3.550	12,2%

Tabela 3: Distribuição das instâncias de personagens em Bamman et al. (2020)

É interessante notar que as duas línguas diferem consideravelmente, o que indica que este trabalho

[4] https://www.linguateca.pt/Gramateca/Literateca/Corref.html

Úrsula	meu minha minha me minha te me te teu tuas te me Meu te te te tu Tancredo
Cortiço	tu la sua ela seu dela na Rita
Canaã	senhor lhe suas eu senhor senhor lo o Milkau

Tabela 4: Um exemplo de cadeia anafórica (invertida) de cada obra. De notar que "senhor"corresponde à forma de tratamento *o senhor*

não poderia ser "importado"ou "adaptado"do inglês. Ou seja, a questão dos sujeitos nulos em português tem quase tanta importância (quantitativa) quantp a anáfora pronominal, que é o principal caso de menção a personagens em inglês.

Outra coisa que podemos calcular depois da marcação manual é o tamanho das cadeias anafóricas empregues pelos autores. Apresentamos na Tabela 4 alguns exemplos destas.

E na Figura 6 mostramos a variação nas três obras (referente às duas personagens de cada obra).

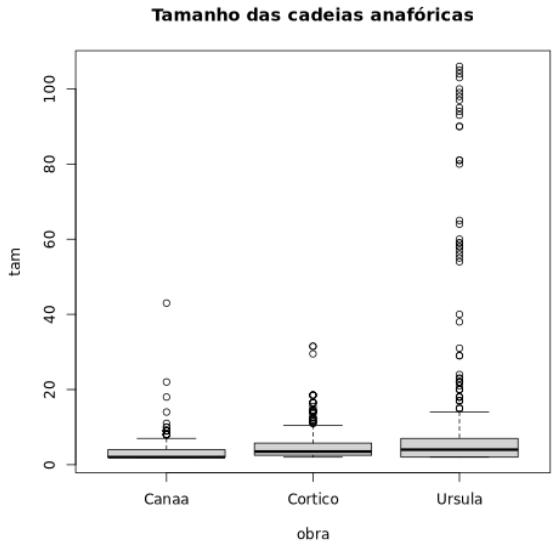

Figura 6: Número de elementos numa cadeia anafórica em cada obra

Embora três obras não sejam evidentemente suficientes para avançarmos generalizações, que terão de ser verificadas num conjunto muitíssimo maior, podemos salientar simplesmente:

- Em casos com dois nomes (em *O Cortiço*), o masculino é quase sempre usado completo (*João Romão*), enquanto o feminino é quase metade das vezes encurtado para o primeiro nome (*Rita Baiana* para *Rita*).
- Nas três obras uma das personagens é muito mais mencionada do que a outra (embora ambas tenham sido escolhidas por serem consideradas principais).
- Na obra escrita por uma mulher, é a personagem feminina a mais frequente. Nas duas obras escritas por homens, é a personagem masculina.

Em relação ao estudo das personagens, é nosso objetivo imediato desenvolver formas de as comparar baseadas nos verbos de que são sujeito, e nos objetos que "possuem", mesmo que ambos (verbos e substantivos) tenham de ser classificados em classes mais genéricas antes da comparação.

Mas com os dados que apresentamos aqui já podemos comparar, por exemplo, as três personagens masculinas em termos dos verbos mais frequentes de que são sujeitos, na Tabela 5.

Uma análise mais detalhada será publicada em Lima (2026).

5 Desenvolvimento de sistemas para a anotação automática de correferência

Para simplificar, usamos a partir de agora "correferência"como um termo abrangente para ambas as questões.

Com base nestes dados queremos desenvolver dois sistemas de correferência em paralelo:

- um baseado em regras, como uma extensão do usado no PALAVRAS-DIP (Bick, 2023b), ou melhor, inspirado por este
- outro baseado em aprendizagem automática, e usando (parte d)o material anotado para treino

Não queremos dar a ideia de que somos os primeiros a tratar de correferência para o português em geral; veja-se Fonseca et al. (2023) para uma panorâmica da situação. Aquilo que tentamos, e que é muito mais específico, é a resolução de correferência de personagens em texto literário.

5.1 Sistema de regras

Este sistema de regras será programado em VislCG3 – veja-se exemplos desse tipo de regras em Bick (2023a), e mesmo anteriormente em Bick (2010). A nossa escolha provém de já existir um

Tancredo		João Romão		Milkau		Úrsula		Rita Baiana		Maria	
ser	15	ter	29	dizer	30	ser	31	ser	17	ter	12
estar	11	fazer	26	ver	18	sentir	16	dar	14	ficar	11
exclamar	11	dar	21	ficar	16	ter	15	ter	14	ser	8
continuar	7	dizer	20	ir	16	amar	14	dizer	11	ver	8
ouvir	7	ser	20	sentir	15	ver	14	querer	9	sentir	7
amar	6	ir	18	estar	12	dizer	13	ir	8	dizer	6
fazer	6	ver	15	saber	11	estar	12	abrir	6	estar	6
sair	5	estar	14	ter	11	saber	12	estar	6	deixar	5
sentir	5	pensar	12	pensar	10	exclamar	10	saber	6	ouvir	5
ter	5	tomar	12	vir	9	voltar	10	sair	6	chegar	4
beijar	4	pôr	11	chegar	8	tornar	9	vir	6	fitar	4
escutar	4	sentir	11	ser	8	compreender	8	atirar	5	fugir	4
interrogar	4	deixar	10	responder	7	estremecer	8	fazer	5	passar	4
ver	4	mandar	9	observar	6	cair	7	meter	5	querer	4
	269		809		536		560		381		278

Tabela 5: Verbos mais frequentes de que são sujeitos cada uma das personagens. O número na última linha corresponde ao total de verbos de que as personagens são sujeito, totalizando 2.833 casos.

módulo do PALAVRAS-DIP (neste momento proprietário) que será assim melhorado e tornado público. De forma que, além de produzirmos um sistema, ele ainda pode ser ajustado por todos os que o queiram usar.

A vantagem deste sistema, evidentemente, é que as regras podem ser baseadas em questões gramaticais e literárias pensadas por um ser humano.

Os dados que compilámos e que apresentamos aqui podem ser usados para avaliar o sistema, que depois também será avaliado para outras personagens e outras obras, e a correção do seu desempenho poderá aumentar os dados anotados e públicos.

5.2 Sistema de aprendizagem automática

Por outro lado, tencionamos simplesmente alimentar um sistema de aprendizagem automática tradicional (usando os dados que descrevemos no presente artigo), e pedir para ele aprender regras que maximizem o seu desempenho.

Além disso, se usarmos um sistema que nos permita ter acesso às regras aprendidas, como por exemplo árvores de classificação (Baayen, 2008), podemos também inspecionar e eventualmente aprender com essas regras.

Contudo, a resolução de correferência em geral exige algum pós-processamento de dados anotados para que estes funcionem como entrada de um módulo automático. Fonseca et al. (2023) mencionam várias abordagens comuns: pares de menções , cadeias de entidades, etc. Sem entrar em grandes pormenores técnicos, o que é relevante aqui é que tais sistemas também necessitam de exemplos negativos, para cuja criação existem vários algoritmos, mas que são menos compreensíveis para os utilizadores típicos de leitura distante.

6 Considerações finais

Para avançar na leitura distante em português, é imprescindível termos um sistema de correferência que permita identificar todas as menções de personagens numa obra, conforme quantificado no presente artigo. Essa identificação permitiu contabilizar de três a oito vezes mais ocorrências da personagem.

Na noção de correferência também incluímos a deteção e atribuição do sujeito nulo – no que parece sermos pioneiros, mas que, salientamos, é extremamente importante para a nossa tarefa. Tencionamos, em trabalhos futuros, indagar também da importância dos objetos nulos para a caracterização das personagens, assim como avaliar a possibilidade da sua obtenção automática.

Cremos que o corpo anotado aqui apresentado e publicamente disponibilizado é um primeiro passo para o estudo das personagens na literatura em português, e para o desenvolvimento de regras que as identifiquem, assim como para o treino de sistemas automáticos. Não descartamos a hipótese de o continuarmos a incrementar e melhorar no futuro, sendo a versão apresentada aqui apenas a primeira versão.

Referências

Harald Baayen. 2008. *Analyzing Linguistic Data: A practical introduction to Statistics using R*. Cambridge University Press.

David Bamman, Olivia Lewke, e Anya Mansoor. 2020. A Dataset of Literary Coreference. Em *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), Marseille, 11–16 May 2020*, página 44–54.

David Bamman, Brendan O'Connor, e Noah A. Smith. 2013. Learning latent personas of film characters. Em *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, páginas 352–361, Sofia, Bulgaria. Association for Computational Linguistics.

David Bamman, Ted Underwood, e Noah A. Smith. 2014. A bayesian mixed effects model of literary character. Em *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, páginas 370–379.

Eckhard Bick. 2000. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Tese de Doutoramento, Aarhus University, Aarhus, Denmark.

Eckhard Bick. 2010. A dependency based approach to anaphora resolution. Em *Extended Activities Proceedings, 9th International Conference on Computational Processing of the Portuguese Language, Apr. 27-30. Porto Alegre, Brazil*.

Eckhard Bick. 2023a. Attribution of Quoted Speech in Portuguese Text. Em *Proceedings of the NoDaLiDa 2023 Workshop on Constraint Grammar - Methods, Tools and Applications (Tórshavn, Faroe Islands)*, páginas 1–9.

Eckhard Bick. 2023b. Extraction of Literary Character Information in Portuguese. *Linguamática*, 15(1):31–40.

Daniela Barreiro Claro, Joaquim Santos, Marlo Souza, Renata Vieira, e Vládia Pinheiro. 2024. Extração de informação. Em H. M. Caseli e M. G. V. Nunes, editores, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, 3 edição, book chapter 22. BPLN.

Regina Dalcastagnè. 2021. Ausências e estereótipos no romance brasileiro das últimas décadas: alterações e continuidades. *Letras de Hoje, Porto Alegre*, 56(1):109–143.

Regina Dalcastagnè. 2025. A personagem do romance brasileiro contemporâneo: 1990-2004. *Estudos de Literatura Brasileira Contemporânea*, 26:13–71.

Evandro Fonseca, Aline Aver Vanin, e Renata Vieira. 2023. Resolução de correferência. Em *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. 1.a edição.

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

Cláudia Freitas, Elivis de Souza, e Luísa Rocha. 2019. Quantificando e qualificando o sujeito oculto em português. Em *STIL 2029, XII Symposium in Information and Human Language Technology and Collocates Events*, páginas 289–293.

Cláudia Freitas e Elvis de Souza. 2021. Sujeito oculto às claras: uma abordagem descritivo-computacional. *Revista de Estudos linguísticos, Belo Horizonte*, 29(2):1033–1058.

Suemi Higuchi, Diana Santos, Cláudia Freitas, e Alexandre Rademaker. 2019. "Distant reading Brazilian history. Em *Proceedings of 4th Conference of The Association Digital Humanities in the Nordic Countries (Copenhagen, March 6-8 2019)*, páginas 190–200.

Mary Kato, Ana Maria Martins, e Jairo Nunes. 2023. *The Syntax of Portuguese*. Cambridge University Press.

Markus Krug, Lukas Weimer, Isabella Reger, Luisa Macharowsky, Stephan Feldhaus, Frank Puppe, e Fotis Jannidis. 2018. Description of a Corpus of Character References in German Novels - DROC [Deutsches ROman Corpus].

Luisa Mara Silva Lima. 2026. A construção da personagem literária maranhense sob uma perspectiva computacional: uma análise baseada em correferências e sujeitos nulos. Tese de Mestrado, Universidade Estadual do Maranhão.

F. Martins e Cláudia Freitas. 2019. Sujeitos ocultos em verbetes biográficos: Contornando dificuldades da extração automática de informações. Em *XI Congresso Internacional da ABRALIN*.

Franco Moretti. 2013. *Distant Reading*. Verso.

Cristina Mota e Diana Santos, editores. 2008. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca.

Diana Santos. 2018. Galeria de personagens.

Diana Santos e Cláudia Freitas. 2019. Estudando personagens na literatura lusófona. Em *STIL - Symposium in Information and Human Language Technology*.

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

Diana Santos, Cristina Mota, Emanoel Pires, Marcia Langfeldt, Rebeca Schumacher Fuão, e Roberto Willrich. 2023. DIP - Desafio de Identificação de Personagens: objectivo, organização, recursos e resultados. *Linguamática*, 15(1):3–30.

Jefferson Fontinelle da Silva. 2011. Resolução de correferência em múltiplos documentos utilizando aprendizado não supervisionado. Tese de Mestrado, USP.