

Periodização automática: Estudos linguístico-estatísticos de literatura lusófona

Automatic literary school assignment: Linguistic-statistical studies of lusophone literature

Diana Santos
Linguatca & Universidade de Oslo
d.s.m.santos@ilos.uio.no

Emanoel Pires
Universidade Estadual do Maranhão
emanoel.uema@gmail.com

Cláudia Freitas
Linguatca & PUC-Rio
maclaudia.freitas@gmail.com

Rebeca Schumacher Fuão
Linguatca
rebischu@gmail.com

João Marques Lopes
Linguatca
marqueslopes1928@hotmail.com

Resumo

Neste artigo usamos um conjunto de características sintático-semânticas da língua portuguesa para classificar em períodos literários dois conjuntos de obras. Em que medida tais características são capazes de refletir distinções relevantes no âmbito dos estudos literários é uma das questões que pretendemos investigar.

O primeiro grupo de obras corresponde à replicação do trabalho relatado em 2009 por Barufaldi et al., que usaram métodos de compressão de dados sobre uma série de obras brasileiras classificadas em quatro períodos literários: barroco, arcadismo, romantismo e realismo, desde o Padre António Vieira até Raul Pompéia, contabilizando 15 autores diferentes e totalizando 37 obras.

O segundo grupo inclui muito mais obras (192), tanto portuguesas como brasileiras, mas apenas integra romances ou novelas publicadas no período de 1840 a 1919. As escolas literárias escolhidas foram o realismo, o romantismo, o simbolismo, o naturalismo, o decadentismo e o modernismo, mas, ao contrário da classificação anterior, permitimos que uma mesma obra pertença a várias escolas.

Usamos técnicas de classificação em R para a primeira tarefa, e análise de correspondências para a segunda. Também aplicamos técnicas de modelos de tópicos à segunda coleção para ver se é possível obter tópicos representativos de escolas literárias diferentes.

Palavras chave

leitura distante, linguística com corpos, literatura lusófona, escola literária, português, literatura brasileira, literatura portuguesa

Abstract

In this paper we use a set of syntactic and semantic features of Portuguese to automatically classify li-

terary works in literary periods and/or schools, and address the issue of their appropriateness, for two different literary collections.

The first task attempts to replicate the work by Barufaldi and colleagues, who applied compression methods on 37 Brazilian works by 15 different authors and classified the works in 4 different literary schools.

The second collection, of 192 novels published in Portugal and Brazil in the period 1840 to 1919, features many works who cannot be singly accommodated in one literary school only, and which have been (not mutually exclusively) classified as romantic, realist, naturalist, symbolist, decadent and modernist.

We use classification techniques in R, such as discriminant analysis and support vector models for the first task, and correspondence analysis for the second collection. We also apply topic modeling to (distinct subsets of) the second collection in order to investigate whether this technique can provide us with recurrent topics for different literary schools.

Keywords

distant reading, corpus linguistics, literary school, Portuguese, Brazilian literature, Portuguese literature, lusophone literature

1 Introdução

O objetivo do presente artigo é avaliar se a informação linguística que temos vindo a associar, em estudos linguísticos da língua portuguesa, a várias obras literárias pode também ser usada para responder a questões do foro dos estudos literários.

Para tal, compilámos uma lista de características sintáticas e semânticas a que temos acesso na Literatca (Santos, 2019b; Santos & Simões, 2019) e usámo-las em dois problemas, que passamos a descrever sucintamente:

- atribuir 37 obras de 15 autores brasileiros diferentes a quatro períodos literários, replicando um trabalho anterior feito com métodos de compressão
- organizar 192 romances ou novelas de autores portugueses e brasileiros publicadas entre 1840 e 1919, tentando apreciar semelhanças entre autores e escolas literárias, seguindo a proposta de Santos et al. (2018b) inspirada por Moretti (2000)

A nossa posição é a de explorar a informação que temos e não a de demonstrar que estes métodos resolvem os problemas literários. Na medida em que for possível encontrar formas de identificar semelhanças e grupos que concordem com a autoridade literária, ou que levem a perguntas pertinentes, estes métodos de leitura distante poderão contribuir para os estudos literários. Se, pelo contrário, indicarem outros agrupamentos, tal não deve ser considerado como uma teoria alternativa, mas apenas como demonstrando que as características escolhidas não eram relevantes para o problema em questão.

Em relação às obras usadas como material para a nossa pesquisa, a lista exata encontra-se em apêndice. Além de material compilado pela própria Linguatca, usamos textos gentilmente cedidos pelos seguintes projetos irmãos *Corpus Histórico do Português Tycho Brahe* (Galves & Faria, 2010) e *Colônia - Corpus of Historical Portuguese* (Zampieri & Becker, 2013). Exceto no caso dos textos provenientes do corpo PANTERA (Santos, 2019c), trabalhamos com textos completos.

2 Características usadas

Na Literateca, além do acesso ao texto em formato eletrônico, temos a vantagem de ter (e disponibilizar para consulta) todo o material anotado gramatical e semanticamente. A anotação morfossintática é feita pelo analisador PALAVRAS (Bick, 2000).

Para ambas as tarefas calculámos um conjunto extenso de características de cada texto (128) que nos pareceram de interesse para uma possível descrição do estilo, que passamos a elencar:

A partir da anotação morfossintática do PALAVRAS, levamos especificamente em conta em nossa análise a presença de adjetivos (índices de qualificação) e nomes próprios (instâncias de classes genéricas como pessoas/personagens e locais, entre outros); a presença e a distribuição de construções como voz passiva ou a forma progressiva; orações relativas e completivas (índices

de complexidade estrutural); a presença de coordenações e conjunções coordenativas, bem como de vírgulas e outros sinais de pontuação (índices de ritmo). As indicações de tempo, modo e aspeto verbal também foram consideradas potenciais elementos caracterizadores (especificamente o modo conjuntivo, o pretérito perfeito composto, o pretérito imperfeito, o perfeito, o mais que perfeito e os aspetualizadores), bem como a presença de verbos na primeira pessoa, e de palavras no género morfológico feminino. Pontos de exclamação, de interrogação e travessões, e elementos de negação também foram utilizados como índices potencialmente caracterizadores de autores, obras e/ou estilos.

De um ponto de vista estilístico, o número de palavras por frase é um elemento que costuma ser utilizado na diferenciação de autores e obras —veja-se, por exemplo, a matéria de Almeida & Mariani (2019), que utiliza este traço para produzir gráficos relativos a obras da literatura brasileira —, e por isso usamos o número de frases por obra.

Além da anotação morfossintática, o material da Literateca conta também com a anotação de diversos campos semânticos¹. Neste trabalho, levamos em conta os campos dos verbos de fala, da saúde/doença, cores, corpo humano, família, roupas e emoções.

Com relação ao campo do dizer Freitas et al. (2016), partimos de um léxico de verbos específicos e de regras que indicam se os verbos estão sendo utilizados para introduzir a fala de alguém (relato direto ou indireto) ou se apenas se trata da menção a algum evento comunicativo (“...e não falou mais no assunto”). Como características, usamos três: verbos de relato direto, verbos de relato indireto, e verbos de fala somente.

O campo semântico das emoções conta com variadas palavras, de diferentes classes gramaticais, distribuídas em 24 grandes grupos, como amor, coragem, desejo, desespero, felicidade, fúria, admiração, inveja e medo, entre outros (veja-se o Emocionário² para sua documentação cabal). O número de palavras em cada um destes grupos é uma característica, assim como o total de palavras de emoção.

No campo da saúde (Santos, 2019a), usamos os seguintes indicadores: o número de palavras

¹Por campo semântico denotamos uma área de conhecimento refletida na língua, como a cor, ou a família. Infelizmente este é um uso completamente distinto daquele que é definido pelo Dicionário Terminológico, conforme nos chamou a atenção Álvaro Iriarte Sanroman.

²<https://www.linguatca.pt/Gramateca/Emocionario.html>

desse campo, a presença do lema dor, e a presença de palavras dos subcampos progressão (da doença), causa (da doença), palavras genéricas sobre saúde ou doença, remédios, acessórios (relacionados com saúde/doença), medicina e saúde psicológica.

No campo da cor (Silva & Santos, 2012), usamos o total de palavras de cor, o total de palavras de cor com sentido de cor, assim como o número de palavras pertencente a cada grupo de cor (Laranja, Vermelho, Dourado, etc.) e o total de palavras de cor com sentido não cor, ou seja, presentes em expressões fixas como *buraco negro*, *luz verde*, *lista negra*.

Para o corpo humano (Freitas et al., 2015), usamos o total de palavras de corpo, o total de palavras de corpo com sentido literal, e o número de palavras pertencentes a cada parte do corpo (Cabeça, Sexual, Pernas, etc.).

Para a roupa (Santos et al., 2011), usamos o total de palavras de roupa, e o número de palavras pertencentes a cada grupo de roupa (Calças, RoupaDormir, Calçado, etc.).

Finalmente, para o campo da família usamos o número de palavras relacionadas com a família, assim como o campo mais específico de parentesco. Veja-se Higuchi et al. (2019) para uma motivação deste campo.

A lista completa, por ordem alfabética, encontra-se no sítio da Linguatca³. Convém esclarecer que a marcação destes campos semânticos é feita automaticamente através de regras, e tem alguma margem de erro.

Uma primeira discussão destes indicadores no contexto da literatura está presente em Santos (2019b). Mas desde esse trabalho, que data de 2017 embora apenas tenha sido publicado em 2019, adicionámos várias características e várias obras.

3 Primeira tarefa: repetir o trabalho de Barufaldi et al.

Barufaldi et al. (2009, 2010) usaram métodos de compressão de dados sobre uma série de obras brasileiras classificadas em quatro períodos literários: barroco, arcadismo, romantismo e realismo, desde o Padre António Vieira até Raul Pompéia, contabilizando 15 autores diferentes e totalizando 37 obras.

Tentando obter exatamente o mesmo material utilizado por Barufaldi et al. (2009, 2010), optamos, sempre que possível, por obras que estão

disponibilizadas em sítios como o da Biblioteca Digital de Literaturas de Língua Portuguesa⁴ e do Domínio Público⁵. Ainda assim, como é restrita a informação mencionada pelos autores no que diz respeito às edições utilizadas, é possível que haja mudanças, ainda que mínimas em alguns casos, nos textos das obras escolhidas. Ademais, a publicação inicial em folhetim e posterior edição em formato livro também pode ser outro fator que acarrete variações nas edições escolhidas. Sobre alterações nas edições das obras de Machado de Assis, por exemplo, conferir Campos (2018).

De todo modo, parece-nos que a possível distinção mais radical entre os arquivos das obras utilizadas possa estar em *14 de Julho na Roça*, de Raul Pompéia. Como se trata de uma coletânea de contos nomeada de *Contos* na Biblioteca de Literaturas de Língua Portuguesa e de *14 de Julho na Roça* no sítio do Domínio Público, restou a dúvida se os autores utilizaram apenas o conto inicial, intitulado de *14 de Julho na Roça*, ou se a obra completa. Como estamos tratando de textos escritos pelo mesmo autor e em um espaço de tempo muito próximo, optamos por utilizar a obra completa, admitindo que todos os contos têm o mesmo estilo de época.

Quanto ao processo computacional, aplicámos duas técnicas (Baayen, 2008) usando o ambiente R (R Development Core Team, 2008), empregando as características descritas acima para o mesmo fim:

- análise de discriminantes com base em componentes principais (ver figuras 1 2 3)
- máquinas de vetores de apoio (support vector machines) (ver tabela 1)

	arcad.	barr.	real.	romant.
arcadismo	3	2	0	0
barroco	0	7	0	0
realismo	0	0	10	3
romantismo	0	1	0	11

Tabela 1: Resultado da classificação com máquinas de vetores de apoio

Ainda não nos está suficientemente claro os motivos por detrás da nossa classificação equivocada das obras *O Uruguai*, de Basílio da Gama, e *Coletânea de obras*, de Alvarenga Peixoto. Em Barufaldi et al. (2009, 2010), a *Coletânea*

³https://www.linguatca.pt/Gramateca/Literateca/lista_caracteristicas.txt

⁴<https://www.literaturabrasileira.ufsc.br/>
⁵<http://www.dominiopublico.gov.br/pesquisa/PesquisaObraForm.jsp>

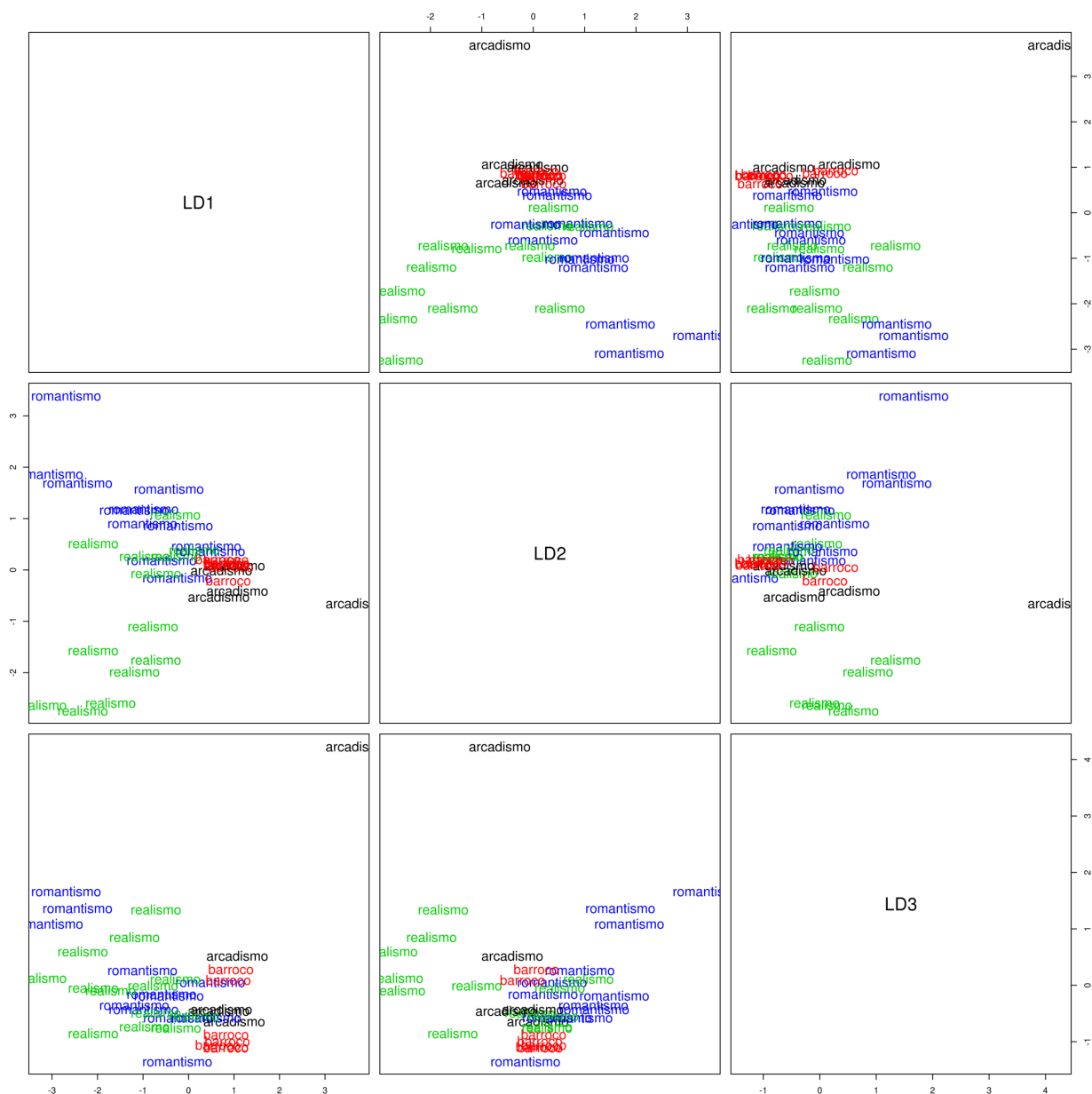


Figura 1: Análise de discriminantes, global

de obras líricas, de Gregório de Matos, também foi classificada erroneamente. As causas podem estar relacionadas ao conjunto de características marcadas nas obras e utilizadas como parâmetro nas análises como, também, à falta de marcações que estejam mais relacionadas com elementos que sinalizem de maneira mais efetiva o estilo na poesia, como os processos de acomodação silábica. Mittmann et al. (2016) utilizam uma ferramenta de escansão automática, o Aoidos (<https://aoidos.ufsc.br/>), que, em estudos futuros, poderá ajudar nos casos de confusão.

Sobre *Ubirajara*, de José de Alencar, ter sido classificado como pertencente ao Barroco, mesmo

sendo em prosa, a hipótese inicial com a qual trabalhamos diz respeito ao tamanho médio das frases do romance de Alencar, que, por serem muito curtas, se diferenciam em muito do estilo empregado nos demais romances do período romântico que foram incluídos na análise.

Seja como for, os resultados indicam que estas características parecem ser apropriadas para distinguir entre os quatro períodos ou escolas literárias selecionados pelos nossos antecessores, embora com uma leve tendência a privilegiar o barroco. Mas pensamos que a tarefa pode ter sido demasiado simples, dado que os diferentes períodos também implicam diferenças tão abissais como poesia vs. prosa e correspondem a

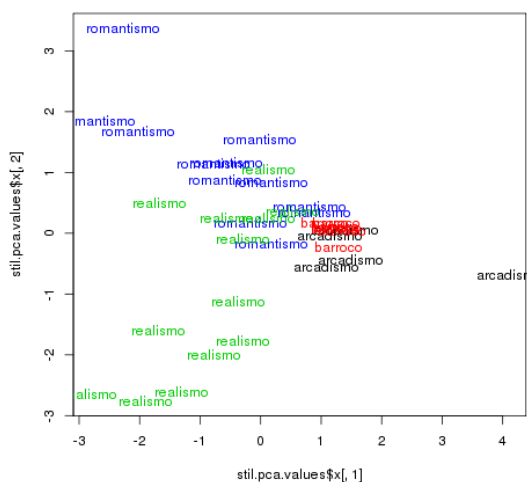


Figura 2: Análise de discriminantes, mostrando o primeiro e o segundo

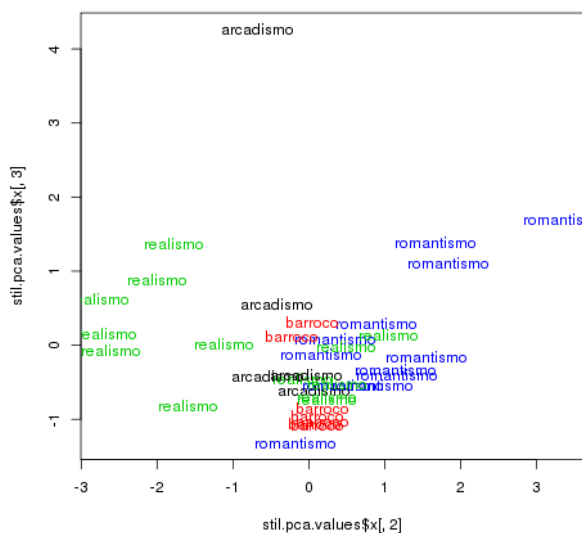


Figura 3: Análise de discriminantes, mostrando o segundo e o terceiro

épocas razoavelmente distintas.

4 Segunda tarefa: romances e novelas portuguesas e brasileiros do período 1840-1919

O segundo conjunto de obras pode ser mais complexo de classificar, visto que se refere a um período de apenas 80 anos, e a duas formas muito semelhantes: o romance e a novela, ambas correspondentes à *novel* inglesa, daí a razão

da amálgama⁶. Contém autores que, devido à sua longevidade e/ou génio, produziram obras que são tradicionalmente consideradas de escolas diferentes, e possui elementos que muitos estudiosos consideram inqualificáveis, por únicos.

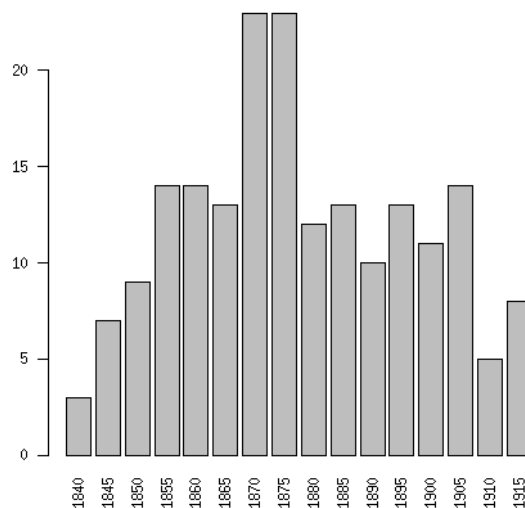


Figura 4: Segundo conjunto de obras por data de publicação

Não cabendo fazer aqui a discussão pormenorizada do assunto, limitamo-nos a sinalizar dois exemplos de complexidade e singularidade, remetendo o leitor para a lista de referências consultadas⁷. O primeiro é *O Ateneu*, do brasileiro Raul Pompeia, que tem uma longa recepção crítica em que uns o consideram naturalista ou realista, outros o tacham de impressionista, outros indicam o predomínio do simbolismo e há ainda quem assinale o expressionismo ou o cruzamento de duas ou mais escolas no seu interior, conforme se pode constatar nos estudos de Araújo (2011) e de Quintale Neto (2007). O segundo é *Os Maias*, do português Eça de Queirós, em cuja classificação Carlos Reis (Reis, s/d), reconhecidamente um dos maiores especialistas neste autor, oscila entre realismo, naturalismo e pós-naturalismo. Na Literateca, e não querendo escolher entre as várias escolas, *O Ateneu* está marcado como **impressionismo-naturalismo-realismo-**

⁶Convirá a este propósito mencionar que o interesse especial por este período vem da ação COST *Distant Reading for European Literary History*, <https://www.distant-reading.net/>, em cujo âmbito estamos a produzir duas coleções de obras em português, uma coleção portuguesa Ore et al. (2020) e uma coleção lusófona, também com obras brasileiras do mesmo período.

⁷Acessível de https://www.literateca.pt/OBRAS/siglas_Literateca.pdf

-simbolismo, e *Os Maias* como realismo-naturalismo-pós-naturalismo⁸.

Portanto, isso fez com que a tarefa de atribuir um rótulo a cada obra não fosse algo linear, indo de casos cuja taxinomia foi efetivamente unívoca (por exemplo, ninguém duvida de que *Eurico, o presbítero*, de Alexandre Herculano, é um marco do romantismo português) a casos mais complicados como os que acabamos de referir.

Devido à facilidade em adicionar a informação de que nos encontrávamos em presença de um romance histórico, essa classificação foi também adicionada aquando da classificação (e cobriu apenas obras classificadas como romantismo).

Vale ainda mencionar que algumas obras não possuem uma classificação dada, e muitas vezes nem são citadas, em nenhuma obra de referência sobre a história da literatura de Portugal ou do Brasil. Isso ocorre porque as mesmas não são obras canônicas. Nesses casos, foi preciso desenvolver um método que nos permitisse classificar essas obras de forma coerente com o conjunto que possuímos.

Primeiramente, realizamos um mapeamento de características que nos permitissem identificar a escola à qual uma obra pertence, tais como: tempo, narrador, espaço, personagens, temas, finais felizes ou infelizes, etc. Após esse mapeamento, partimos para a leitura de trechos ou da obra completa para então discutir e determinar em que escola poderíamos enquadrá-la.

O resultado deste trabalho encontra-se na tabela 2.

Dito isso, a coleção que usámos é a seguinte: todos os romances e novelas em português em formato eletrónico a que tínhamos acesso à data de 25 de outubro de 2019 —estão em curso diversas iniciativas para aumentar este acervo, mas estes são os que pudemos coligir nessa altura e que foram publicados no período já mencionado (1840-1919).

Isso corresponde a 192 obras (listadas no anexo), das quais 123 portuguesas e 69 brasileiras. O autor com mais obras é Camilo com 37, seguido de Machado de Assis com 13, Aluísio de Azevedo com 11, Eça de Queirós com 11, José de Alencar com 9, Júlio Dinis com 8, e Alexandre Herculano com 6. Os restantes autores têm entre uma a quatro obras nesta coleção. (Cinco obras são traduzidas, duas por Machado de Assis,

⁸De notar que a escolha das classes foi feita com base nos especialistas que sobre os autores e obras se pronunciaram, o que resultou em que por exemplo apenas um autor, Eça de Queirós, tem (em algumas obras) a classificação *pós-naturalismo*, e uma autora, Ana de Castro Osório, *pós-romantismo*.

Escola literária	Quantos
decadentismo	1
expressionismo	1
expressionismo-simbolismo	1
ficção	1
histórico	3
historico-romantismo	2
impress.-natural.-realismo-simbol.	1
indianismo-romantismo	2
modernismo	2
naturalismo	12
naturalismo-realismo	2
naturalismo-realismo-romantismo	1
naturalismo-regionalismo	1
picaresco-realismo	1
realismo	20
realismo-naturalismo	8
realismo-pos-naturalismo	1
realismo-posnaturalismo	5
realismo-regionalismo-romantismo	3
realismo-romantismo	1
regionalismo	2
romantismo	78
romantismo-decadentismo	1
romantismo-histórico	16
romantismo-indianismo	1
romantismo-indianismo-histórico	1
romantismo-realismo	15
romantismo-realismo-naturalismo	5
romantismo-regionalismo	3
simbolismo	1

Tabela 2: Escolas literárias atribuídas

uma por Eça de Queirós, outra por Camilo Castelo Branco, e uma adaptada por Pedro Supico de Moraes. Estas obras são úteis para servirem de teste.)

Na figura 4 pode ver-se a distribuição da data de publicação destas obras por períodos de cinco anos.

Uma análise de correspondências mostra-nos como as características que seleccionámos colocam os diferentes autores, e as diferentes escolas, no plano definido por estas.

Na figura 5 vê-se cada obra com uma cor diferente por autor, além de apresentar as características mais discriminadoras neste conjunto de obras a vermelho, nomeadamente o número de completivas, de interrogativas, a menção de humildade e a referência a medicina ou progressão de uma doença.

Vemos que há autores, como Aluísio de Azevedo e Júlio Dinis, que são bem fiéis a si próprios, definindo portanto áreas bem claras no plano, en-

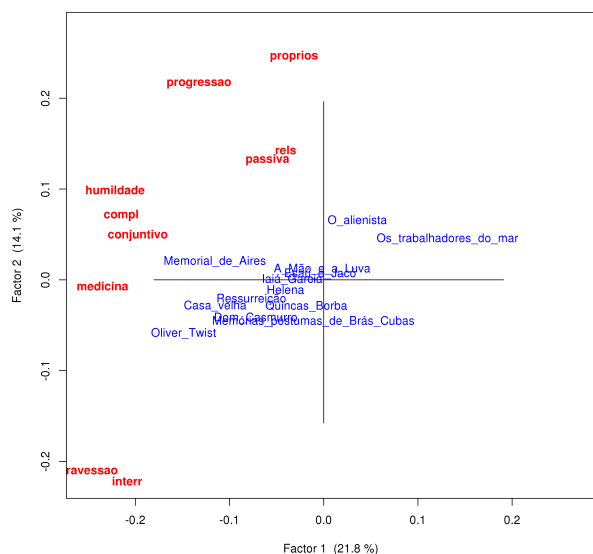


Figura 6: Análise de correspondências mostrando apenas as obras de Machado de Assis

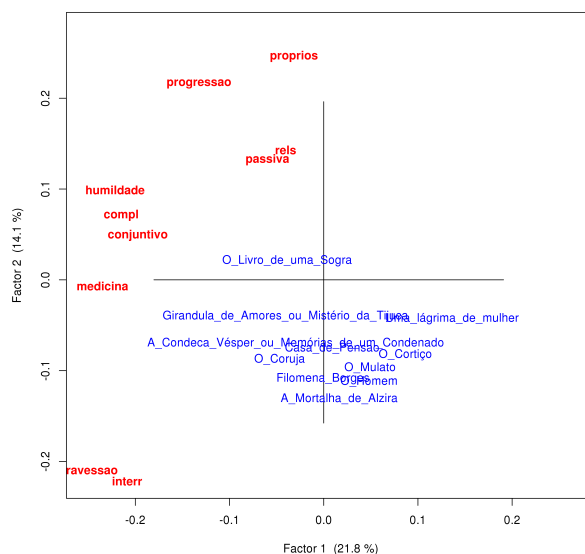


Figura 8: Análise de correspondências mostrando apenas as obras de Aluísio de Azevedo

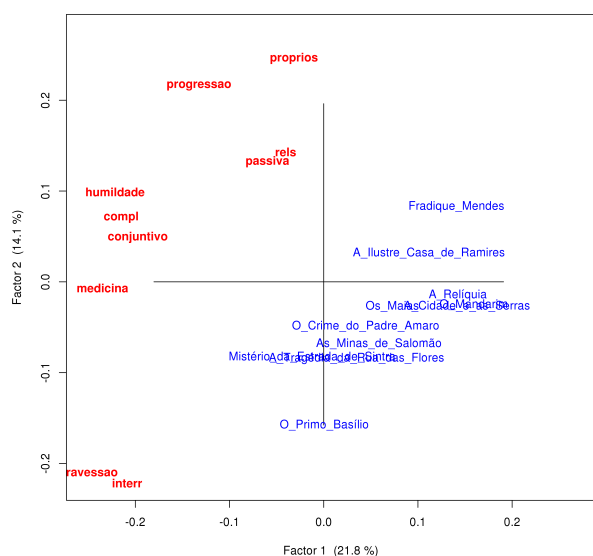


Figura 7: Análise de correspondências mostrando apenas as obras de Eça de Queirós

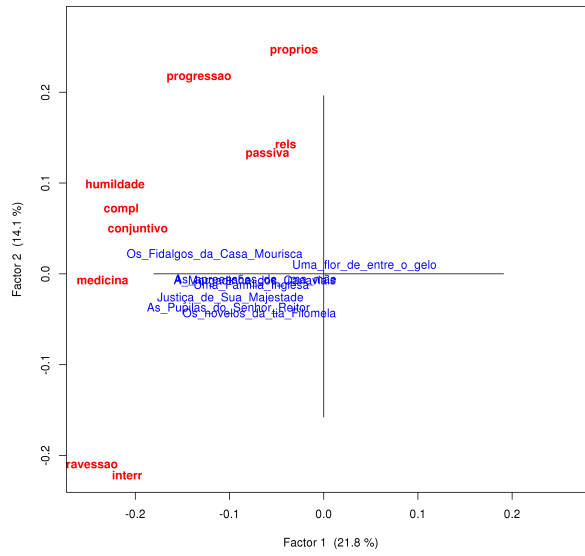


Figura 9: Análise de correspondências mostrando apenas as obras de Júlio Dinis

camente, usámos as seguintes regras para “traduzir” a pertença para apenas uma escola: qualquer menção a romantismo, indianismo ou histórico ganhava a classificação de romantismo. Depois, qualquer menção a realismo ou regionalismo ficavam realismo puro. Em seguida, se naturalismo era mencionado, ficava naturalismo, enquanto simbolismo e decadentismo eram amalgamados em simbolismo. Obviamente outras formas de reclassificar seriam possíveis, por exemplo usando a primeira classificação em vez de ordenar a decisão da forma que fizemos.

Seja como for, temos claramente regiões em que as escolas se localizam, mesmo que não haja regiões sem sobreposição. A figura 11 mostra a situação sem simplificações, ou seja, cada obra aparece com o conjunto de escolas que lhe foram atribuídas (veja-se de novo a tabela 2).

5 Análise de tópicos

Desviando-nos um pouco da análise linguística, resolvemos também usar o método estatístico mais comum dos estudos literários: a análise de

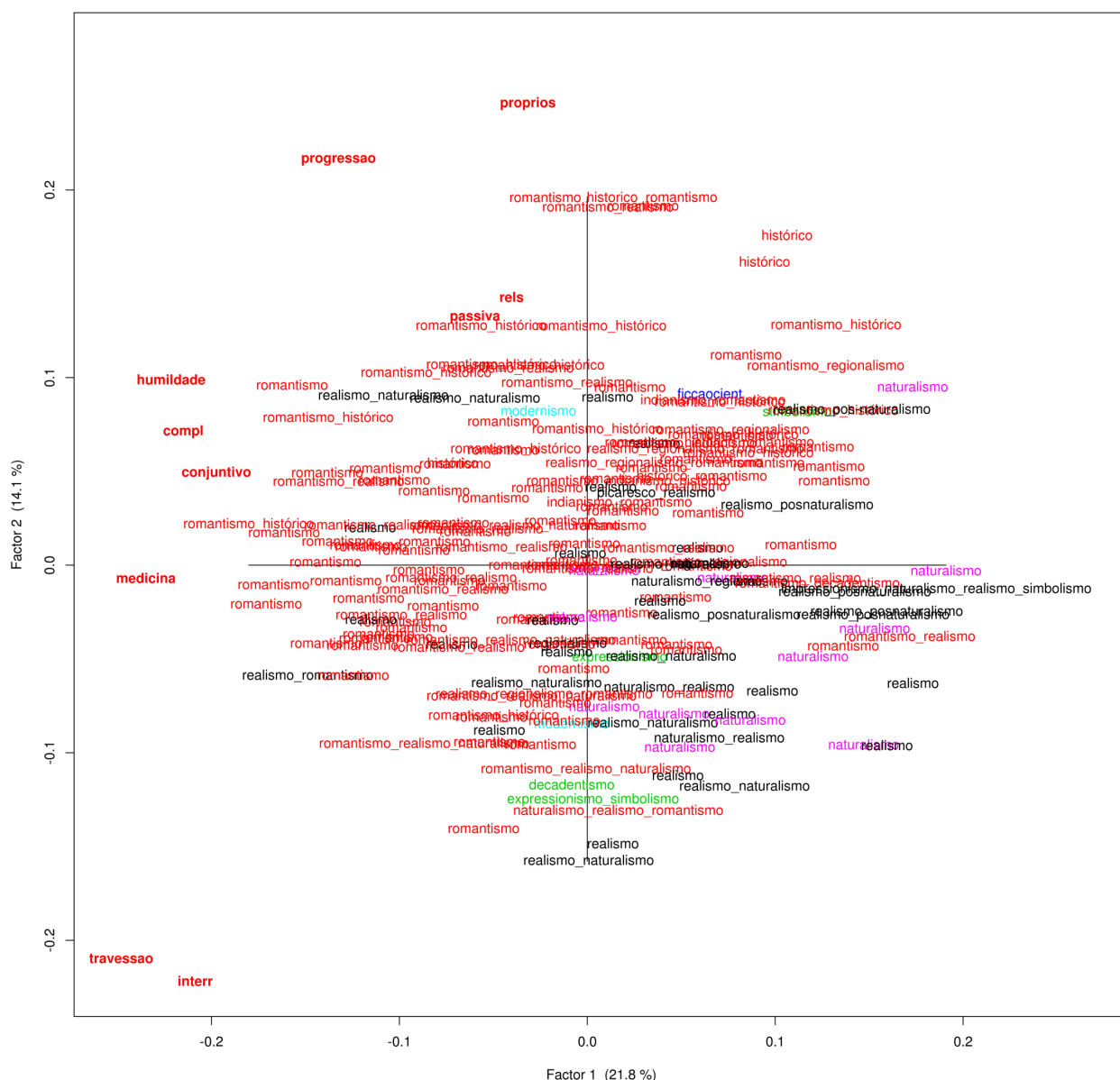


Figura 10: Análise de correspondências mostrando cada escola com uma cor diferente

Nova escola literária	Quantos
expressionismo	1
ficcaocient	1
modernismo	2
naturalismo	12
realismo	42
romantismo	131
simbolismo	3

Tabela 3: Escola literária simplificada

tópicos (*topic modelling*), ver Jockers (2013), que apenas usa as palavras e calcula os tópicos sem acesso a outras classificações (exceto que as pa-

lavras usadas são exclusivamente as das classes gramaticais substantivo, adjetivo e advérbio, obtidas pela análise do PALAVRAS).

Usando blocos de 500 dessas palavras consecutivas, e pedindo 100 tópicos, o sistema *mallet* (McCallum, 2002) produziu a lista em https://www.linguateca.pt/Gramateca/Literateca/artigoAPL/topicos_todas_as_obras/nomestopicosNA_todos_tam500_apl.txt para a coleção completa.

Apresentamos alguns tópicos que nos parecem esclarecedores, pela consistência e facilidade de interpretação, na tabela 4.

Alguns destes apresentamos também em nuvem de palavras, nas figuras 12 e 13.

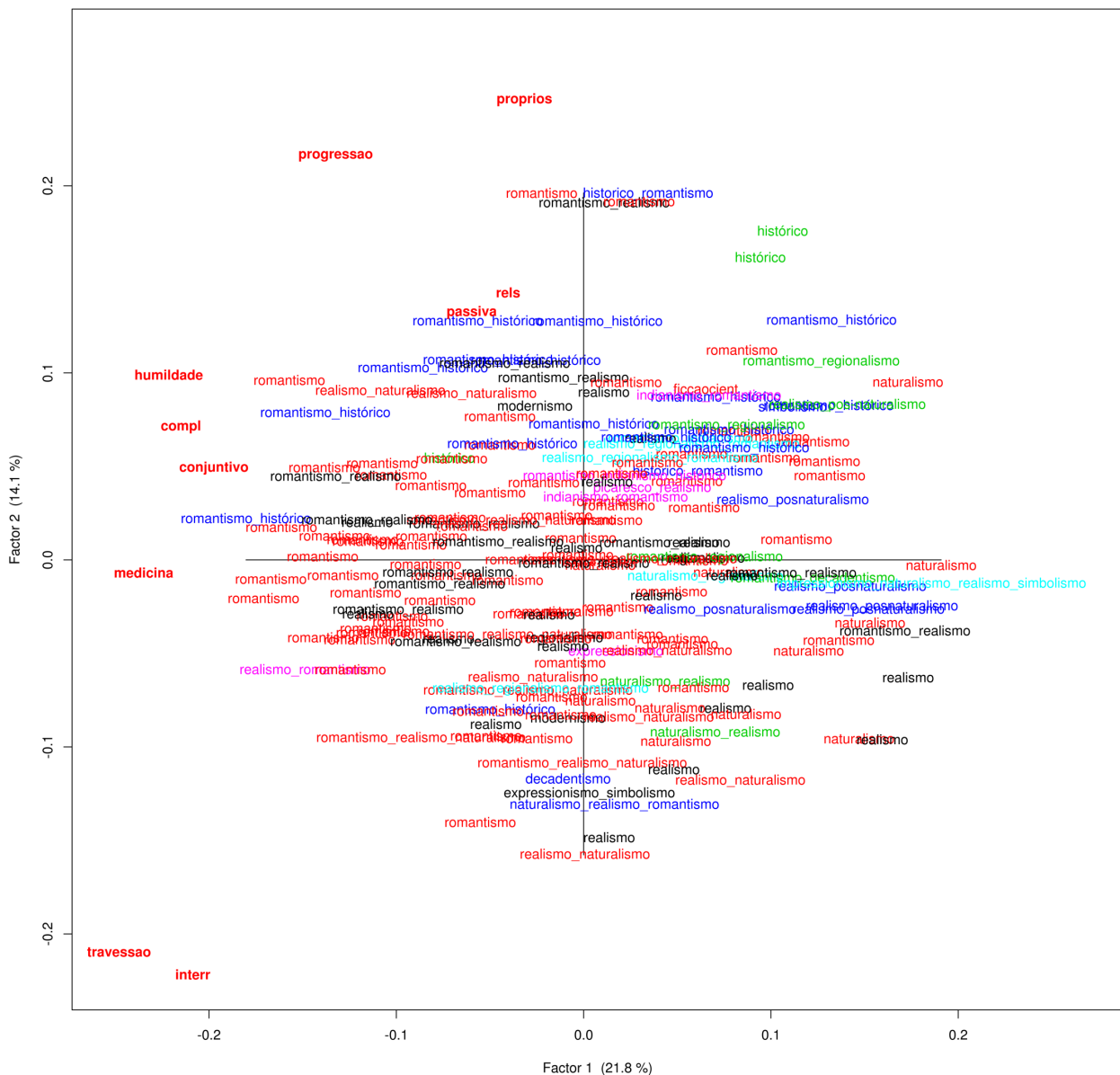


Figura 11: Análise de correspondências mostrando todas as classificações de escola literária

num.	tópico
13	porta rua janela parede luz
16	homem arma inimigo soldado guerra
52	padre santo igreja missa religião
59	cavalo caminho estrada homem cavaleiro
90	sala baile festa salão sociedade
93	livro poeta romance verso obra

Tabela 4: Tópicos obtidos sobre as obras classificadas como românticas

Outros há que não são facilmente interpretáveis, enquanto outros ainda são mais específicos, como 4 (romano lusitano povo exército

cidade) ou 80 (gaúcho sertanejo vez animal fazenda).

Selecionando apenas as obras marcadas (não necessariamente exclusivamente) com a classificação de românticas (124 obras), por um lado, e realistas e/ou naturalistas, por outro (68 obras), obtemos as seguintes novas listas: https://www.linguateca.pt/Gramateca/Literateca/artigoAPL/topicos_romantismo/nomestopicosNA_romantismo_tam500_apl.txt e https://www.linguateca.pt/Gramateca/Literateca/artigoAPL/topicos_realismo/nomestopicosNA_realismo_tam500_apl.txt. Na Tabela 5 apresentamos tópicos realistas/naturalistas, e na Tabela 6 românticos.



Figura 12: Palavras que constituem o tópico 13



Figura 13: Palavras que constituem o tópico 90

6 Comentários e Trabalho futuro

Este trabalho marca o início de um programa de colaboração entre estudos literários e linguística computacional para mutuamente enriquecer ambas as disciplinas. Dessa forma, em vez de muitos resultados, temos muitas interrogações, e vias de desenvolvimento futuro.

Se por um lado pensamos ter mostrado que as características linguísticas (nas secções 3 e 4) e o conteúdo lexical (na secção 5) são úteis para a exploração e estudo da literatura, estamos plenamente conscientes de que muito mais trabalho tem de ser feito em relação à identificação e correta anotação de muitas destas características, e

num.	tópico
25	médico dia doente febre saúde
44	dinheiro conto negócio real carta
55	mulher amor paixão vida beijo
60	estudante colégio professor diretor livro
73	casa porta noite sala quarto

Tabela 5: Tópicos obtidos sobre as obras classificadas como realistas ou naturalistas

num.	tópico
1	mar vento praia onda tempestade
14	guerreiro chefe virgem cabana taba
19	cavaleiro homem rei batalha namorado
68	flor sombra sol doce jardim
89	navio homem marinheiro bordo capitão
90	leito doente quarto corpo morte

Tabela 6: Tópicos obtidos sobre as obras classificadas como românticas

pretendemos efetuar muito brevemente estudos de algumas em particular, como as emoções e o corpo.

Por outro lado, foi evidente que a noção de escola literária não era uma questão simples, e que muitas outras características e interrogações seriam possíveis, desde o género do autor, data de escrita, local de escrita (por exemplo Brasil ou Portugal) ao tipo de obra (romance histórico, romance de costumes, etc.).

Além disso, o facto de termos um número considerável de obras que caíram no esquecimento, e que provavelmente nunca foram colocadas numa escola literária pelos teóricos da literatura, pode também levantar a questão de que as escolas do cânone literário poderão não ser as únicas, e que outras fações ou movimentos podem ser sugeridos através do estudo de mais obras — uma das vantagens da leitura distante.

O peso de determinados autores assim como a possibilidade de discordância em relação à escola ou escolas atribuídas mostra como muito do que por exemplo encontrámos no romantismo poderia mudar se seguissemos a opinião de João Gaspar Simões (Simões, 1967) e considerássemos Camilo um autor por si só (e não romântico!)¹⁰

Será pois relevante, de um ponto de vista literário, experimentar fazer outras escolhas (mesmo entre as obras que temos) para criar novas (sub)coleções, para não deixar que um autor, uma época ou um tipo de escrita pese demais no

¹⁰Convém referir que muitos autores, incluindo, aliás, Camilo, tiveram obras classificadas em escolas diferentes... Cada obra recebeu uma classificação distinta.

grupo.

Muitas outras técnicas estatísticas, assim como tarefas, podiam ter sido realizadas com este material, que se encontra público¹¹ para que outros o possam experimentar e avançar no estudo da literatura lusófona¹². Mencionamos como uma das mais evidentes a tentativa de classificação da segunda coleção através do método das árvores de decisão.

Concluindo, este trabalho é apenas um primeiro passo no uso de métodos estatísticos e linguísticos para reconsiderar a literatura lusófona. Ao tornarmos públicos os documentos e as análises, assim como os primeiros resultados, esperamos que alguns nos sigam no destringir de características, influências e semelhanças entre muitos autores que escreveram em português, assim como desejamos que este tipo de explorações nos dê mais conhecimento sobre o estilo e a “alma” linguística da língua portuguesa.

Agradecimentos

Estamos muito gratos a Alckmar Luiz dos Santos pelas sugestões e críticas feitas em Oslo, à audiência da APL em Braga pelas perguntas pertinentes, e aos revisores Miguel Anxo Portela e Álvaro Iriarte Sanroman pela revisão atuante de uma primeira versão deste trabalho.

Agradecemos à FCCN pelo alojamento da Linguatca nos seus servidores, ao grupo de Research Computing da Universidade de Oslo pelo apoio informático, e à UNINETT Sigma2 - the National Infrastructure for High Performance Computing and Data Storage in Norway pelos recursos computacionais cedidos para o processamento dos corpos e a obtenção de resultados.

Este artigo não existiria se não tivesse sido desencadeado pela ação COST “Distant reading for European literary history”, financiada pelo EU Framework Programme da União Europeia, Horizon 2020.

Finalmente, Emanuel Pires agradece à FAPEMA pelo apoio ao projeto “Estudos estatístico-literários em literatura lusófona: junção de esforços entre a Linguatca e o Portal Maranhão”.

¹¹Em <https://www.linguatca.pt/Gramatca/Literateca/artigoAPL/> podem encontrar-se quer os dados como os comandos utilizados em R.

¹²Vários dos corpos de onde estas obras foram retiradas estão acessíveis da Linguatca, veja-se por exemplo o OBras (Santos et al., 2018a), ou de outros projetos como o Tycho Brahe (Galves & Faria, 2010).

Referências

- Almeida, Rodolfo & Daniel Mariani. 2019. O ritmo e o estilo de diferentes obras literárias brasileiras. <https://www.nexojournal.com.br/grafico/2017/01/30/0-ritmo-e-o-estilo-de-diferentes-obras-liter%C3%A1rias-brasileiras>.
- Araújo, Francisco Magno da Silva de. 2011. *O Ateneu e a nostalgia da forma*: Centro de Ciências Humanas, Letras e Artes da Universidade Federal do Rio Grande do Norte, Natal. Tese de Mestrado.
- Baayen, Harald. 2008. *Analyzing Linguistic Data: A practical introduction to Statistics using R*. Cambridge University Press.
- Barufaldi, Bruno, Eduardo F. Santana, José Rogério B. B. Filho, Jan Kees van der Poel, Milton Marques Júnior & Leonardo Vidal Batista. 2009. Classificação Automática de Textos por Período Literário Utilizando Compressão de Dados Através do PPM-C. Em *STIL 2009, The 7th Brazilian Symposium in Information and Human Language Technology, September 8-10, 2009*, http://www.nilc.icmc.usp.br/til/stil2009_English/Proceedings/stil/Barufaldi-57866_1.pdf.
- Barufaldi, Bruno, Eduardo F. Santana, José Rogério B. B. Filho, Jan Kees van der Poel, Milton Marques Júnior & Leonardo Vidal Batista. 2010. Classificação Automática de Textos por Período Literário Utilizando Compressão de Dados Através do PPM-C. *Linguamática* 2(1). 35–44.
- Bick, Eckhard. 2000. *The Parsing System “Palavras”: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus, Denmark: Aarhus University. Tese de Doutorado.
- Campos, Alex Sander Luiz. 2018. Edições de Machado de Assis: por quê, para quê? *Machadiana Eletrônica* 1(1). 131–150.
- Freitas, Cláudia, Bianca Freitas & Diana Santos. 2016. QUEMDISSE?: Reported speech in Portuguese. Em *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 4410–4416.
- Freitas, Cláudia, Diana Santos, Cristina Mota, Bruno Carriço & Heidi Jansen. 2015. O léxico do corpo e anotação de sentidos em grandes corpora: o projeto E squeue. *Revista de Estudos da Linguagem* 23(3). 641–680.

- Galves, Charlotte & Pablo Faria. 2010. Tycho Brahe Parsed Corpus of Historical Portuguese. <http://www.tycho.iel.unicamp.br/~tycho/corpus/en/index.html>.
- Higuchi, Suemi, Diana Santos, Cláudia Freitas & Alexandre Rademaker. 2019. Distant reading Brazilian history. Em Constanza Navarreta, Manex Agirrezabal & Bente Maegard (eds.), *Proceedings of 4th Conference of The Association Digital Humanities in the Nordic Countries (Copenhagen, March 6-8 2019)*, 190–200.
- Jockers, Matthew L. 2013. *Macroanalysis: Digital methods and literary history*. University of Illinois Press.
- McCallum, Andrew Kachites. 2002. Mallet: A machine learning for language toolkit. <http://mallet.cs.umass.edu>.
- Mittmann, Adiel, Aldo von Wangenheim & Alckmar Luiz dos Santos. 2016. A system for the automatic scansion of poetry written in portuguese. Em *Computational Linguistics and Intelligent Text Processing, 17th International Conference, CICLing 2016, Konya, Turkey, April 3–9, 2016, Revised Selected Papers, Part II*, 611–628. Springer.
- Moretti, Franco. 2000. Conjectures on world literature. *New Left review* 54–68.
- Ore, Christian-Emil, J. Berenike Herrmanns, Carolin Odebrecht & Diana Santos. 2020. ELTeC: a comparable corpus of novels in many European literatures. Em *DHN2020*, .
- Quintale Neto, Flávio. 2007. *Idéias estéticas e filosóficas nos romances O Ateneu, de Raul Pompéia e Die Verrirungen des Zöglings Törless, de Robert Musil*: USP, São Paulo. Tese de Doutorado.
- R Development Core Team. 2008. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.
- Reis, Carlos. s/d. Trajeto literário. <https://queirosiana.wordpress.com/trajeto-literario/>.
- Santos, Diana. 2019a. Distant reading health: A pilot study on health and disease in lusophone literature. Illness and disability in literary and cultural texts: an international seminar (Univ. de Oslo, 17 June 2019). <https://www.linguateca.pt/Diana/download/DRHealth.pdf>.
- Santos, Diana. 2019b. Literature studies in Lite-rateca: between digital humanities and corpus linguistics. Em Martin Doerr, Øyvind Eide & Oddrun Grønvik ans Bjørghild Kjelsvik (eds.), *Humanists and the digital toolbox: In honour of Christian-Emil Smith Ore*, 89–109. Novus forlag.
- Santos, Diana. 2019c. PANTERA: a parallel corpus to study translation between Portuguese and Norwegian. *Bergen Language and Linguistics Studies* 10(1). doi:DOI10.15845/bells.v10i1.1372.
- Santos, Diana, Cláudia Freitas & Eckhard Bick. 2018a. Obras: a fully annotated and partially human-revised corpus of brazilian literary works in the public domain. Em *Latin American and Iberian Languages Open Corpora Forum (OpenCor) 2018*, .
- Santos, Diana, Cláudia Freitas & João Marques Lopes. 2018b. Ler e estudar a literatura lusófona como parte da literatura mundial: recursos para leitura distante em português. Em Suemi Higuchi & Cláudio José Silva Ribeiro (eds.), *HdRio2018 – I Congresso Internacional em Humanidades Digitais no Rio de Janeiro (HdRio2018)*, 375–383.
- Santos, Diana, Augusto Soares da Silva & Cristina Mota. 2011. Guarda-fatos: notas sobre a anotação do campo semântico do vestuário em português. Relatório técnico. Linguat-eca. <http://www.linguateca.pt/acesso/GuardaFatos.pdf>.
- Santos, Diana & Alberto Simões. 2019. Towards a computational environment for studying literature in portuguese. Apresentação em DH Budapest 2019, Digital Humanities Conference (Budapest, 25-27 de setembro de 2019).
- Silva, Rosário & Diana Santos. 2012. Arco-íris: notas sobre a anotação do campo semântico da cor em português. <http://www.linguateca.pt/acesso/ArcoIris.pdf>.
- Simões, João Gaspar. 1967. *História do Romance Português*. Estúdios Cor.
- Zampieri, Marcos & Martin Becker. 2013. Colonia: Corpus of Historical Portuguese. *ZSM Studien* 5. 77–84. Special Volume on Non-Standard Data Sources in Corpus-Based Research.

Lista de textos

1843 *O Bobo*, de Alexandre Herculano

1844 *A Moreninha*, de Joaquim Manuel de Macedo

- 1844** *Eurico o Presbítero*, de Alexandre Hercu-
lano
- 1845** *O Arco de Santana*, de J. B. da Silva L. de
Almeida Garrett
- 1845** *O moço louro*, de Joaquim Manuel de Ma-
cedo
- 1846** *O Galego*, de Alexandre Herculano
- 1846** *Viagens na Minha Terra*, de J. B. da Silva
L. de Almeida Garrett
- 1848** *O Monge de Cister I*, de Alexandre Hercu-
lano
- 1848** *O Monge de Cister II*, de Alexandre Her-
culano
- 1848** *Os Dois Amores*, de Joaquim Manuel de
Macedo
- 1851** *Anátema*, de Camilo Castelo Branco
- 1851** *O Pároco de Aldeia*, de Alexandre Hercu-
lano
- 1852** *Memórias de um sargento de milícias*, de
Manuel de Almeida
- 1853** *Coisas que só eu sei*, de Camilo Castelo
Branco
- 1854** *A Filha do Arcediágo*, de Camilo Castelo
Branco
- 1854** *Helena*, de J. B. da Silva L. de Almeida
Garrett
- 1854** *Mistérios de Lisboa I*, de Camilo Castelo
Branco
- 1854** *Mistérios de Lisboa II*, de Camilo Castelo
Branco
- 1854** *Mistérios de Lisboa III*, de Camilo Castelo
Branco
- 1855** *Livro Negro de Padre Dinis I*, de Camilo
Castelo Branco
- 1855** *Livro Negro de Padre Dinis II*, de Camilo
Castelo Branco
- 1855** *O Cura de São Lourenço*, de M M S A e
Vasconcelos
- 1856** *Carolina*, de Casimiro de Abreu
- 1856** *Onde Esta a Felicidade*, de Camilo Castelo
Branco
- 1856** *Um Homem de Brios*, de Camilo Castelo
Branco
- 1857** *A viuvinha*, de José de Alencar
- 1857** *O Guarani*, de José de Alencar
- 1857** *O soldado de Aljubarrota*, de Matilde Isa-
bel de Santana e Vasconcelos Moniz Betten-
court
- 1857** *Os tripeiros: Crónica do século XIV*, de
António José Coelho Lousada
- 1858** *A Vingança*, de Camilo Castelo Branco
- 1858** *O Que Fazem Mulheres*, de Camilo Castelo
Branco
- 1859** *Maria ou a menina roubada*, de Antônio
Gonçalves Teixeira e Souza
- 1859** *Úrsula*, de Maria Firmina dos Reis
- 1861** *A chave do enigma*, de António Feliciano
de Castilho
- 1861** *Romance dum Homem Rico*, de Camilo
Castelo Branco
- 1862** *Amor de Perdição*, de Camilo Castelo
Branco
- 1862** *Coisas Espantosas*, de Camilo Castelo
Branco
- 1862** *Coração Cabeça e Estômago*, de Camilo
Castelo Branco
- 1862** *Infestas Aventuras de Mestre Marçal Es-
touro: Vítima duma paixão*, de José da Silva
Mendes Leal
- 1863** *Adelina*, de Ana Plácido
- 1863** *Aventuras de Basílio Fernandes Enxer-
tado*, de Camilo Castelo Branco
- 1863** *O Bem e o Mal*, de Camilo Castelo Branco
- 1864** *A Filha do Doutor Negro*, de Camilo Cas-
telo Branco
- 1864** *A pálida estrela*, de Bulhão Pato
- 1864** *Amor de Salvação*, de Camilo Castelo
Branco
- 1864** *No Bom Jesus do Monte*, de Camilo Cas-
telo Branco
- 1864** *Vinte Horas de Liteira*, de Camilo Castelo
Branco
- 1865** *Iracema, lenda do Ceará*, de José de Alen-
car
- 1866** *A Queda dum Anjo*, de Camilo Castelo
Branco
- 1866** *A conquista de Lisboa*, de Carlos Pinto de
Almeida
- 1866** *Os trabalhadores do mar*, de Machado de
Assis

- 1867 *A Doida do Candal*, de Camilo Castelo Branco
- 1867 *As Pupilas do Senhor Reitor*, de Júlio Dinis
- 1867 *Henriqueta*, de Maria Peregrina de Sousa
- 1868 *A Morgadinha dos Canaviais*, de Júlio Dinis
- 1868 *O Retrato de Ricardina*, de Camilo Castelo Branco
- 1868 *O ermitão do Muquém*, de Bernardo Guimarães
- 1868 *Uma Família Inglesa*, de Júlio Dinis
- 1869 *A luneta mágica*, de Joaquim Manuel de Macedo
- 1869 *Os Brilhantes do Brasileiro*, de Camilo Castelo Branco
- 1870 *A Rosa do Adro*, de Manuel Maria Rodrigues
- 1870 *A ermida de Castromino*, de Antonio Augusto Teixeira de Vasconcellos
- 1870 *A pata da gazela*, de José de Alencar
- 1870 *As apreensões de uma mãe*, de Júlio Dinis
- 1870 *Justiça de Sua Majestade*, de Júlio Dinis
- 1870 *Mistério da Estrada de Sintra*, de José Maria Eça de Queirós
- 1870 *O gaúcho*, de José de Alencar
- 1870 *Oliver Twist*, de Machado de Assis
- 1870 *Os romances da tia Filomela*, de Júlio Dinis
- 1870 *Uma flor de entre o gelo*, de Júlio Dinis
- 1871 *Herança de lágrimas*, de Lopo de Sousa
- 1871 *Os Fidalgos da Casa Mourisca*, de Júlio Dinis
- 1872 *A Infanta Capelista*, de Camilo Castelo Branco
- 1872 *Inocência*, de Visconde de Taunay
- 1872 *O Carrasco de Vitor Hugo*, de Camilo Castelo Branco
- 1872 *O seminarista*, de Bernardo Guimarães
- 1872 *Ressurreição*, de Machado de Assis
- 1873 *A alma de Lázaro*, de José de Alencar
- 1873 *A filha do Cabinda*, de Alfredo Campos
- 1873 *O Anel Misterioso: Scenas da Guerra Peninsular*, de Alberto Pimentel
- 1873 *Um conto portuguez: episódio da guerra civil: a Maria da Fonte*, de Miguel J T Mascarenhas
- 1874 *A Mão e a Luva*, de Machado de Assis
- 1874 *Ubijarara*, de José de Alencar
- 1875 *A Escrava Isaura*, de Bernardo Guimarães
- 1875 *A Filha do Regicida*, de Camilo Castelo Branco
- 1875 *A Freira no Subterraneo*, de Camilo Castelo Branco
- 1875 *A senhora viscondessa*, de S de Magalhães Lima
- 1875 *Novelas do Minho I*, de Camilo Castelo Branco
- 1875 *O Crime do Padre Amaro*, de José Maria Eça de Queirós
- 1875 *O sertanejo*, de José de Alencar
- 1875 *Os selvagens*, de Francisco Gomes de Amorim
- 1875 *Senhora*, de José de Alencar
- 1876 *A Caveira da Mártir*, de Camilo Castelo Branco
- 1876 *Helena*, de Machado de Assis
- 1876 *O Cabeleira*, de Franklin Távora
- 1876 *O Christão novo*, de Diogo de Macedo
- 1877 *Alice*, de Maria Amália Vaz de Carvalho
- 1877 *Novelas do Minho II*, de Camilo Castelo Branco
- 1878 *A Tragédia da Rua das Flores*, de José Maria Eça de Queirós
- 1878 *Iaiá Garcia*, de Machado de Assis
- 1878 *O Matuto*, de Franklin Távora
- 1878 *O Primo Basílio*, de José Maria Eça de Queirós
- 1879 *Eusébio Macário*, de Camilo Castelo Branco
- 1879 *O Romance da Rainha Mercedes*, de Alberto Pimentel
- 1879 *O Sacrifício*, de Franklin Távora
- 1879 *Uma lágrima de mulher*, de Aluisio Azevedo
- 1880 *A Corja*, de Camilo Castelo Branco
- 1880 *O Mandarim*, de José Maria Eça de Queirós

- 1881 *Memórias póstumas de Brás Cubas*, de Machado de Assis
- 1881 *O Mulato*, de Aluisio Azevedo
- 1882 *A Brasileira de Prazins*, de Camilo Castelo Branco
- 1882 *A Condeca Vésper ou Memórias de um Condenado*, de Aluisio Azevedo
- 1882 *As jóias da Coroa*, de Raul Pompéia
- 1882 *Girandula de Amores ou Mistério da Tijuca*, de Aluisio Azevedo
- 1882 *O alienista*, de Machado de Assis
- 1882 *Uma tragédia no Amazonas*, de Raul Pompéia
- 1884 *Casa de Pensao*, de Aluisio Azevedo
- 1884 *Filomena Borges*, de Aluisio Azevedo
- 1885 *Casa velha*, de Machado de Assis
- 1886 *O Brasileiro Soares*, de Luís Magalhães
- 1886 *Quincas Borba*, de Machado de Assis
- 1886 *Vulções de Lama*, de Camilo Castelo Branco
- 1887 *A Relíquia*, de José Maria Eça de Queirós
- 1887 *O Homem*, de Aluisio Azevedo
- 1888 *A Carne*, de Júlio Ribeiro
- 1888 *Mais Uma*, de Conde de Ficalho
- 1888 *O Ateneu*, de Raul Pompéia
- 1888 *Os Maias*, de José Maria Eça de Queirós
- 1888 *Uma Eleição Perdida*, de Conde de Ficalho
- 1889 *No declínio*, de Visconde de Taunay
- 1889 *O Coruja*, de Aluisio Azevedo
- 1890 *O Cortiço*, de Aluisio Azevedo
- 1891 *As Minas de Salomão*, de José Maria Eça de Queirós
- 1891 *Dona Guidinha do Poço*, de Manuel de Oliveira Paiva
- 1891 *O Barão de Lavos*, de Abel Botelho
- 1891 *O missionário*, de Inglês de Sousa
- 1891 *O último cartuxo da Scala Caeli de Évora: Romance histórico (1808-1865)*, de António Francisco Barata
- 1892 *Noites de Cintra*, de Alberto Pimentel
- 1892 *O Dr. Luiz Sandoval*, de Alice Moderno
- 1893 *A Normalista*, de Adolfo Caminha
- 1894 *A Mortalha de Alzira*, de Aluisio Azevedo
- 1895 *A viúva Simões*, de Júlia Lopes de Almeida
- 1895 *Miragem*, de Coelho Neto
- 1895 *O Bom-Crioulo*, de Adolfo Caminha
- 1895 *O Livro de uma Sogra*, de Aluisio Azevedo
- 1895 *O mundo no ano 3000*, de Pedro José Supico de Moraes
- 1896 *Tentacao*, de Adolfo Caminha
- 1897 *Pero da Covilhan: Episódio Romântico do Século XV*, de Zeferino Norberto Gonçalves Brandão
- 1898 *A descoberta e conquista da Índia pelos portugueses: romance historico*, de Artur Lobo d'Avila
- 1899 *A afilhada*, de Manuel de Oliveira Paiva
- 1899 *A conquista*, de Coelho Neto
- 1899 *Dom Casmurro*, de Machado de Assis
- 1899 *Elle*, de Claudia de Campos
- 1899 *Transviado*, de Jayme de Magalhães Lima
- 1900 *A Ilustre Casa de Ramires*, de José Maria Eça de Queirós
- 1900 *Fradique Mendes*, de José Maria Eça de Queirós
- 1900 *O exilado*, de Maurícia C de Figueiredo
- 1901 *A Cidade e as Serras*, de José Maria Eça de Queirós
- 1901 *A falência*, de Júlia Lopes de Almeida
- 1901 *Amanhã*, de Abel Botelho
- 1903 *A Farsa*, de Raúl Brandão
- 1904 *Esaú e Jacó*, de Machado de Assis
- 1904 *Os filhos do padre Anselmo*, de António da Costa Couto Sá de Albergaria
- 1904 *Turbilhão*, de Coelho Neto
- 1904 *Viriato*, de Teófilo Braga
- 1905 *A Ala dos Namorados*, de António Campos Junior
- 1905 *A Intrusa*, de Júlia Lopes de Almeida
- 1906 *A Divorciada*, de José Augusto Vieira
- 1906 *A Lenda da Meia-Noite*, de Manuel Joaquim Pinheiro Chagas
- 1906 *Os Bravos do Mindelo*, de Faustino da Fonseca
- 1906 *Os Pobres*, de Raúl Brandão

- 1908** *A Casa dos Fantasma*s, de Luís Augusto Rebelo da Silva
- 1908** *A feiticeira*, de Ana de Castro Osório
- 1908** *A vinha*, de Ana de Castro Osório
- 1908** *Diário de uma criança*, de Ana de Castro Osório
- 1908** *Memorial de Aires*, de Machado de Assis
- 1908** *Sacrificada*, de Ana de Castro Osório
- 1909** *O Salústio Nogueira*, de Teixeira de Queirós
- 1909** *Recordações do escrivo*ã Isaías Caminha, de Lima Barreto
- 1910** *Maria Dusá*, de Lindolfo Rocha
- 1911** *Triste Fim de Policarpo Quaresma*, de Lima Barreto
- 1913** *A Confissão de Lúcio*, de Mário de Sá-Carneiro
- 1914** *A Marquesa de Vale Negro*, de Maria O'Neill
- 1914** *Por bom caminho*, de Maria O'Neill
- 1915** *A capital federal*, de Coelho Neto
- 1915** *A engomadeira: novela vulgar lisboeta*, de José Sobral de Almada Negreiros
- 1916** *A morte vence*, de João José Grave
- 1916** *Decameron*, de Virgínia de Castro e Almeida
- 1916** *Innocente*, de Virgínia de Castro e Almeida
- 1916** *O Solar dos Pavões*, de Virgínia de Castro e Almeida
- 1919** *Amor crioulo*, de Abel Botelho
- 1919** *Húmus*, de Raúl Brandão