Comparando a literatura lusófona com outras literaturas: recursos para leitura a distância em português

Comparing lusophone literature with other literatures: resources for distant reading in Portuguese

Diana Santos [1]

Cláudia Freitas [2]

João Marques Lopes [3]

Resumo. *Após apresentar alguns recursos existentes para o estudo da literatura lusófona, nomeadamente os corpos OBras, Vercial, NOBRE, Tycho Brahe e Colonia, explicamos como juntos dão origem ao que chamamos Literateca. Apresentamos de seguida o plano de criação de um recurso multilíngue para comparar várias literaturas no âmbito da ação COST 16204, DISTANT READING, financiada pela União Europeia. Terminamos sugerindo brevemente vários usos literários dos mesmos recursos e relatando trabalho em progresso.*
Palavras-chave: *literatura lusófona, leitura distante, leitura a distância, literaturas de língua portuguesa, corpos linguísticos*

Abstract. *After presenting some resources for studying literature in Portuguese (lusophone literature), namely the following corpora: Vercial, OBras, NOBRE, Tycho Brahe and Colonia, we explain what we call Literateca. We introduce the EU-funded COST action 16204 (DISTANT READING) and its aim to compare different European literatures. We end by describing some literary uses of these corpora and work in progress.*
Keywords: *Lusophone literature, literature in Portuguese, Portuguese literature, Brazilian literature, distant reading, language corpora.*

1. Corpos literários

A literatura é um dos melhores exemplos de língua (os escritores são, por definição, aqueles que a manejam melhor), e por isso há muitos corpos[4] linguísticos que contêm textos literários, sobretudo se são paralelos, visto que a literatura é o género de texto mais traduzido.

[1]Doutora, Linguateca e Universidade de Oslo, Portugal e Noruega; d.s.m.santos@ilos.uio.no

[2]Doutora, PUC-Rio e Linguateca, Brasil, Portugal; claudiafreitas@puc-rio.br

[3]Doutor, Linguateca e Universidade de Lisboa, Portugal; marqueslopes1928@hotmail.com

[4] Ao longo do artigo, e propositadamente, usamos a grafia corpo/corpos, como advogado em Santos (2008) e a grafia corpus/corpora. Considerando a co-autoria e as diferentes nacionalidades dos autores, optamos por não uniformizar o texto segundo uma das variantes (do Brasil ou de Portugal).

Contudo, até há muito pouco tempo os corpos literários não eram criados para responder especificamente a perguntas de viés literário. Muito pelo contrário, eram apenas empregues para problemas linguísticos, mesmo quando esses problemas linguísticos tinham como objetivo iluminar a literatura.

Tal tem começado a mudar com os estudos propostos por Franco Moretti (Moretti, 2000) e Mathew Jockers (Jockers, 2013) dando ensejo a uma nova disciplina de leitura a distância, ou leitura distante, ou, ainda, leitura distanciada[5]. Para esses autores, os problemas que pretendem estudar são genuinamente literários, a nível de escala e a nível de perguntas, e debruçam-se sobre a literatura ou as literaturas como um todo e não sobre obras ou autores em particular. (Embora não excluam, evidentemente, a consideração de obras e assuntos específicos.) Usando a metáfora do macroscópio, instrumento conceitual que permite ver um dado objeto ou paisagem a diferentes distâncias, e àquela que for pertinente para a pergunta em questão, a leitura a distância permitirá responder a perguntas sobre grandes quantidades de obras literárias, tal como a linguística com corpos antes dela permitiu responder a tendências e a caracterizações de tipos de textos distintos recorrendo à quantidade.

Para podermos ler a distância em português, precisamos de ter bastantes obras literárias acessíveis e identificadas como obras literárias, e é por essa necessidade básica que começamos, portanto, o nosso artigo: a descrição dos corpos literários já existentes para a nossa língua, e a sua junção num projeto infraestrutural que os reúne e caracteriza.

1.1. O corpo Vercial

Este corpo foi obtido usando a digitalização de obras de literatura portuguesa pelo projeto Vercial, através da autorização do seu promotor, Leon de Carvalho, concedida à Linguateca em 1999. Informação detalhada sobre o conteúdo, com base em que versões, e revista por quem, encontra-se em https://www.linguateca.pt/acesso/lista_de_obras_Vercial.txt

1.2. O corpo OBras

O corpo OBras (Obras Brasileiras) resultou da nossa vontade de termos um acervo equivalente para a literatura brasileira, e foi iniciado num sistema de voluntariado entre Anya Campos e as duas autoras deste artigo, que conseguiram alunos e bolsistas para fazer esse

[5] Nenhuma destas traduções parece consensual. Para uma das autoras, “a distância” , como em “ensino a distância”, atribui frequentemente à dimensão não-presencial um caráter paliativo, trazendo a conotação de “maneira remediada de leitura”. Para outra das autoras, “distanciada” remete à palavra “distanciamento”, atitude frequentemente invocada em atividades analíticas e críticas, dando a impressão errada de que é ou deve ser menos emocional. E o mesmo se verifica em relação à palavra “distante”. Ora a nossa concepção de “distant reading” denota apenas uma outra maneira de olhar e analisar a literatura, complementar a um “close reading”. Por isso, e dada este desconforto com os vários termos possíveis, é bem provável que ainda não tenha sido encontrada a palavra certa para esta atividade em língua portuguesa!

trabalho. Este corpo está portanto em constante atualização e aumento, veja-se https://www.linguateca.pt/acesso/lista_de_obras_OBras.txt

1.3 O corpo NOBRE

Este corpo, denominado Novas OBRas publicadas na Europa, foi criado já em 2018 para resolver o problema de aumentar o número de obras de literatura portuguesa ou lusófona, vindas de outras origens que não o projeto Vercial, projeto de que recebemos todas as obras na mesma altura, em 1999. Uma das explicações para o título provém da iniciativa COST chamada *Distant Reading for European Literary History* em que é necessário, como explicado na secção 2, coligir obras em português que tenham sido publicadas na Europa (quer como primeira edição, quer até dez anos depois da publicação noutro país, como poderá ser o caso de vários clássicos brasileiros). Da mesma forma que os dois corpos anteriores, a sua composição está em https://www.linguateca.pt/acesso/lista_de_obras_NOBRE.txt

1.4 O corpo Tycho Brahe

O Corpus Histórico do Português Tycho Brahe é um corpo eletrônico anotado, composto de textos em português escritos por autores nascidos entre 1380 e 1845, compilado pela Universidade de Campinas (Galves et al., 2017). A sua página principal é http://www.tycho.iel.unicamp.br/corpus/index.html. Ainda que o objetivo deste corpo não fosse primordialmente o estudo da linguagem literária, contém muitas obras relevantes nessa categoria, além de ter um cuidado filológico muito salutar em relação aos textos e às versões usadas. Selecionamos deste corpo as obras de cariz narrativo, dissertativo e dramatúrgico que consideramos fazerem parte do espólio literário lusófono, embora não necessariamente ficcionais.

1.5 O corpo COLONIA

O corpo COLONIA é um corpo eletrônico anotado compilado para pesquisa sobre a história da língua portuguesa, com textos escritos entre 1500 e 1936, desenvolvido pela Universidade de Colónia (Zampieri & Becker, 2013). A sua página principal é http://corporavm.uni-koeln.de/colonia/. Existem muitos textos que também fazem parte dos corpos anteriores, o que levou à criação da infraestrutura Literateca, explicada na secção 3.

2. A ação COST 16204, para a história da(s) literatura(s) europeia(s)

Os presentes autores encontram-se associados, em graus diversos, à ação COST *Distant Reading for European Literary History* (COST 16204), descrita em www.distant-reading.net. As ações COST são iniciativas financiadas pela União Europeia

para apoiar projetos com vários intervenientes, cujos objetivos necessitem da perícia e do conhecimento existente em diferentes países e línguas. Tal serviu como catalisador para iniciar a criação de mais um corpo literário, neste caso multilíngue e também multiliterário, que permitisse comparar e desenvolver estudos sobre várias literaturas, e contribuísse também para o desenvolvimento de ferramentas computacionais e para a proposta de novas hipóteses sobre a literatura.

Inspirada pelos já citados trabalhos de Moretti, Jockers e pelo de Schöch (o principal proponente dessa ação, ver Schöch (2017)), uma das vertentes desse trabalho é criar o ELTEC com cem (ou mais) romances e novelas (“novels” em inglês) num conjunto apreciável de línguas e literaturas. Embora como projeto europeu tenha o foco na literatura europeia, foi nossa opinião de que não poderíamos deixar passar esta oportunidade para também estudar a literatura lusófona, ou melhor, todas as literaturas em português, construindo recursos semelhantes como parte do projeto, e aplicando os mesmos métodos e ferramentas.

Duas razões principais temos para esta postura: a primeira, a de que as literaturas “nacionais” não coexistem necessariamente com independências políticas. Por exemplo, as literaturas angolana e moçambicana podem ter um traço africano muito antes da independência destes países. Por outro lado, o facto de serem publicados na Europa (neste caso, obras africanas ou timorenses em Portugal ou, no caso de muitos livros brasileiros, publicados inicialmente em França) não legitima a consideração destes como parte da literatura europeia.

A outra, e mais importante do ponto de vista prático, tem a ver com a filosofia da Linguateca, que é um projeto de recursos para a língua portuguesa, independentemente da sua variedade ou variante nacional; veja-se, por exemplo, Santos (2009) ou Santos (2015). Estamos convencidos de que a língua portuguesa, que os escritores lusófonos têm em comum, os aproxima mais do que a qualquer outra literatura, por mais importante e marcante que esta seja. Essa é uma convicção que poderá ser confirmada ou infirmada com este projeto. Mas, de um ponto de vista de engenharia, as ferramentas podem e devem ser as mesmas. Por isso, para nós faz todo o sentido considerar a literatura lusófona como um todo, para depois observar as diferenças segundo vários eixos (temporais, de género, de escola literária, etc.), e não só ou especialmente o nacional.

Daí resulta que, embora o COST tenha de produzir uma coleção exclusivamente europeia que se referirá provavelmente a obras apenas de autores portugueses, e que ficará muito provavelmente a cargo dos representantes de Portugal na ação COST, nós propomo-nos

criar, simultaneamente, uma coleção alargada (“extended collection” na terminologia do projeto), que abranja e represente a literatura lusófona, e não apenas a literatura portuguesa. É sobre essa coleção que nos debruçamos aqui, começando por explicar o que é a Literateca, que será a infraestrutura na qual basearemos a nossa coleção.

3. A Literateca, ambiente para o estudo da literatura em português[6]

A Literateca pode ser descrita como uma infraestrutura que seleciona as obras literárias dos corpos acima referidos (e de outros corpos que aqui não serão discutidos, porque englobam apenas excertos de obras), garantindo a não duplicação de textos e o seu tratamento uniforme, com uma série de ferramentas associadas.

Todo este acervo, anotado sintaticamente pelo PALAVRAS (Bick, 2000), contendo informação sobre vários campos semânticos descritos noutras publicações, assim como indicações de autor, obra, data e género literário, pode ser consultado e usado para estudos de literatura de língua portuguesa. Parece-nos, por isso, importante apresentá-lo aqui, indicando que está em contínuo desenvolvimento, aberto a mais obras e a melhoria e adição nas anotações. Consideramos a Literateca mais do que um corpo, visto que é um ambiente especializado de consulta e pesquisa, inspirado na Gramateca (Santos, 2014).

De momento (junho de 2018), a Literateca contém 716 obras distintas (de 169 autores diferentes), incluindo crônicas, cartas (como a do Descobrimento do Brasil), sermões e mesmo atas de congregações (coligidas pelo projeto Tycho Brahe); contém atualmente cerca de 200 obras em prosa (romances e novelas) publicadas no período 1850-1919 (o período sobre o qual a ação COST se debruça). Uma questão associada é justamente a contabilização das obras, quando temos formas originalmente publicadas como coletâneas: em um livro de novelas[7], por exemplo, devemos considerar como obra o livro completo, ou cada novela constitui uma obra? E quando um romance foi publicado em vários volumes, é uma obra só, ou são tantas quantos os volumes?

Num primeiro momento, é possível fazer uma série de contagens sobre os textos que depois podem ser alvo de tratamento estatístico, veja-se Santos (2008); outras tecnologias, como os modelos de tópicos (“topic models”) e as redes de personagens, estão de momento sendo desenvolvidas de forma a também serem disponibilizadas como um serviço ao usuário.

[6] É certo que neste caso estamos a estudar apenas a literatura escrita originalmente em português, mas a Literateca, por incluir nomeadamente também já algumas traduções para português efetuadas por escritores portugueses e brasileiros, pode ser mais geralmente descrita como literatura em português (original ou traduzida).

[7] Veja-se, por exemplo, as *Novelas do Minho* de Camilo Castelo Branco, ou os *Serões da Província* de Júlio Dinis.

Aqui vamos nos concentrar sobre algumas anotações que estamos, neste preciso momento, adicionando.

3.1 A indicação da escola literária

Uma das variáveis que pode ser relevante para o estudo da história da literatura é exatamente a corrente literária em que uma dada obra se enquadra, e essa foi uma das nossas primeiras preocupações. Sabendo que existem opiniões diferentes entre os especialistas, e que, além disso, uma dada obra pode ser incorporada em mais do que uma escola, decidimos marcá-la com todas as correntes literárias nas quais havia sido classificada pelos teóricos.

Como é sobremaneira sabido, o enquadramento das obras na taxinomia dos géneros e dos movimentos literários não é uma tarefa linear. Para ficarmos apenas com um exemplo referente à problemática da inserção de uma obra canónica no respetivo movimento literário, pensemos nos debates e nas dificuldades suscitados por *O Ateneu*, de Raul Pompeia.

Não cabendo fazer aqui a discussão pormenorizada do assunto, limitamo-nos a sinalizar que esse romance tem uma longa recepção crítica em que uns o consideram naturalista ou realista, outros o taxam 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. Em Araújo (2011) e Quintale Neto (2007), encontra-se uma discussão apurada dessas polémicas taxinómicas a respeito de *O Ateneu*, pelo que remetemos o leitor interessado para tais referências. Na Literateca, e não querendo escolher entre as várias escolas, esta obra está marcada como “impressionismo_naturalismo_realismo_simbolismo”.

3.2 A indicação das personagens

Uma questão importante (pelo menos na literatura do período que estamos a tratar) é a identificação das personagens de cada obra, garantindo, adicionalmente, que as várias formas de a identificar são entendidas como uma só personagem. No romance *Dom Casmurro*, por exemplo, não há uma única personagem cuja menção seja feita sempre da mesma maneira. Bentinho é o campeão de alcunhas: *Bentinho*, *Dom Casmurro*, *Santiago*, *Bento*, *Doutor Santiago*, *Padre Bentinho* e *Sr. Bentinho*. Assim, é preciso escolher um nome que funcione como agregador das diferentes menções. Por outro lado, e ainda na mesma obra, tivemos como desafio desambiguar personagens de mesmo nome (pois um foi nomeado em homenagem ao outro), caso de *Ezequiel Escobar* e *Ezequiel Santiago*.

Ainda com relação às personagens, lembramos que nem todas as referências a pessoas, em uma obra, são personagens em sentido estrito: nomes próprios podem se referir a

personagens históricas ou fictícias de outras épocas, evidenciando influências e referências. Detalhamos este ponto a seguir.

3.3 Uma pré-anotação de entidades mencionadas

Contrastando com os textos jornalísticos ou científicos, os nomes próprios em textos literários representam geralmente os seguintes tipos de entidades: personagens fictícias (de literatura ou mitologia), escritores, personagens históricas, por um lado, e localizações, muitas vezes ficcionais. Resolvemos pois fazer uma primeira separação entre três tipos de pessoa: Personagem, Pessoa ficcional e Pessoa histórica[8]. O primeiro tipo refere-se às personagens da obra em questão; pessoas ficcionais referem-se a personagens de outras obras, como *Otelo* e *Ulisses*; e pessoas históricas referem-se a figuras como *Camões* e *Homero*.

Quanto às personagens da obra propriamente, indicamos que consideramos como personagens, numa primeira análise, qualquer pessoa (que não seja nem histórica nem ficcional) mencionada mais do que uma vez ao longo do texto[9]. Também observamos os toponímicos, marcados com o sema Local, que já no HAREM (Mota & Santos 2008) se nos tinham revelado complicados, devido ao uso metonímico frequente.

4. Primeiros estudos

Nesta secção pretendemos tão só mostrar aquilo que já se pode fazer, abrindo um mundo de pesquisas e interrogações.

Começamos por mostrar que a revisão de verbos de dizer (veja-se Freitas et al., 2016) pode ser importante para o estudo de particularidades de escrita de diversos escritores, bem como para a caracterização das próprias personagens conforme apresentadas pelo narrador (a passividade de personagens que *concordam, assentem* e *justificam-se,* em oposição àquelas que *interrompem, propõem, acentuam, teimam, insistem e exclamam*, por exemplo). Tomando por base, novamente, *Dom Casmurro*, e associando a indicação de personagens com os verbos do dizer, vemos a diferença entre a quantidade de falas atribuídas a *Capitu*, por um lado, e a *Bentinho* – narrador da história, por outro[10]. A tabela 1 apresenta a distribuição de falas por personagem.

[8] Na sintaxe de anotação do AC/DC, os três tipos estão codificados no atributo sema da seguinte maneira: Pessoa:ficc, Pessoa:hist, e Pessoa:PERSONAGEM:NomedaPersonagem, em que o nome da personagem refere-se ao seu identificador único, que escolhemos de forma a minimizar a sobreposição com outras obras.

[9] Conscientes do que esta concetualização e taxinomia de "personagem" resultam de uma aproximação básica, remetemos o leitor interessado numa definição própria da narratologia para Reis e Lopes (1988: 215-219).

[10] No apêndice apresentamos todas as expressões de busca usadas ao longo da preparação do artigo.

Tabela 1: Distribuição de falas por personagem em Dom Casmurro[11]

Personagem	Número de verbos de elocução associados
Bentinho	**24 (37%)**
Capitu	**13 (20%)**
José Dias	**10 (15%)**
Escobar	**7 (10%)**
Padre Cabral	**5 (7%)**
Fortunata	**1 (1,5%)**
Dona Glória	**1(1,5%)**
Sancha	**1(1,5%)**
Ezequiel Santiago	**1(1,5%)**
Cosme	**1(1,5%)**
Sr.Pádua	**1(1,5%)**

A existência, no contexto da Gramateca, da anotação do corpo humano (veja-se Freitas et al., 2015) também agiliza a busca por predicadores humanos. A figura 1 apresenta a distribuição das palavras do corpo humano nas obras *Memórias póstumas de Brás Cubas*, *Quincas Borba* e *Dom Casmurro*. Se quisermos aproximar o olhar (embora lendo a distância) vemos que para *olho*, uma das partes do corpo humano mais mencionada nestas obras, os adjetivos que o qualificam dizem respeito, principalmente, a traços psicológicos das personagens. A figura 2 apresenta os adjetivos modificadores de *olho*, respectivamente, nas três obras.

Figura 1: Distribuição das palavras do corpo humano em *Memórias Póstumas de Brás Cubas, Quincas Borba* e *Dom Casmurro*

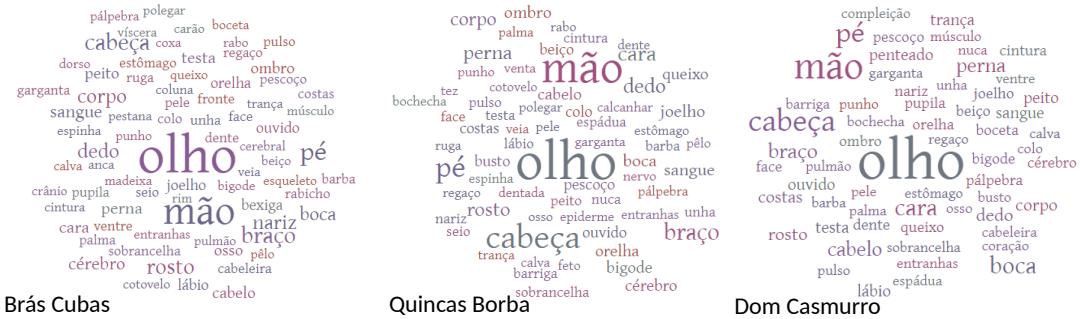


Figura 2: Distribuição dos modificadores adjetivos de *olho* em Quincas Borba, Memórias Póstumas... e Dom Casmurro

[11] Estes dados são referentes à versão 4.1 do OBras, e foram obtidos no dia 3 de junho de 2018.

Ainda quanto aos predicadores humanos, buscamos por estruturas sintáticas capazes de indicar caracterizações humanas, como predicativos e apostos, como ilustra a figura 3:

Figura 3: Linhas de concordância de predicadores humanos

id="Grãos_de_mostarda prosa:conto HC 1926": A Carminha era **bonita e nova**.
id="O_Mulato prosa:romance AA 1881": Sabemos que **ela está tão pura** como dantes!
id="Ressurreição prosa:romance MdA 1872": Há dias em que se levanta meiga e alegre, outros em que toda **ela é irritação e melancolia** .
id="Uma_lágrima_de_mulher prosa:romance AA 1880": E **Rosalina, meiga**, encarava com chorosa ternura o olhar sombrio de Miguel

Apresentamos na figura 4 caracterizadores humanos associados a personagens masculinas e femininas em alguns romances de Machado de Assis.

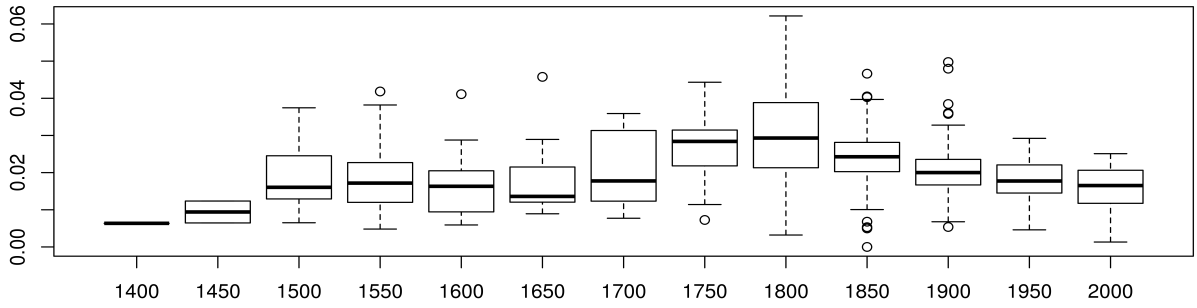
Figura 4: Distribuição de predicadores humanos quanto ao gênero da personagem

Outra exploração possível é procurar a existência de palavras ou expressões remetentes para obras anteriores, ou para fontes de inspiração, como é o caso de a menção de personagens de Camões ou de Shakespeare em obras mais modernas. Ou de sátiras ou comentários desdenhosos a rivais (a influência, afinal, pode ser positiva ou negativa). É interessante reparar que, apenas nos 9 romances de Machado de Assis na Literateca, existem referências a 27 pessoas históricas (a maioria autores!) e a 16 personagens literárias (incluindo personagens bíblicas).

Outra possibilidade é a marcação de algumas questões específicas (por exemplo a maternidade, ou a religião), além dos tópicos eventualmente obtidos automaticamente. Já marcámos um campo saúde, que pode ser interessante explorar: verificar a correlação deste campo com os sentimentos, com o corpo humano, e com os avanços da ciência médica e psiquiátrica, dado que temas como a loucura ou personagens médicas serem bastante frequentes na literatura em português ao longo dos tempos – na época abrangida pelo COST, veja-se *O Alienista* ou *As Pupilas do Senhor Reitor*.

A questão da assinatura emocional, de um autor, de uma época ou de uma obra, é outro assunto que poderá contribuir para os estudos literários. E, como mostramos na figura 5 e argumentamos em Santos e Maia (2018), a diferença de atenção dada a diferentes emoções em diferentes períodos (literários) pode também iluminar a própria importância dessas emoções em português.

Figura 5: Referência a emoções ao longo do tempo na literatura lusófona, na Literateca, em valores relativos



Na apresentação oral que realizamos durante o HD-Rio (Santos, Freitas & Lopes, 2018), damos exemplos de mais trabalhos de outros autores, que convidamos a consultar.

5. Considerações finais

Apresentamos aqui os primeiros passos relativos à exploração, através da leitura distante, de uma coleção de obras de literatura de língua portuguesa, que insistimos em caracterizar como literatura lusófona. Descrevemos a necessidade de adicionar marcação e anotação específica de estudos literários, como gênero, escola, personagens, e tipo de nomes próprios envolvidos (pessoas históricas, personagens de ficção consideradas de cultura geral e, portanto, não explicadas, e outras pessoas ou entes, assim como locais).

Apresentamos também, centrando-nos principalmente na obra de Machado de Assis, uma série de caracterizações preliminares com base na anotação semântica existente nos corpos, seja sobre o corpo humano, sobre referência às emoções, ou sobre o próprio discurso.

Gostaríamos de enfatizar que nos encontramos apenas no início do que esperamos ser uma série de descobertas pelos mares nunca dantes navegados da leitura a distância na literatura em português, nas quais queremos convidar todos os leitores a tomar parte.

6. Agradecimentos

O trabalho aqui relatado enquadra-se na ação COST *Distant Reading for European Literary History* (*DISTANT READING*) CA16204, que é financiada pelo programa da comunidade Europeia *Horizon 2020*. Agradecemos à FCCN portuguesa o alojamento dos recursos nos seus servidores, e à UNINETT Sigma2, *the National Infrastructure for High Performance Computing and Data Storage in Norway*, pela atribuição de tempo computacional na rede de computadores abel. Além disso estamos gratos a toda a equipa da Linguateca pela sua existência e pelos recursos que usamos.

7. Referências

ARAÚJO, Francisco Magno da Silva. ***O Ateneu* e a nostalgia da forma.** Dissertação de mestrado, Centro de Ciências Humanas, Letras e Artes, Universidade Federal do Rio Grande do Norte, Natal, 2011.

BICK, Eckhard. **The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework**. Aarhus University Press, 2000.

COST ACTION 16204. Disponível em: <https://e-services.cost.eu/files/domain_files/CA/Action_CA16204/mou/CA16204-e.pdf>. Acesso em 19 nov 2017.

FREITAS, Cláudia; FREITAS, Bianca; SANTOS, Diana. QUEMDISSE?: Reported speech in Portuguese. In: CALZOLARI, Nicoletta et al. (Eds.). **Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)**. ELRA, p. 4410-4416, 2016.

FREITAS, Cláudia; SANTOS, Diana; MOTA, Cristina; CARRIÇO, Bruno; JANSEN, Heidi. O léxico do corpo e anotação de sentidos em grandes corpora: o projeto Esqueleto. **Revista de Estudos da Linguagem**, v.23, n.3, p. 641-680, 2015.

GALVES, Charlotte; Andrade, Aroldo Leal de; and Faria, Pablo. **Tycho Brahe Parsed Corpus of Historical Portuguese.** Dezembro, 2017. Disponível em: <http://www.tycho.iel.unicamp.br/~tycho/corpus/texts/psd.zip>. Acesso em 19 nov 2017.

JOCKERS, Matthew L. **Macroanalysis: Digital methods and literary history**. University of Illinois Press, 2013.

MORETTI, Franco. Conjectures on world literature. **New Left review** 1, Jan-Feb 2000, p. 54-68, 2000.

MOTA, Cristina; SANTOS, Diana (Eds.). **Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM**. 2008. Disponível em: <https://www.linguateca.pt/HAREM/actas/Livro-MotaSantos2008.pdf>

QUINTALE NETO, Flávio. **Idéias estéticas e filosóficas nos romances *O Ateneu*, de Raul Pompéia e *Die Vervirrungen des Zöglings Törless*, de Robert Musil.** Tese de doutorado, USP, São Paulo, 2007.

REIS, Carlos; LOPES, Ana Cristina M. **Dicionário de teoria da narrativa**, São Paulo, Ática, 1988.

SANTOS, Diana. Corporizando algumas questões. In: TAGNIN, Stella E. O. & VALE, Oto Araújo (Eds.), **Avanços da Lingüística de Corpus no Brasil**, Editora Humanitas/FFLCH/USP, São Paulo, p.41-66, 2008.

SANTOS, Diana. Caminhos percorridos no mapa da portuguesificação: A Linguateca em perspectiva. **Linguamática**, 1, 1, Maio de 2009, p. 25-58, 2009.

SANTOS, Diana. Gramateca: corpus-based grammar of Portuguese. In: BAPTISTA, Jorge; MAMEDE, Nuno; CANDEIAS, Sara; PARABONI, Ivandré; PARDO, Thiago A.S. Pardo; NUNES, Maria das Graças Volpe (Eds.), **PROPOR 2014**, LNAI 8775. Heidelberg: Springer, p. 214-219, 2014.

SANTOS, Diana. Corpora at Linguateca: Vision and roads taken. In: BERBER SARDINHA, Tony; FERREIRA, Telma de Lurdes São Bento (Eds.). **Working with Portuguese Corpora**. Bloomsbury, p. 219-236, 2014.

SANTOS, Diana. Português internacional: alguns argumentos. In: TEIXEIRA, José (Ed.). **O Português como Língua num Mundo Global: problemas e potencialidades.** Centro de Estudos Lusíadas da Universidade do Minho, p. 49-66, 2016.

SANTOS, Diana; MAIA, Belinda. Language, emotion, and the emotions: A computational introduction. **Language and Linguistics Compass.** 2018. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/lnc3.12279>

SANTOS, Diana; FREITAS, Cláudia; LOPES; João Marques. Ler e estudar a literatura lusófona como parte da literatura mundial: recursos para leitura distante em português. Apresentação no HD-Rio 2018, Rio de Janeiro, 9 a 13 de abril de 2018. Disponível em: <https://www.linguateca.pt/Diana/download/SantosFreitasLopesHDRio2018.pdf>

SCHÖCH, Christof. Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama. **Digital Humanities Quarterly** v.11, n.2. 2017. Disponível em: <http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html>. Acesso em 19 nov 2017.

ZAMPIERI, Marcos; BECKER, Martin. Colonia: Corpus of Historical Portuguese. In: ZAMPIERI, Marcos; DIWERSY, Sascha (Eds.). **Non-standard Data Sources in Corpus-based Research**, Volume 5 de ZSM-Studien, Schriften des Zentrums Sprachenvielfalt und Mehrsprachigkeit der Universität zu Köln Shaker, p. 77-84, 2013.

8. Apêndice

Comandos usados para a obtenção dos dados, no contexto do serviço AC/DC, escolhendo o corpo Obras: <https://www.linguateca.pt/acesso/corpus.php?corpus=OBRAS>

Em todos os casos, o tipo de resultado pretendido deve ser “Distribuição de lemas”:

Verbos de elocução associados a Capitu:
(@[sema="dizer.*"] [pos!="V.*"]* [pos!="V.*"]* [func="<SUBJ.*" & sema="Pessoa:PERSONAGEM:Capitu"])|([func="SUBJ>" & sema="Pessoa:PERSONAGEM:Capitu"] [pos!="V"]* [pos!="V.*"]* @[sema="dizer.*"]) within s;

Verbos de elocução associados a Bentinho:
(@[sema=".*dizer_.*" & obra="Dom.*" & pessnum="1S"] [pos!="V.*"]* [func!="<SUBJ.*"])|([func!=".*SUBJ.*" & obra="Dom.*"] [pos!="V"]* @[sema=".dizer_.*" & pessnum="1S"]) within s;

Comentários: (1) Como Bentinho é o narrador, a expressão é diferente, e especifica que os verbos devem estar na 1ª pessoa. (2) O fato de a língua portuguesa permitir a omissão do sujeito traz um desafio adicional à associação entre falas e personagens.

Verbos de elocução associados a todas as personagens de Dom Casmurro:
([sema=".*dizer_.*" & obra="Dom.*"] [pos!="V.*"]* @[func="<SUBJ.*" & sema="Pessoa:PERSONAGEM.*"]) |(@[func="SUBJ>" & sema="Pessoa:PERSONAGEM.*" & obra="Dom.*"] [pos!="V"]* [sema=".*dizer_.*"]) within s;

Palavras do corpo em Memórias Póstumas de Brás Cubas:
[sema="corpo" & obra="Memórias.*"]

Modificadores de “olho” em Memórias Póstumas..., Quincas Borba e Dom Casmurro:
([sema="corpo" & autor="MdA" & obra="Memórias.*|Quincas.*|Dom.*" & lema="olho"] [pos!="V.*|N"]* @[func="N<" & pos!="PRP"])|(@[func=">N" & pos="ADJ|V"] [sema="corpo" & autor="MdA" & obra="Memórias.*|Quincas.*|Dom.*" & lema="olho"]) within s;

Algumas expressões para buscar predicadores humanos apenas femininos:
[pos="PROP" & func="SUBJ>"] [lema="ser|estar"] [pos="ADV.*"]* @[temcagr!=".*PASS.*" & pos="ADJ|N|V" & gen="F" & func="<SC"]
[lema="ela" & func="SUBJ>"] [lema="ser|estar"] [pos="ADV.*"]* @[temcagr!=".*PASS.*" & pos="ADJ|N|V" & gen="F" & func="<SC"]
[pos="PROP" & func!="P<"] "," [pos="ADV.*"]* @[func="N<PRED|.*APP.*" & gen="F" & pos="ADJ"]

Expressão para buscar, por exemplo, pessoas ficcionais nos romances de Machado de Assis:
[autor="MdA" & classe="Prosa:romance" & sema=".*Pessoa:ficc.*"].