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

Diana Santos, Luisa Lima, Emanoel Pires

d.s.m.santos@ilos.uio.no, luisamaralima01@gmail.com, emanoeluema@gmail.com

PROPOR2026, Salvador, 14 de abril de 2026

Estudos literários computacionais

Estudar a literatura em português usando métodos e ferramentas de processamento de linguagem natural (PLN) e de linguística com corpos (LC)

- Os dados são literários
- As perguntas de pesquisa são literárias
- Os métodos são de PLN ou de LC
- O ambiente computacional é "gémeo" do para estudar linguística ou língua: a Literateca é uma parte da Gramateca (Santos, 2014)

Dados: Literateca, corpo do AC/DC

Vercial Textos canónicos da literatura portuguesa (prosa, poesia e teatro). Corpo compilado pelo projeto Vercial (Braga, Portugal), que também atualizou a grafia

OBras Textos de literatura brasileira (prosa). Corpo compilado pela Linguateca como um subprojeto do AC/DC, com a colaboração de várias pessoas e instituições (Anya Campos, PUC-Rio e UEMA)

NOBRE Textos não canónicos de literatura portuguesa (prosa). Corpo compilado pela Linguateca, com a colaboração da UEMA (de certa forma, complemento ao Vercial)

Colonia Textos em português (Brasil e Portugal) do século XVI ao XIX. Corpo compilado por Marcos Zampieri (Univ. Köln)

Tycho Corpus Histórico do Português Tycho Brahe (autores nascidos entre 1380 e 1845), criado pela Univ de Campinas

PANTERA Corpo paralelo português-norueguês, com excertos de textos traduzidos entre as duas línguas, criado pela Linguateca. Parte portuguesa.

Quantidade de texto

A 23 de março de 2026 (os corpos OBras, NOBRE e PANTERA continuam a crescer, e a última versão alojada do Tycho Brahe é de 2014)

Corpo	Obras	Autores	tamanho em unidades
Vercial	339	56	15.777.136
OBras	336	67	11.076.840
NOBRE	172	112	9.784.245
Tycho Brahe	39	36	1.736.251
Colonia	35	22	1.601.036
PANTERA	37	23	160.995
Total	958	281	40.142.464

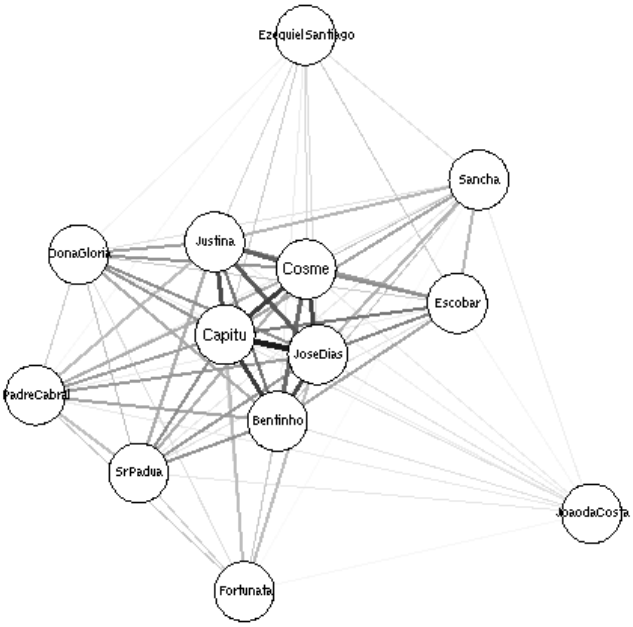
Convém lembrar que há autores cujas obras se encontram em vários corpos.

As personagens numa obra literária

São evidentemente importantes

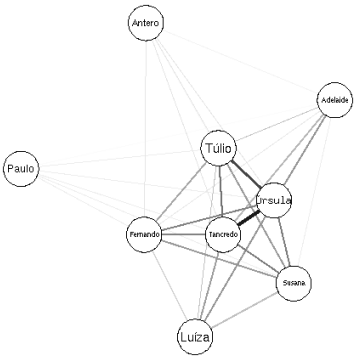
- Há três tipos: principais, secundárias, e figurantes/tipos
- O seu número e as suas ligações
- As suas ocupações
- As relações de parentesco
- A forma como são caracterizadas

Veja-se o Desafio de Identificação de Personagens (DIP, Santos et al. 2023), e o artigo Freitas & Santos (2023)

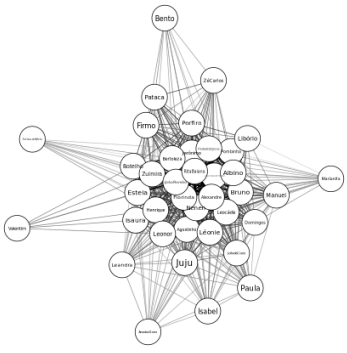


A rede de personagens correspondente a *Dom Casmurro*

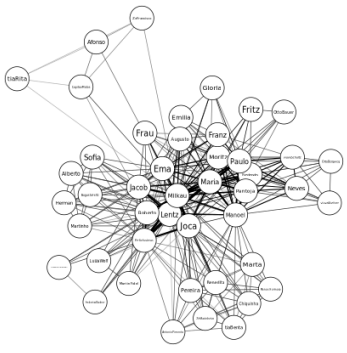
Três obras maranhenses



Úrsula, Maria Firmina dos Reis, 1859, romantismo, 53.568 unidades, 9 personagens

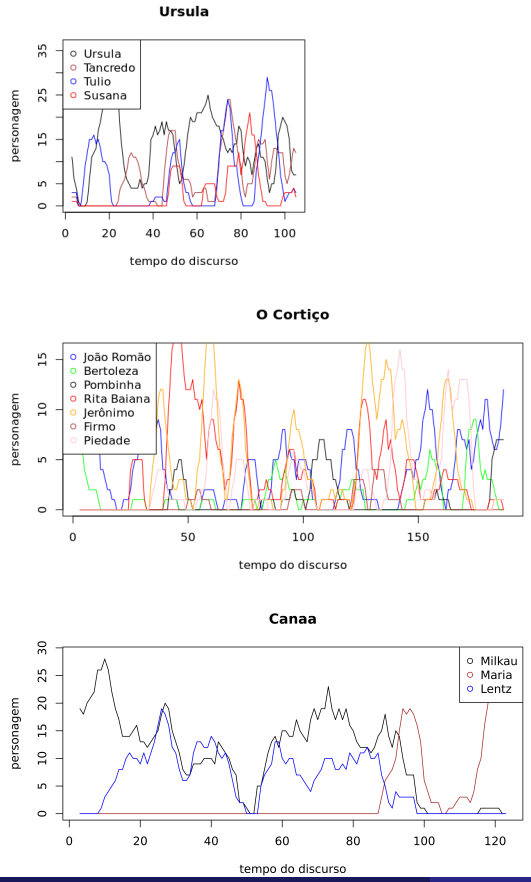


O Cortiço, Aluísio Azevedo, 1890, naturalismo, 95.472 unidades, 38 personagens



Canaã, Graça Aranha, 1902, pré-modernismo, 90.167 unidades, 45 personagens

Três obras maranhenses (2)



Limitações do apresentado até agora e como resolvê-las

- Baseado exclusivamente em casos de nomes próprios
- Sabemos que existem muitos outros casos em que as personagens não são descritas pelos seus nomes próprios
- Quisemos obter um sistema em que TODAS as ocorrências das personagens eram identificadas, e não apenas as mencionadas com um nome próprio
- Também queremos ver se existem diferenças na forma como as personagens são mencionadas

Marcação da correferência

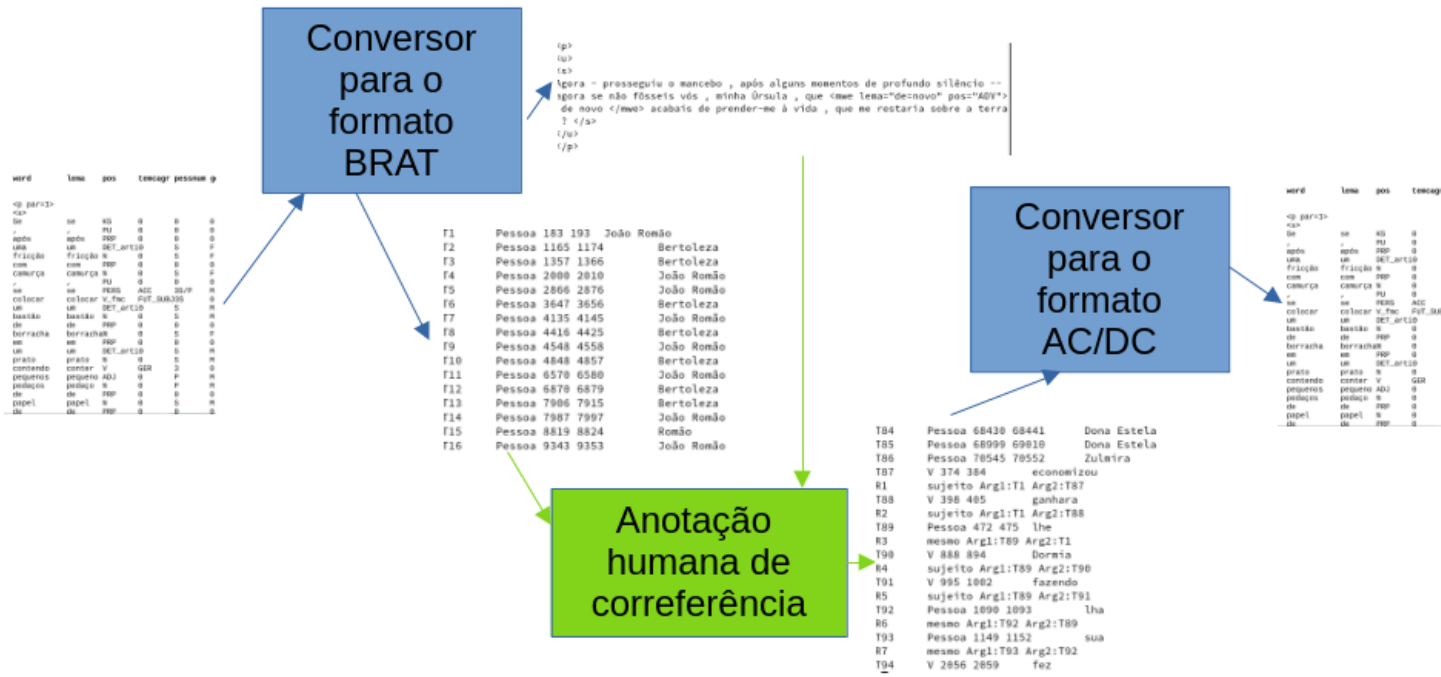
Perguntas de pesquisa

- Qual o aumento de referências que se obtém se marcarmos a correferência?
- Qual a contribuição dos sujeitos nulos, dos pronomes, e de outras formas de correferência?
- O que se aprende sobre as personagens?

Desafios computacionais

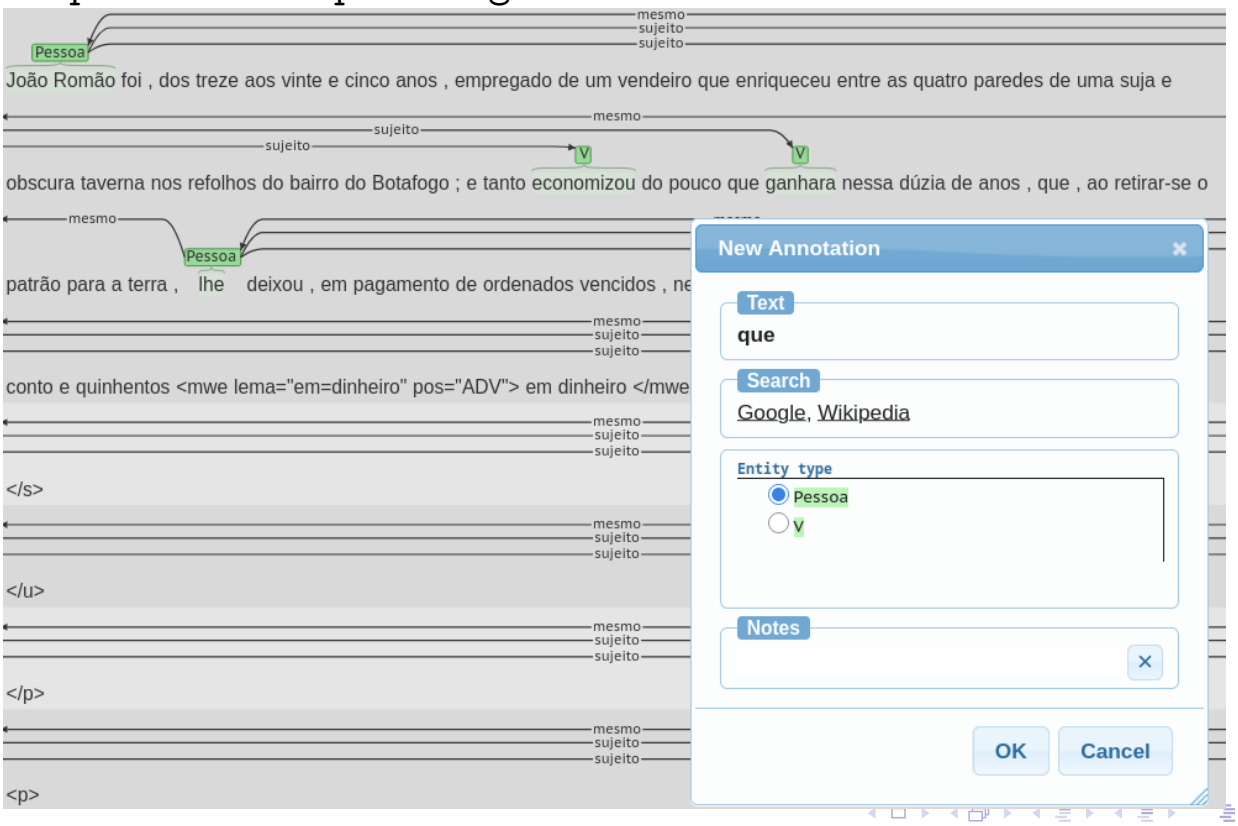
- Como integrar a marcação de correferência no AC/DC?
- Como definir um processo de marcação amigável para anotadores provenientes dos estudos literários?
- Como usar um corpo anotado por seres humanos para desenvolver um sistema automático de anotação?

Sistema completo para a integração pessoa-máquina



BRAT annotation tool

https://brat.nlplab.org/



Novo corpo especializado no AC/DC

https://www.linguateca.pt/acesso/corpus.php?corpus=CORPIREF

Procura: **[coref="V-SUBJ-João.*"]**.

---

id="O Cortiço Prosa:romance AA 1890 naturalismo masc ": 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 .

---

id="O Cortiço Prosa:romance AA 1890 naturalismo masc ": 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 .

---

id="O Cortiço Prosa:romance AA 1890 naturalismo masc ": **Dormia** sobre o balcão da própria venda, em cima de uma esteira, fazendo travesseiro de um saco de estopa cheio de palha .

---

id="O Cortiço Prosa:romance AA 1890 naturalismo masc ": Dormia sobre o balcão da própria venda, em cima de uma esteira, **fazendo** travesseiro de um saco de estopa cheio de palha .

As instruções de anotação detalhadas encontram-se em Lima (2026).

Dados que se podem obter

Verbos de que João Romão é sujeito (nulo) (os mais frequentes):

Houve **331** valores diferentes de **lema**.

ter	19
fazer	16
ser	12
ir	11
pensar	11
dar	10
dizer	10
estar	9
tomar	9
sentir	8
deixar	7
pôr	6

Comparação das personagens

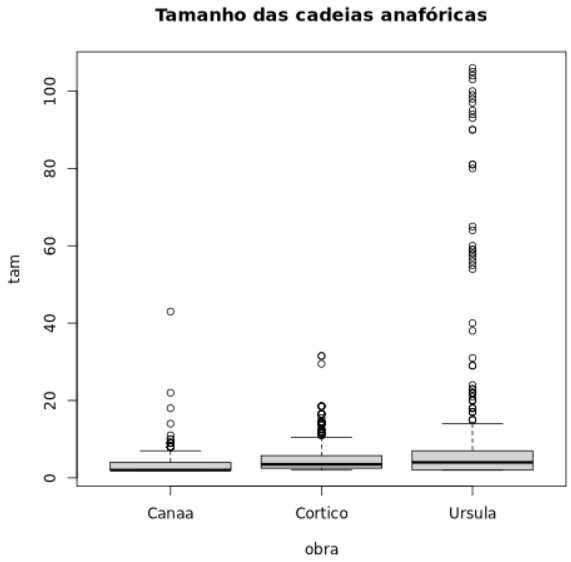
Verbos mais frequentes de que são sujeitos.

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

Cadeias de correferência

Também calculamos as cadeias de correferência. Alguns exemplos:

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



Mais alguns resultados

Instâncias das seis personagens nas três obras:

nomes próprios	1.082	21,6%
sujeitos nulos	1.717	34,3%
pronomes	2.200	44,1%

- Marcar a correferência das personagens principais corresponde a um acréscimo de 362%!
- Diferenças interessantes entre as obras e as personagens

Todos os dados e programas estão acessíveis de
`https://www.linguateca.pt/Gramateca/Literateca/Corref.html`

Trabalho futuro

- Avaliar a importância da correferência nominal (não contemplada no presente trabalho)
- Criar classes de verbos de forma a poder comparar mais facilmente personagens (verbos de dizer, verbos de emoção, verbos de construção, etc.)
- Criar, em conjunto com Eckhard Bick, um sistema público de correferência baseado em regras
- Criar outro sistema público de correferência baseado em aprendizagem automática
- Analisar mais personagens das mesmas e de outras obras, para o objetivo último de ter um sistema de leitura distante que identifica e categoriza personagens em obras literárias em língua portuguesa

Referências

- Cláudia Freitas Diana Santos. "Gender Depiction in Portuguese: Distant reading Brazilian and Portuguese literature". *Journal of Computational Literary Studies* 2, 1, 2023.
- Diana Santos. "Gramateca: corpus-based grammar of Portuguese". In Jorge Baptista et al., *Proceedings of PROPOR 2014*, pp. 214-219.
- Diana Santos, Cristina Mota, Emanoel Pires, Marcia Langfeldt, Rebeca Schumacher Fuão & Roberto Willrich. "DIP - Desafio de Identificação de Personagens: objectivo, organização, recursos e resultados". *Linguamática* 15, 1, 2023, pp. 3-30.
- Luisa Mara Silva Lima. "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, 2026.