Ler e estudar a literatura lusófona como parte da literatura mundial: recursos para leitura distante em português

Tackling lusophone literature as part of world literature: resources for distant reading in Portuguese

Resumo. Após apresentar três recursos existentes para o estudo da literatura lusófona, nomeadamente OBras, Vercial e Literateca, e o plano de criação de um recurso multilíngue para comparar várias literaturas no âmbito da ação COST 16204, discutimos possíveis critérios de escolha dessa amostra (100 romances), assim como apresentamos brevemente usos literários dos mesmos recursos.

Abstract. After presenting three existing resources for studying lusophone literature, namely Obras, Vercial and Literateca, and the planned creation of a multilingual literary resource under the scope of COST action 16204, we discuss several possible criteria for an informed choice of 100 novels, as well as present briefly some literary uses of the aforementioned resources.

1. Apresentação

Neste artigo apresentamos três corpos[1] de literatura em português: o OBras, o Vercial, e a Literateca, no âmbito da recém-criada inciativa europeia "Distant Reading for European Languages" (COST Action 16204).

Neste contexto, e com o objetivo de estudar correlações literárias entre literaturas e autores escrevendo em línguas diferentes, será constituído um corpo de obras em dez línguas/literaturas, ao qual serão aplicados métodos de descoberta e de anotação originais, a serem desenvolvidos e orquestrados ao longo do processo. A escolha das obras – estão previstos, na proposta do projeto, 100 romances por língua com datas de publicação entre 1800 e 1920 – será também objeto de estudo e discussão alargada.

1.1. O corpo OBras

Este corpo nasceu da reflexão de que, embora houvesse muito mais texto brasileiro acessível em

1 Ao longo do artigo, utilizamos a grafia "corpo"; "corpos" como defendido em Santos (2008).

termos de outros géneros (como blogues, resenhas, texto didático, correio eletrónico, etc.) na Linguateca, havia uma falta total de texto literário, o que criava um viés em relação a possíveis comparações entre a variante brasileira e a portuguesa, por exemplo. E o projeto OBras (Obras BRASileiras) surgiu como um projeto voluntário entre três atores, a saber: uma universidade brasileira, uma universidade norueguesa e uma pesquisadora brasileira, obtendo, revendo e marcando uma série de obras consideradas canônicas no Brasil, e para as quais o problema dos direitos autorais estava resolvido. Como modelo, usou-se o corpo do projeto Vercial.

1.2. O corpo Vercial

O corpo Vercial foi um dos primeiros criados pela Linguateca, graças ao apoio do projeto Vercial, que cedeu a versão textual de cerca de 500 obras de literatura portuguesa, no início dos anos 2000. Todas as obras foram tratadas por programas desenvolvidos por Paulo Rocha, no âmbito da Linguateca, que também modelou a forma de codificar os seus metadados, baseado numa tripartição entre Prosa, Teatro e Poesia. Ao longo do tempo foram sido feitas algumas melhorias esporádicas, mas apenas neste momento se está a incluir a informação sobre género literário, e a considerar adicionar escola literária a cada uma das obras incluídas.

3. O corpo Literateca e a própria Literateca

O corpo Literateca não é mais do que a união de todos os textos literários que a Linguateca tem acessíveis. Foi criado para facilitar o estudo de textos literários e permitir o uso de anotações relacionadas a gêneros literários, assim como resolver o problema da repetição de textos literários em corpos diferentes. Com a indicação clara da origem, e para sublinhar a autorização dada, foram acrescentados os textos do corpo Tycho Brahe (Galves e Faria, 2010), do Colonia (Zampieri e Becker, 2014), do OBRAS e do Vercial, assim como os excertos originários de corpos paralelos que incluam o português.

Todo este acervo, anotado sintaticamente com o PALAVRAS (Bick, 2000) e semanticamente com vários campos descritos noutras publicações, e com as indicações de autor, obra, data e género, pode ser consultado e usado para estudos de literatura em português. 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 (ou análogo) na Gramateca (Santos, 2014).

De momento, a Literateca contém 462 obras distintas (de 157 autores diferentes), incluindo crónicas, cartas (como a do Descobrimento do Brasil), sermões e mesmo atas de congregações (coligidas pelo projeto Tycho Brahe); há atualmente cerca de 200 obras em prosa publicadas no período 1800-1920 .

2. A escolha do subconjunto para o COST

A escolha de um conjunto de obras para pertencer a um corpo literário numa das línguas europeias, neste caso o português, não deve ser sinónimo da limitação de tal escolha ao espaço da Europa, sobretudo quando a língua em questão também veiculou uma literatura "nacional" noutros continentes. Daí um primeiro critério ser o da não restrição das obras às que foram produzidas em Portugal, alargando-as ao Brasil e, se viável, outros países.

Um segundo critério resulta da necessidade de definir um horizonte temporal em que as obras já sejam de domínio público. Daí a escolha de obras anteriores a 1920 e o recuo até 1800.

Exceto estes dois critérios, não há certeza quanto aos outros a definir. Deixam-se de seguida algumas hipóteses.

2.1. Outros critérios de escolha

A primeira questão que se põe é o que é realmente um romance, ou melhor, se há critérios fora de cada tradição literária que permitem identificar um género em literaturas diferentes. Todos sabemos que enquanto em português existem tradicionalmente romances, novelas e contos, em inglês trabalha-se apenas com *novels* e *short stories*.

Outra questão é imediatamente que tipo de romance vamos incluir ou excluir: serão preferidas as obras consideradas obras-primas pela academia, ou as obras com mais impacto (leitores, edições) no seu tempo?

2.2. Critérios motivados por considerações estatísticas

Considerando um corpo como uma amostra que deve ser variada e representativa, é preciso compreender que a escolha de uma amostra tem critérios independentes que podem estar em conflito entre si. Por um lado seria natural escolher um leque de autores tão vasto quanto possível

(e daí incluir uma obra de cada), mas por outro perder-se-ia a possibilidade de distinguir entre autor e obra; por outro lado, é possível fazer uma amostragem relativamente uniforme por década – e nesse caso perder a oportunidade de incluir várias obras primas, todas elas publicadas na mesma época; por razões extra-literárias poderíamos decidir por uma divisão arbitrária entre obras de Portugal e do Brasil, ou entre obras de escritoras femininas e de autores masculinos. Todas essas opções corresponderiam a diferentes escolhas e diferentes conjuntos de 100 obras.

A nossa sugestão inicial propõe a seguinte metodologia, chamada de estratificação parcial: seleção de alguns autores absolutamente canónicos, e escolha de duas obras de cada (se escreveram mais do que uma). Seleção aleatória entre todos os outros, colocando como requisito que mantenham a proporção estatística da proporção entre homens e mulheres, e eventualmente entre décadas, correntes literárias, "estatuto" (em inglês, *high-brow* e *low-brow*) e proveniência geográfica.

3. Exemplos de estudos realizáveis com este método e material

Numa conferência de humanidades digitais, mais importante do que apresentar recursos é explicar como recorrer a eles, e é o que faremos no artigo final. Aqui listamos brevemente alguns estudos possíveis tanto na língua portuguesa, apenas, como comparando influências e "tropos".

3.1. O estudo de obras literárias em português

Embora o título deste artigo contenha o termo "leitura distante" batizado por Moretti (2000), é importante sublinhar que também a leitura próxima ou aturada (*close reading*) sai facilitada. Por isso, pretende-se começar 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 dos próprios personagens conforme apresentados pelo narrador (a passividade de personagens que *concordam*, *assentem* e *justificam-se*, em oposição àqueles que *interrompem, propõem, acentuam, teimam, insistem e exclamam*, por exemplo). Ainda quanto à caracterização de personagens, para além da correta anotação de nomes próprios e formas de tratamento, indicamos que o fato de já contarmos com a análise sintática fornecida pelo PALAVRAS, bem como a existência, no contexto da Gramateca, da anotação do corpo humano (veja-se Freitas et al., 2015), agiliza a busca por predicadores humanos, como iremos mostrar.

3.2. O estudo das relações entre obras literárias em português

Outra técnica – e domínio de estudo – associada naturalmente à leitura distante é a noção de "topic modeling", veja-se por exemplo Jockers (2013) ou Schöch (2017), em que é possível associar automaticamente tópicos (conjunto de palavras relacionadas) a partes de obras literárias, permitindo uma comparação baseada nesses "tópicos" automáticos entre obras e entre autores.

Outro método mais associado à linguística com corpos seria o de 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).

Neste caso poderia ser interessante tomar em consideração não só os textos literários em si, mas também os paratextos (dedicatórias, citações, comentários à primeira e a novas edições), que se encontram neste momento apenas acessíveis nas obras provindas dos corpos Vercial e OBras.

3.3. O estudo das relações entre obras literárias em geral

Muito possivelmente uma primeira estratégia será usar uma forma de "traduzir" ou "alinhar" os tais tópicos automáticos em línguas diferentes, e tentar fazer um mapeamento entre várias literaturas. Mas evidentemente que outros rumos também poderão ser tomados em paralelo, tomando em conta o estudo e conhecimentos dos próprios autores e das influências expressas e possíveis (por razões espaciotemporais), assim como a marcação de algumas questões (por exemplo a maternidade, ou a religião) especialmente.

Referências

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, 2016, p. 4410-4416.

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, pp. 641-680, 2015.

GALVES, Charlotte; FARIA, Pablo. Tycho Brahe Parsed Corpus of Historical Portuguese. 2010. Disponível em: <http://www.tycho.iel.unicamp.br/~tycho/corpus/en/index.html>. 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.

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, 2014, p. 214-219.

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, 2014, p. 219-236.

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, 2008, pp.41-66.

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, 2013, p. 77-84.