FUTURO RISONHO: PROLEGÓMENOS PARA UMA COLABORAÇÃO ENTRE A LINGUATECA E O NUPILL

Diana Santos[*]

Foi no *Primeiro Encontro sobre a Leitura Distante em Português*, em outubro de 2019 em Oslo (Santos et al 2020b), que estes dois projetos de pesquisa (ou instituições) se encontraram e decidiram colaborar numa atividade que pudesse significativamente ultrapassar aquilo que ambos faziam cada um por seu lado, para o estudo e glória da literatura em português. Os dois projetos são provenientes – no sentido de ambições para o futuro, não no sentido administrativo – do NuPILL (Núcleo de Pesquisas em Informática, Literatura e Linguística (NuPILL), e da Linguateca, e a atividade de que estamos a falar, e que se encontra ainda no futuro dos deuses, é a criação de um acervo digitalizado de obras literárias lusófonas que conjugue harmoniosamente a informação sobre as mesmas e o acesso ao seu texto completo para um leitor humano, com a possibilidade de fazer estudos baseados em corpos tanto em cada obra como no conjunto de todas as obras. Por outras palavras, que permita leitura distante e próxima do património literário em língua portuguesa, através de uma colaboração entre essas duas instituições com longos anos de experiência com a língua e a literatura.

[*] Universidade de Oslo, Noruega. E-mail: d.s.m.santos@ilos.uio.no. ORCID: http://orcid.org/0000-0002-3108-7706.

Visto que este texto é escrito do lado da Linguateca, para uma obra que celebra o NuPILL e cuja audiência é primordialmente de amigos e membros do NuPILL, tenho como principal objetivo explicar o que a Linguateca tem e já fez em relação à literatura lusófona, assim como a minha visão – unilateral, portanto – do que poderíamos fazer em conjunto. Mas esse contexto pode também permitir uma clarificação e tomada de consciência de vieses e de lacunas que poderão ser colmatadas através de um diálogo que pretende abrir um caminho. Tentarei portanto evitar ao máximo o nós por oposição ao vós, no sentido de dar a falsa impressão de que cada um traz uma coisa diferente, e enfatizar a ideia de que juntos podemos fazer sim algo novo, a que chamarei "literatura computacional.

A Linguateca

A Linguateca começou no ano 2000, após dois anos de um projeto chamado "Processamento computacional do português", cujo objetivo era planear/projetar aquilo que deveriamos fazer para melhorar a presença da língua portuguesa na informática, desenvolvendo ferramentas e criando recursos que pudessem melhorar a pesquisa e o desenvolvimento em áreas que tivessem a ver com o conhecimento e a comunicação. A Linguateca, financiada pelo Ministério da Ciência e da Tecnologia em Portugal, foi assim formada para tornar mais fácil progredir em métodos e programas que compreendessem (ou pelo menos manipulassem) o português, concebida como um serviço à comunidade acadêmica e empresarial, produzindo e disponibilizando recursos, e desenvolvendo avaliações conjuntas.

Avaliações conjuntas são iniciativas que, junto a uma comunidade de interessados numa tecnologia e numa tarefa específica, permitem avaliar consensualmente uma dada atividade ou sub-área tecnológica, e têm uma já longa história para a língua inglesa, organizadas pela DARPA ou pelo NIST, organismos de pesquisa ou de estandardização ligados ao governo americano. De 2002 a 2012 a Linguateca organizou quatro dessas avaliações para o português, tendo uma delas sido no âmbito do CLEF, um projeto europeu de recolha cruzada de informação entre várias línguas. Para uma resenha histórica dessa atividade, veja-se Santos (2021b).

Além dessa atividade de avaliação, a Linguateca também criou vários recursos públicos para fomentar a pesquisa e o desenvolvimento em várias áreas: a linguística com corpos (o projeto AC/DC, Santos (2014)), os estudos contrastivos (os corpos COMPARA (Frankenberg-Garcia & Santos, 2002), CorTrad (Teixeira et al., 2012) e PANTERA (Santos, 2019)), a análise sintática (Floresta Sintá(c)tica, Afonso et al. 2002), os estudos léxico-semânticos (PAPEL, Gonçalo Oliveira et al. (2010)), e a resposta automática a perguntas (o sistema Esfinge, Costa (2016)), entre outros.

Questões conjunturais levaram a que a Linguateca tivesse de se reestruturar de forma muito mais modesta em 2010, deixando de receber financiamento e passando a funcionar como uma rede de interessados que deveriam obter o financiamento para a pesquisa nos seus locais de trabalho. Tal não impediu, contudo, que os recursos continuassem públicos (graças ao acolhimento informático da FCCN, Fundação para o Cálculo Científico Nacional) e que projetos menores continuassem a ser desenvolvidos, muitas vezes graças à ajuda de voluntários.

A Linguateca foi, contudo, exclusivamente um projecto de linguística e de informática, até que o seu envolvimento nas Humanidades Digitais na Faculdade de Letras da Universidade de Oslo a arrebatou também para a área da literatura computacional, através da ação COST "Distant Reading for European Literary History". Este projeto, iniciado em 2017 e prestes a terminar, tem três vertentes: a da linguística com corpos, a das ferramentas computacionais, e a dos estudos literários. O seu desenrolar levou a que fosse criado um corpo de literatura portuguesa para leitura distante segundo normas comuns para todas as literaturas presentes no projeto (veja-se Schöch et al. 2020), assim como potenciou o desenvolvimento e afinação de ferramentas para algumas tarefas necessárias aos estudos literários, como será descrito mais adiante. Além disso, abriu-nos os olhos para as limitações e faltas que existem em relação à digitalização e ao acesso ao acervo literário em língua portuguesa, que é um problema que desejariamos colmatar em parceria com o NuPILL.

Mas antes de nos debruçarmos sobre isso, vamos descrever o pouco o que temos já feito na área da literatura computacional em português, como "cartão de visita" para uma cooperação. Começamos por descrever a Literateca.

A Literateca

Se dermos uma atenção especial aos textos literários incluídos nos variados corpos a que a Linguateca dá acesso, podemos identificar um conjunto de dados suficientemente numeroso para dar origem a uma novo entidade, a Literateca, que reúne textos literários como um corpo linguístico – mas que pode também ser objeto de estudos literários. O que significa

que os textos (além das características linguísticas que possuem como textos) podem ser anotados, e vistos, com os olhos de um estudioso da literatura. Sem deixar de lembrar que os estudos literários têm objetivos e métodos muito diferentes dos linguísticos, e que portanto não pode ser "só" substituir a anotação, e esperar que todo o resto se adapte.

Donde é que vêm esses textos literários? Muito rapidamente, vêm de várias fontes/projetos. A primeira leva foi o projeto Vercial, de qual tivemos autorização para dar acesso aos seus textos através do projeto AC/DC já em 1999. O projeto Vercial[1] digitalizou e reviu/atualizou a grafia[2] de uma grande quantidade de textos canónicos portugueses, lírica, prosa e drama, desde as crónicas de Fernão Lopes em 1385 até às memórias de Raul Brandão em 1933.[3]

O projeto OBras (Obras Brasileiras) surgiu como uma resposta ao Vercial, visto que muitos utilizadores do AC/DC pediam também textos literários brasileiros e não só portugueses. O corpo Obras[4] surgiu de uma colaboração entre Anya Campos, Cláudia Freitas da PUC-Rio e eu da Universidade de Oslo, tentando obter alunos e/ou bolsas para incorporar obras brasileiras no domínio público no AC/DC. A partir de 2019, Emanoel César Pires de Assis, da Universidade Estadual do Maranhão, juntou-se ao grupo, contribuindo principalmente com obras maranhenses, e em 2020 foi a vez de Marcia Langfeldt, com obras amazônicas. Uma das grandes vantagens do OBras é que, além de ser acessível através do

[1] Ver http://alfarrabio.di.uminho.pt/vercial/
[2] Infelizmente, para a grafia anterior ao Acordo Ortográfico atual.
[3] A composição do corpo Vercial encontra-se em https://www.linguateca.pt/acesso/lista_autores_vercial.html
[4] A composição do corpo OBras encontra-se em https://www.linguateca.pt/acesso/lista_autores_obras.html

AC/DC, também pode ser levantado por completo (Santos et al. 2018).

O projeto NOBRE (Novas OBRas publicadas na Europa) é por sua vez uma contrapartida ao OBras envolvendo obras portuguesas no domínio público que ainda não se encontravam no corpo do Vercial. Era preciso termos obras não canônicas para a coleção COST, e como as não podíamos incluir nem no Vercial nem no Obras, tivemos de criar mais um corpo. As obras nele contidas [5] têm a particularidade de apresentarem, em sua maioria, uma grafia antiga, ao contrário dos dois corpos anteriores (embora o OBras também contenha alguns livros na grafia antiga, a esmagadora maioria das obras tem a grafia atualizada).

Damos além disso acesso no AC/DC a vários outros corpos criados por projetos distintos da Linguateca, e dois desses projetos também lidam com textos literários, nomeadamente o *Tycho Brahe*[6], e o Colonia[7], ambos reunindo obras portuguesas e brasileiras publicadas ao longo de vários séculos.

Como não poderia deixar de ser, projetos diferentes compilados por pessoas diferentes e com objetivos diferentes muitas vezes vão escolher os mesmos textos. Além disso, não seria muito prático para um utilizador ter de consultar separadamente cinco corpos diferentes, nem seria possível agregar os resultados. Por isso, decidimos criar um novo corpo, a Literateca, que reúne todos os textos literários acessíveis da Linguateca num único local, e apenas uma vez. (Cabe indicar

[5] A composição do corpo NOBRE encontra-se em https://www.linguateca.pt/acesso/lista_autores_nobre.html
[6] Ver http://www.tycho.iel.unicamp.br/corpus/index.html
[7] Ver http://corporavm.uni-koeln.de/colonia/

aqui que não é pacífico o que é um texto literário, e que nós deixamos essa escolha aos compiladores dos corpos iniciais. Apenas no caso do *Tycho Brahe*, que reúne uma grande variedade de géneros, nós retiramos os textos classificados como dissertativos, epistolares, e atas de sociedades.)

Além disso, juntamos à Literateca os excertos de obras constantes nos corpos paralelos a que damos acesso, nomeadamente os do PANTERA [8]. Esta opção pode ser discutível pelo fato de irem ombrear com obras completas, mas é sempre possível nas buscas escolher o material sobre o qual procuramos (e daí retirar esses excertos).

Para dar uma ideia da constituição e do tamanho da Literateca em junho de 2021, veja-se a seguinte tabela, em que, além do total, também separamos o material português e brasileiro (mas refira-se que existem autores moçambicanos, angolanos e caboverdianos no PANTERA, por isso o total é um pouco maior do que a soma das duas literaturas):

Tabela 1 - Quantidades na Literateca, versão 6.21, 20 de junho de 2021

	Total	PT	BR
Palavras	29.049.724	21.406.087	7.631.910
Unidades[9]	35.996.294	26.529.665	9.447.509
Obras	857	537	315
Autores	244	185	58

[8] Ver https://www.linguateca.pt/PANTERA

[9] Unidades é tudo o que constitui o texto: além de palavras, sinais de pontuação, e números. Em inglês usa-se *tokens*.

Estamos conscientes de que há discordância em relação à classificação de vários textos luso-brasileiros, como por exemplo os sermões do Padre António Vieira, a carta de Pero Vaz de Caminha sobre o achamento do Brasil, ou o romance Aventuras de Diófanes. Não querendo entrar no debate, indico apenas que todos estes textos, por se encontrarem tratados pelo projeto Vercial, ou terem sido considerados pelo corpo *Tycho Brahe* como portugueses, assim foram marcados na Literateca. (Mas, mais uma vez, é perfeitamente exequível "desfazer" esta marcação através de buscas mais elaboradas.)

Também temos de chamar a atenção para o conceito fluido de obra, empregue na tabela acima: inclui, conforme o critério dos compiladores dos respetivos corpos ou dos curadores de coleções e textos, casos de livros de contos, livros de poesia, mas também contos separados, e mesmo crónicas ou poemas separados.[10] Para uma descrição exaustiva, veja-se a página do conteúdo da Literateca.[11]

É preciso de qualquer maneira indicar que todos os corpos – e correspondentes coleções de obras subjacentes – mas sobretudo o OBras e o NOBRE, que são da responsabilidade direta da Linguateca – continuam a ser alimentados e a aumentar a um ritmo equilibrado, e o nosso desejo é que essa alimentação passe, ou melhor seja dirigida, pelo NuPILL – que foi a instituição que tratou da digitalização e atualização do acervo de Machado de Assis já mencionado. Assim, é preciso não esquecer que a Literateca é algo dinâmico que continua em

[10] Isto acontece sobretudo no OBras, por causa da digitalização da obra completa de Machado de Assis ter feito essa escolha (separar as crónicas e os contos como obras distintas. Mas também a obra 14 de julho na roça, de Raul Pompéia, nos levantou dúvidas sobre se se deveria referir ao conto homónimo ou a todo o livro, quando tentávamos repetir um trabalho publicado por outros pesquisadores, em Santos (2020a).
[11] Ver https://www.linguateca.pt/acesso/lista_autores_literateca.html

expansão, e que um dos meus objetivos neste capítulo é convencer o NuPILL a ajudá-los a enriquecê-la.

Primeiros estudos na Literateca

Na Literateca fizemos já alguns estudos e aplicamos algumas técnicas de literatura computacional, a saber: será que a identificação de várias características sintáticas e semânticas (linguísticas, portanto) permite a identificação da escola literária? Em Santos et al. (2020a) – com um subconjunto das obras que agora constituem a Literateca, visto que esta se encontra em permanente expansão – classificamos 192 obras literárias do género romance ou novela, publicadas no período 1840 a 1919, nas diversas escolas literárias presentes nesse período (nomeadamente romantismo, realismo, naturalismo, decadentismo, regionalismo, indianismo, pós-naturalismo, expressionismo e modernismo) e também em alguns casos em romance histórico ou de ficção científica. É preciso salientar que muitas obras receberam mais de uma classificação, quer por corresponderem a períodos de transição, quer por haver real discordância entre os teóricos. A grande maioria pertencia, contudo, à escola romântica. Para podermos aplicar métodos estatísticos, atribuímos-lhes depois uma escola "simplificada" em que cada obra apenas podia receber uma classificação, ficando com 131 obras pertencentes ao romantismo, 42 ao realismo e 12 ao naturalismo, ou, juntando estas duas últimas, 54 obras de cariz realista-naturalista.

Usando 128 características sintático-semânticas que calculamos para cada obra, identificamos, como mais discriminantes para localizar as obras em diferentes regiões do plano, imperfeitamente separando as várias escolas, o uso de

pronomes interrogativos, de passivas, de orações relativas, de nomes próprios, de completivas, do conjuntivo, de travessões, de referência à medicina e à emoção humildade.

Podemos encontrar modelos de tópicos que reflitam diferentes escolas literárias? Usando a classificação simplificada descrita no ponto anterior, e aplicando a análise de tópicos (veja-se por exemplo Jockers & Mimno, 2013) a ambos os conjuntos de romances (românticos e realistas-naturalistas), obtivemos alguns tópicos que nos pareceram típicos de cada escola literária, também em Santos et al. (2020a).

A identificação de personagens através do uso de nomes próprios, a criação de redes de personagens e a visualização da sua presença ao longo da obra permitem caracterizar de alguma forma um romance? Marcamos, até agora em 13 romances, os nomes próprios com a personagem a que se referiam, e desenhamos a presença das personagens ao longo da obra, assim como as relações entre as diversas personagens, quantificadas pelo número de vezes que aparecem em conjunto em passagens da obra, criando assim redes de personagens[12]. Além disso, classificamos todos os outros nomes de pessoas em pessoas históricas, pessoas mitológicas ou ficcionais, e entidades religiosas, de forma a podermos comparar várias obras e autores. Uma primeira descrição dos resultados com ênfase na invocação de Deus e do diabo, pode encontrar-se em Santos & Freitas (2019).

Dado que a identificação automática de nomes de pessoas é uma subtarefa do reconhecimento de entidades mencionadas, e as personagens de um romance tendem a ser as pessoas cujo nome é mais usado, não parece impossível

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

automatizar a identificação automática de personagens em grandes acervos, permitindo estudos quantitativos sobre, por exemplo, o gênero ou a ocupação das personagens principais em milhões de obras (veja-se Vala et al. 2015). O desenvolvimento e avaliação desse tipo de sistemas é algo que queremos desafiar o NuPILL a organizar conosco (ver adiante a seção "Pistas para cooperação").

A identificação dos locais na literatura permite caracterizar diferentes obras? Uma outra constante de uma obra literária em prosa é a localização ou localizações do seu enredo, tanto a nível real como ficcional. Ou seja, uma ação passa-se sempre em qualquer lado, seja dentro dos contornos de uma casa ou de uma aldeia, seja viajando por uma região ou país inteiro. Identificar as localizações (com nome próprio ou apenas comum) e comprender a área coberta, as viagens, os pontos de interesse ou de ação, ajuda a caracterizar e a compreender uma obra – e, se estivermos interessados em leitura distante, um conjunto de obras. Vários autores, usando reconhecimento de entidades geográficas, já tentaram responder a perguntas caras aos estudos literários, como Elson et al. (2010) ou Cooper & Gregory (2011).

Em colaboração com o projeto do *Atlas das Paisagens Literárias de Portugal Continental*[13], a Linguateca participou em 2019-2020 no projeto BILLIG, cujo objetivo era explorar o triplo "sistemas de informação geográfica, linguística e literatura". Nesse âmbito estamos a classificar os nomes próprios que correspondem a lugares no AC/DC, e a atribuir-lhes (quando não são ficção) coordenadas geográficas, o que já permite caracterizar as obras em termos de tipos de locais e

[13] Ver http://litescape.ielt.fcsh.unl.pt/

regiões a que fazem referência.[14] Temos neste momento mais de 6000 lugares distintos geo-referenciados, correspondendo, na Literateca, a cerca de 100 mil casos. Além disso, desenvolvemos um protótipo de visualização de mapas com base em procuras no AC/DC, ver Santos & Alves (2021).

A menção de profissões, obras de arte, e demônimos ou gentílicos, também pode caracterizar uma literatura? Mas não são só pessoas/personagens e locais que nos podem dar pistas, de longe, para classificar uma obra. O conjunto de descrições socio-profissionais, e de gentílicos, é igualmente significativo para representar o ambiente social em que uma dada obra se desenrola. E a relação com outros textos e outras obras de arte é importante para identificar a influência, explícita ou nem isso, que os autores sofreram. Alguns comentários sobre essa questão, assim como com a possibilidade de o fazer com um sistema de reconhecimento de entidades mencionadas (REM), o PALAVRAS-NER, são discutidos em Santos, Bick & Wlodek (2000).

O estudo do léxico associado à saúde é relevante para uma nova visão do texto literário? Uma característica interessante das literaturas portuguesa e brasileira é a sua íntima relação com os médicos, como é observado por exemplo em Santos (2019) e em Langfeldt (2021). Por esse motivo, a identificação e o estudo das atitudes e da presença da doença na literatura podem ser uma excelente ferramenta para a história da medicina. Por outro lado, a forma como episódios de doença e morte são referidos e integrados num texto literário pode ser uma característica autoral.

[14] Ver https://www.linguateca.pt/Gramateca/Viagem.html

Existem diferenças na descrição dos gêneros? Um dos assuntos mais abordados nestes últimos tempos é a perspetiva de género numa obra literária. Como os diferentes gêneros são descritos, classificados, conceitualizados, e empregues na trama passou a estar sob a lupa dos pesquisadores, assim como o resgate de obras escritas por mulheres que tenham caído no esquecimento. Um primeiro trabalho, de muito valor, usando exatamente o OBras foi a tese de mestrado da Flávia Silva (2021), em que a caracterização de personagens femininas e masculinas foi estudada e comparada. Abordagens análogas podem ser seguidas em relação a fenômenos como etnicidade ou regionalidade. Ou seja, vejam-se as seguintes perguntas pertinentes: como são caracterizadas na literatura pessoas de diferentes etnias, de diferentes regiões ou nacionalidades, de diferentes religiões, ou mesmo de diferentes orientações sexuais?

Obras diferentes referem emoções diferentes? Certamente que uma das consequências indiscutíveis de uma obra literária é a de suscitar emoções no leitor. Mas até que ponto a própria menção das emoções (sentidas pelas personagens da obra, ou pelo narrador) permite classificar e compreender um livro ou um dado autor? Um trabalho preliminar é a anotação dos textos com referências a emoção, algo que está sendo feito no AC/DC, veja-se Santos, Mota & Simões (2020). Um mapeamento de sentimentos positivos ou negativos ao longo de uma obra, feito entre outros por Archer & Jockers (2016), poderia ser mais uma.

Existem outros campos semânticos que podem iluminar a literatura? Tanto o vestuário como a famíliar foram selecionados para anotação no AC/DC por razões exteriores aos estudos literários, mas podem ser usados (como o fizemos em

Santos et al. 2020) para caracterizar textos. Em Santos (2021) faço umas observações preliminares sobre a roupa na literatura. E a existência ou não de laços familiares descritos nas obras é também uma janela sobre a sociedade e visão de família que nos é oferecida.

Da mesma forma, trabalho sobre o relato (direto, indireto ou misto) e atribuição do falante, descrito em Freitas et al. (2018), e anotado em todos os corpos do AC/DC, e portanto também na Literateca, pode ser valioso para caracterizar o discurso literário em grandes traços. Mencione-se de passagem que, no âmbito da já mencionada ação COST, estamos de momento a comparar as várias coleções em termos da preseça de vida interior, de acordo com uma ideia de Tamara Radak, e sugestões de Fotis Iannidis e Pieter François.

Seja como for, na Linguateca enveredamos por este ramo, de tentar aplicar as técnicas e as informações linguístico-computacionais ao estudo da literatura lusófona há bem pouco tempo, e precisamos da ajuda e orientação do NuPILL, que está nestas lides há 25 anos, para que o esforço seja útil para estudos literários na literatura lusófona. Passo portanto a sugerir alguns esforços colaborativos concretos.

Pistas para cooperação

Um primeiro passo é obter uma "linha de montagem", em que o NuPILL, primordialmente, mas talvez ajudado pela Linguateca, execute as seguintes tarefas "preliminares", muitas das quais já executa no âmbito da Biblioteca Digital de Literatura de Países Lusófonos[15]: a) escolha das obras; b)

[15] Ver https://www.literaturabrasileira.ufsc.br/

digitalização das mesmas; c) revisão; d) documentação (pública) dessa revisão e da sua classificação; e) codificação estrutural (capítulos, epígrafes, dedicatórias, prefácios, posfácios, etc.).

Seria imaginável – diria mesmo sensato – desenvolver sistemas informáticos que agilizassem significativamente as fases da digitalização e da revisão humana, que deveriam ser financiados pelos grandes atores da digitalização, seja a Google, sejam as Bibliotecas Nacionais dos países lusófonos, como foi feito na Suécia (Borin et al., 2016). Não começando do nada, mas sim usando já como material de treino as centenas de obras passadas por reconhecimento ótico de carateres (ROC) e a sua correção humana.

Na documentação e categorização do resultado seria também muito interessante desenvolver uma ontologia literária, algo em que, se não me engano, o NuPILL já trabalha[16].

O segundo passo é identificar problemas *literários* que possam ser resolvidos, ou ajudados, por uma aproximação de leitura distante, ou por uma aproximação de micro-leitura. Por exemplo, e retomando muitos dos temas já mencionados em secções anteriores: a) a identificação das personagens de uma obra; b) a identificação do discurso direto, indireto e indireto livre; c) a identificação do tipo de narrador; d) a identificação dos locais em que a ação se passa; e) a identificação do perfil temporal de uma obra; f) a identificação de cenas; g) a identificação do tipo de final (final feliz? Qual o destino de cada personagem?); h) a identificação de descrições de personagens e suas semelhanças (Bamman et al., 2014); i) a identificação de tópicos tratados; j) a identificação de ideologias; k) a deteção de

[16] Veja-se http://dados.literaturabrasileira.ufsc.br/

mudanças no tempo, etc. etc. Este tipo de propriedades, depois de (semi)automaticamente detetadas, poderiam ser adicionadas à base de dados sobre as obras, de forma a acumular um conhecimento vasto sobre a biblioteca (aqui considerada como o conjunto de livros elencado e "lidos" de forma distante). Para isso, seria preciso fazer várias "tempestades de ideias" (do inglês *brainstorming*) e usar inicialmente anotação humana, talvez recorrendo ao DLnotes, uma ferramenta de anotação literária criada pelo NuPILL (Mittmann et al. 2013). Desse processo acredito que surgiriam aulas e módulos de ensino da literatura com um viés muito interessante.

Uma outra pista que poderíamos seguir, e que já não se refere à caracterização das obras como um todo, mas sim a passagens ou a frases, era a articulação com a procura em contexto de questões mais específicas de vocabulário ou sintaxe, como cores, relações familiares, corpo humano, roupa ou emoções, ou adjetivação, orações relativas, passivas, etc., dentro de uma obra específica, ou de conjuntos de obras, algo para o qual o AC/DC está vocacionado.

Depois de ter uma ideia mais concreta do que se poderia com vantagem procurar e marcar nos textos, assim como proceder a uma definição de prioridades, poderíamos tentar interessar vários grupos (além da Linguateca e do NuPILL) no desenvolvimento de programas que detetassem essas características. A melhor forma para o fazer é, na minha opinião, organizar uma avaliação conjunta. E por isso proponho desde já avançarmos para a primeira: a deteção de personagens, e sua caracterização simples (a definir melhor com os participantes, mas que poderia envolver: o seu gênero, as suas relações familiares com as outras personagens, o seu estatuto sócio-profissonal, as variadas formas por que são

mencionadas). Para levar este acontecimento a bom termo seria preciso que a organização, constituída pela Linguateca e pelo NuPILL, selecionasse um conjunto de obras de vários tipos (canônicas, não canônicas, com grafias variadas, etc.) que fossem usadas para avaliação e que garantisse ter a resposta certa para um subconjunto (a coleção dourada), subconjunto esse desconhecido dos participantes. Proponho que esse trabalho seja feito nos meses de setembro a novembro do presente ano.

À laia de conclusão

Neste texto tentei esboçar algumas formas de cooperação entre o NuPILL e a Linguateca, que me parecem úteis para a área da literatura computacional em língua portuguesa, e que derivam de algum trabalho preliminar ou experiência existente na Linguateca, mais especificamente na Literateca. É, pois, apenas uma metade daquilo que pode ser feito ou proposto, a nossa metade.

Muito brevemente, seguimos o modelo de colaboração entre "estados soberanos", em que nenhum dos projetos contrata o outro, mas em que ambos têm vantagens na cooperação e concordam que não podem (ou devem) fazer o trabalho sozinhos: além das questões técnicas que irão ser desenvolvidas em conjunto, cada projeto terá responsabilidade principal em algumas áreas de atuação. Por exemplo, a escolha da literatura e dos problemas literários a atacar seria do NuPiLL, enquanto o processamento corpóreo-linguístico seria da Linguateca. O que não quer dizer que o outro parceiro não pudesse contribuir significativamente. Proponho aqui assim duas coisas distintas, idealmente a fazer em paralelo.

A primeira é a definição de uma "linha de montagem" para entrelaçar a Biblioteca Digital de Literaturas em Língua Portuguesa do NuPiLL com a Literateca: como receber as obras escolhidas e identificadas em termos de metadados pelo NuPILL na Literateca, e como produzir nesta dados quantitativos, e outras informações que enriqueçam a base de dados da biblioteca digital. Além disso, como apontar claramente entre os dois projetos (ambos acessíveis na rede), de forma a ser fácil para um usuário mudar de plataforma conforme os seus interesses de pesquisa.

A segunda é a organização da primeira avaliação conjunta de ferramentas de leitura distante, para a tarefa de deteção automática de personagens e algumas suas características. E fica também a sugestão de tentar um projeto mais tecnológico em conjunto com atores de digitalização globais, para melhorar o reconhecimento automático de caracteres nas várias grafias do português ao longo dos (pelo menos) dois últimos séculos, e para proporcionar um ambiente de revisão humana que expedite e simplifique o processo.

E voto para que o futuro da nossa colaboração seja risonho, dando razão ao título deste texto híbrido entre uma proposta de colaboração, e uma homenagem a um parceiro muito desejado.

REFERÊNCIAS

Afonso, Susana, Eckhard Bick, Renato Haber & Diana Santos. "Floresta sintá(c)tica: a treebank for Portuguese". *In*: Manuel González Rodríguez & Carmen Paz Suárez

Araujo (eds.), *Proceedings of LREC 2002, the Third International Conference on Language Resources and Evaluation* (Las Palmas de Gran Canaria, Spain, 29-31 May 2002), ELRA, 2002, pp. 1698-1703. http://www.lrec-conf.org/proceedings/lrec2002/pdf/1.pdf

Archer, Jodie & Matthew L. Jockers. *The Bestseller Code: Anatomy of the Blockbuster Novel*. Sr. Martins's Press, 2016.

Ardanuy, Mariona Coll & Caroline Sporleder. "Structure-based Clustering of Novels". *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLfL) @ EACL 2014, Gothenburg, Sweden, April 27, 2014*, pp. 31–39, ACL, 2014. http://www.aclweb.org/anthology/W14-0905

Bamman, David, Ted Underwood & Noah A. Smith. "A Bayesian Mixed Effects Model of Literary Character". *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, Maryland, USA, June 23-25, 2014, pp. 370–379. https://www.aclweb.org/anthology/P14-1035.pdf

Borin, Lars, Gerlof Bouma & Dana Dannélls. "Free cloud service for OCR / En fri molntjänst för OCR: Project report". Research Reports from the Department of Swedish, University of Göteborg, GU-ISS-2016-01, 2016. https://gupea.ub.gu.se/bitstream/2077/42228/1/gupea_2077_42228_1.pdf

Cooper, David & Ian N. Gregory. "Mapping the English Lake District: A literary GIS". *Transactions of the Institute of British Geographers*, **36**, 2011, pp. 89-108.

Costa, Luís. "Esfinge - A Question Answering System in the Web using the Web". In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL*

2006) (Trento, Itália, 3-7 de Abril de 2006), pp. 127-130. https://www.aclweb.org/anthology/E06-2011.pdf

de Does, Jesse, Katrien Depuydt, Karina VanDalen-Oskam & Maarten Marx. "Namescape: named entity recognition from a literary perspective". In Jan Odijk & A. Van Hessen (eds.), *CLARIN in the Low Countries*. Ubiquity Press, 2017, pp. 361–370.

Elson, David K., Nicholas Dames & Kathleen R. McKeown. "Extracting Social Networks from Literary Fiction". In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, 11-16 July 2010, pp. 138–147. https://www.aclweb.org/anthology/P10-1015.pdf

Frankenberg-Garcia, Ana & Diana Santos. "COMPARA, um corpus português-inglês na Web". *Cadernos de Tradução* **9** (2002), Universidade Federal de Santa Catarina, Brasil, 2003, pp. 61-79. https://www.linguateca.pt/Diana/download/FrankenbergGarciaSantosCadTrad.pdf

Freitas, Cláudia, Bianca Freitas & Diana Santos. "QUEMDISSE?: Reported speech in Portuguese". In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.). *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 4410-4416. http://www.lrec-conf.org/proceedings/lrec2016/pdf/417_Paper.pdf

Gonçalo Oliveira, Hugo, Diana Santos & Paulo Gomes. "Extracção de relações semânticas entre palavras a partir de um dicionário: o PAPEL e sua avaliação". *Linguamática* **2**, 1, Abril 2010, pp. 77-94. https://linguamatica.com/index.php/linguamatica/article/view/39/61

Jockers, Matthew L. & David Mimno. "Significant themes in 19th-century literature", *Poetics* **41** (6), 2013, pp. 750-769.

Langfeldt, Marcia Caetano. "Entre médicos e charlatães: A ascensão da medicina na formação da literatura brasileira". Apresentação no *III Encontro Nacional de Estudos Linguísticos e Literários (ENAELL), I Encontro Internacional de Pesquisas em Letras (ENIPEL)* (UEMA, 25-27 de maio de 2021). https://www.linguateca.pt/documentos/Langfeldt2021.pdf https://www.youtube.com/watch?v=XFEVaZC_ibU

Lee, John & Chak Yan Yeung. 2012. Extracting networks of people and places from lite-rary texts. Em *Pacific Asia Conference on Language, Information and Computation*, pp. 209–218. https://www.aclweb.org/anthology/Y12-1022.pdf

Mittmann, Adiel, Roberto Willrich & Renato Fileto. "DLNotes2: Ferramenta de anotações estruturadas e semânticas voltada ao ensino da literatura". In L. P. Núñez (ed.), *Escritorios electrónicos para las literaturas: nuevas herramientas digitales para la anotación colaborativa*. Universidad Complutense de Madrid. 2013, pp. 137-152.

Santos, Diana. "Corpora at Linguateca: Vision and Roads Taken". In Tony Berber Sardinha & Telma de Lurdes São Bento Ferreira (eds.), *Working with Portuguese Corpora*. Bloomsbury, 2014, pp. 219-236.

Santos, Diana. "PANTERA: a parallel corpus to study translation between Portuguese and Norwegian". In Jon Askeland, Marco Gargiulo & Synnøve Ones Rosales (eds.). *ROM17: Anais da XX Conferência de Romanistas Escandinavos*, BeLLS 10, 1, 2019, s/pp. https://bells.uib.no/index.php/bells/article/view/1372/2844

Santos, Diana. "Doctors in lusophone literature". Blog post in Digital Literary Stylistics (SIG-DLS). 2019. https://dls.hypotheses.org/952

Santos, Diana. "Explorando o vestuário na literatura em português". *TradTerm* **37**, 2, 2021, pp. 622-643. https://www.revistas.usp.br/tradterm/article/view/170266

Santos, Diana & Cláudia Freitas. "Estudando personagens na literatura lusófona". In *STIL 2019 – XII Symposium in Information and Human Language Technology and Collocates Events, October 15-18, 2019, Salvador, BA, Proceedings of conference*, pp. 48-52. https://www.linguateca.pt/Diana/download/STIL2019_SantosFreitas.pdf

Santos, Diana, Cláudia Freitas & Eckhard Bick. "OBras: a fully annotated and partially human-revised corpus of Brazilian literary works in public domain". In *CorLex,* 24 de Setembro de 2018. https://www.linguateca.pt/Diana/download/CorLex.pdf

Santos, Diana, Emanoel Pires, João Marques Lopes, Rebeca Schumacher Fuão & Cláudia Freitas. "Periodização automática: Estudos linguístico-estatísticos de literatura lusófona". *Linguamática* 12 (1), 2020a, pp. 80-95. https://linguamatica.com/index.php/linguamatica/article/view/314/465

Santos, Diana, Daniel Alves, Raquel Amaro, Isabel Araújo Branco, Olivia Fialho, Cláudia Freitas, Suemi Higuchi, Marcia Langfeldt, João Marques Lopes, Alckmar Luiz dos Santos, Emanoel Pires, Barbara Ramos, Danielle Sanches, Rebeca Schumacher Fuão, Paulo Silva Pereira & Paula Terra. "Leitura Distante em Português: resumo do primeiro encontro". *MAT-LIT - Materialidades da Literatura* **8**, 1, 2020b, pp.

279-298. https://impactum-journals.uc.pt/matlit/article/view/2182-8830_8-1_16/6763

Santos, Diana, Alberto Simões & Cristina Mota. "Estudo de sentimentos: algumas direções". *Workshop Empirical Research on Portuguese*, Univ. de Viena, 11-12 December 2020, 2020c. https://www.linguateca.pt/Diana/download/WERP.pdf

Santos, Diana, Eckhard Bick & Marcin Wlodek. "Avaliando entidades mencionadas na coleção ELTeC-por". *Linguamática* 12 (2), 2020d, pp. 29-49. DOI: 10.21814/lm.12.2.336 https://linguamatica.com/index.php/linguamatica/article/view/336/470

Santos, Diana, Cristina Mota & Alberto Simões. "Broad coverage emotion annotation". Em apreciação. https://www.linguateca.pt/Diana/download/emosLRE.pdf

Santos, Diana. "Evaluation contests in Portuguese: Linguateca's contribution". Em apreciação. https://www.linguateca.pt/Diana/download/AvalConjLRE.pdf

Santos, Diana & Daniel Alves. "Placing GIS and NLP in literary geography: experiments with literature in Portuguese". Em apreciação. Preprint: https://www.linguateca.pt/Diana/download/AlvesSantosBILLIG_2020rascunho.pdf

Schöch, Christof, Tomaž Erjavec, Roxana Patras & Diana Santos. "Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives". *Modern Languages Open*. No prelo. Preprint: http://doi.org/10.5281/zenodo.4742420

Silva, Flávia Martins da Rosa Pereira da. "Caracterização de personagens na literatura brasileira quanto ao gênero: uma proposta metodológica". Dissertação de

Mestrado, PUC Rio, 2021. https://www.linguateca.pt/documentos/TeseMestradoFlaviaSilva2021.pdf

Teixeira, Elisa D., Diana Santos & Stella E. O. Tagnin. "CorTrad: um novo corpus paralelo multiversão para o par de línguas português-inglês". In Tania Shepherd, Tony Berber Sardinha & Marcia Veirano Pinto (eds.), *Caminhos na Linguística de Corpus*, Mercado de Letras, 2012, pp. 151-176.

Vala, Hardik, David Jurgens, Andrew Piper & Derek Ruths. "Mr. Bennet, his coachman, and the Archbishop walk into a bar but only one of them gets recognized: On The Difficulty of Detecting Characters in Literary Texts". Em *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 769–7.