Preparação para Leitura Distante em português: diálogos entre PLN e Humanidades Digitais

Luísa Rocha
Departamento de Letras
PUC-Rio & Linguateca
Rio de Janeiro, Brasil
l.rocha7@globo.com

Cláudia Freitas
Departamento de Letras
PUC-Rio & Linguateca
Rio de Janeiro, Brasil
claudiafreitas@puc-rio.br

Diana Santos
Linguateca & ILOS
Universidade de Oslo
Oslo, Noruega
d.s.m.santos@ilos.uio.no

***Abstract*—Leitura Distante é uma técnica criada por Franco Moretti que se propõe a estudar os padrões na literatura. Tendo em vista a utilização da leitura distante em obras literárias de língua portuguesa, a familiarização com a área foi importante, com o levantamento de alguns trabalhos. Deste modo, percebeu-se necessário ajustes e cuidados no tratamento do corpus OBras. A correção do gênero gramatical dos nomes próprios, a adição do papel semântico e correções de segmentação foram os ajustes feitos e abordados neste artigo para que as explorações no corpus aconteçam da melhor maneira possível.**

***Index Terms*—Distant Reading, Leitura à distância, Corpus, Literatura, Humanidades Digitais**

I. Introdução

Humanidades Digitais é uma área de atividade acadêmica que junta ciências humanas e recursos digitais para criar novos pontos de vista e métodos para esses estudos. Leitura Distante, ou ”ler à distância”, ou “distant reading” [1], é uma das técnicas nas humanidades digitais, criada por Franco Moretti, para estudar literatura de uma forma diferente. Consiste em distanciar-se das obras literárias, não como um obstáculo, mas sim como uma forma de obter novos ângulos, vendo menos detalhes e mais relações, padrões e formas em uma obra ou entre obras. Para Moretti, entender padrões entre os romances de determinado período de tempo, pode revelar uma regularidade do período, o que ele chama de “estrutura temporária”, e o gênero do romance seria seu reflexo literário.

Com o intuito de uma maior familiarização com a área, tendo em vista a utilização da leitura distante em obras literárias de língua portuguesa, foi feito um levantamento de alguns trabalhos para se reunir exemplos e ideias de pesquisas para se fazer no corpus OBras [2]. Composto atualmente por 246 obras de literatura brasileira em domínio público e em permanente expansão, o Corpus OBras foi criado para ser a contraparte brasileira do Corpus Vercial. Todo material está anotado com informação morfossintática e semântica, e está disponível para consulta e para ser baixado [3]. No entanto, para que as explorações aconteçam da melhor maneira possível, são necessários ajustes e cuidados no tratamento das obras, de modo a não comprometer as pesquisas, como a atribuição do gênero correto dos nomes próprios, e o acréscimo de informação semântica (semas), local, pessoa ou religião [4]. A adição de novas obras, principalmente de autoras, ao corpus, também é uma tarefa em andamento.

II. Familiarização com a área: alguns trabalhos relacionados

Os três trabalhos resumidos abaixo são exemplos de estudos que utilizam Leitura Distante para analisar algumas obras. As leituras foram importantes para futuros trabalhos com obras brasileiras.

O trabalho de [5] propõe-se a detectar diferentes emoções nos textos literários. Detecção de sentimentos é um método muito usado para comentários de usuários em redes sociais e avaliação de produtos online, então os autores se perguntaram se sua aplicação em textos literários seria tão útil quanto para a outra área. Para isso, foi criado um dicionário de alemão das palavras associadas às sete emoções de Plutchik e Ekman, e o aplicaram em um estudo de caso com dois romances de Franz Kafka. Utilizando-se do dicionário, 300 palavras foram anotadas como sentimentos pertencentes a uma das sete emoções, correspondendo aproximadamente a: raiva, desgosto, medo, felicidade, tristeza, surpresa e satisfação. Dois entre três anotadores concordaram na anotação de 85% das palavras, enquanto todos os três só concordaram em 46%, o que indica como pode ser difícil a anotação de sentimentos. Com os resultados, os autores tentaram detectar mudanças como aumento ou diminuição de sentimentos ao longo dos romances, assim como quais sentimentos são mais presentes em cada personagem.

Já o trabalho de [6] pretende examinar as redes sociais dos personagens em nove romances de Jane Austen e Charles Dickens. A análise de redes sociais (Social Network Analysis - SNA) permite pesquisas com um nível único de abstração, ao mesmo tempo em que mantém a estrutura coletiva dos personagens dentro do enredo.

Os textos vêm do Projeto Gutenberg e foram anotados manualmente por estudiosos de literatura com uma metodologia “radicalmente inclusiva”. Para anotar cada obra, primeiro foi criado um dicionário de personagem contendo uma única entrada para cada um e seus apelidos. Cada personagem se torna um nó e cada coocorrência entre nós forma uma aresta no gráfico. Foram contabilizadas todos os tipos de coocorrências e não só conversas diretas, pois essa maneira permitia capturar um maior número de interações e associações entre personagens.

Um dos resultados foi observar que as sociedades construídas por Austen parecem ser mais compactas do que as de Dickens. Outro ponto observado foi o uso de muitas micronarrativas por Dickens, apesar dos autores acreditarem que tal técnica não afeta o enredo central, sendo só um estilo pessoal do autor.

Por fim, [7] utilizam-se dos romances ingleses do século XIX para mapear as emoções da cidade de Londres. Os autores empregaram um programa de REM (Reconhecimento de Entidade Mencionada) para selecionar as passagens que se referiam à cidade, como ruas ou bairros. Essas passagens são compostas por uma sequência de duzentas palavras, tendo a localização de Londres no centro. Em seguida, 'voluntários' deveriam identificar (anotar) as emoções expressas ali. Houve muita discordância entre os anotadores 'voluntários' (foi um experimento de anotação *crowdsource*) e a anotação feita por alunos de graduçao, que serviram como grupo de controle. Por isso, os autores decidiram reduzir as opções de emoções para os opostos: "Frightening" (assustador) e "Happy" (feliz). Essas etiquetas só foram atribuídas às passagens se pelo menos metade dos anotadores concordassem na resposta. Essas mesmas passagens também foram analisadas pelo "Sentiment Analysis Program" de Standford, que utiliza um dicionário de termos positivos e negativos.

Antes de discutir os resultados, o artigo apresenta um contexto da expansão geográfica da cidade de Londres, e como era esperado que tal crescimento fosse representado na literatura. Contudo, só o centro e a parte oeste da cidade, que crescia para longe do rio, foram as mais citadas. Os bairros mais frequentes foram Westminster e The City, o que reflete a parcialidade nas representações da cidade de Londres.

Com isso, o primeiro achado da pesquisa foi a ausência do crescimento da cidade na Londres Ficcional, já que o foco nos dois bairros permanecia o mesmo. A característica heterogênea do bairro The City é um dos motivos apontado para sua permanência como local dos romances, enquanto, por outro lado, a homogeneidade de Westminister, bairro das classes mais altas, pode ser apontado como o principal motivo para sua permanência.

Um ponto interessante, ressaltado rapidamente pelos autores, é a "semântica do espaço", ou seja, o léxico que mais foi associado àquele lugar. Referente a Westminister, o léxico mostrava opulência (parques, praças e jardins), patriarcado e servidão (condes, servos, ordem) e também rituais de "socialização formal" (encontros e visitas).

Ao finalmente apresentar a "geografia do massustador ou da felicidade" na cidade de Londres, se descobriu, na verdade, uma cidade sem emoções. A maior parte das passagens foi anotada como "emocionalmente neutra" (67% com anotadores humanos e 78% com a ferramenta de *Sentiment Analysis*). Todavia, naquelas passagens marcadas com emoções, novamente West End e The City foram as regiões que mais apareceram marcadas com *feliz* e *assustador*, respectivamente.

III. Correções feitas no OBras

Os textos lidos oferecem múltiplas possibilidades de se trabalhar com a técnica de leitura distante e são inspiração para pesquisas com o corpus OBras [2], composto por obras de literatura brasileira dos séculos XVII até XX. Em pesquisas preliminares, percebemos a necessidade de fazer algumas modificações e correções na anotação para que as buscas obtivessem resultados mais precisos.

A. Correção de gênero gramatical dos Nomes Próprios sem gênero definido

Utilizando a expressão de busca [pos="PROP.*hum" & gen="M/F"] no serviço AC/DC [8] – ferramenta de acesso e interrogação de corpus –, a pesquisa retornou 4185 palavras, todos nomes próprios que o parser PALAVRAS [9] classificou como "humanos" [10] e para os quais atribuiu um gênero gramatical indefinido (Masculino/Feminino). O objetivo desta etapa foi revisar essa lista para atribuir o gênero gramatical correto e revisar a classificação semântica do nome próprio, caso fosse um erro (isto é, se o nome próprio não fosse "humano"). Abaixo, em negrito, estão exemplos de palavras anotadas como nome próprio e com gênero indefinido.

Para tal revisão, as 4185 palavras foram transformadas em um arquivo de lista tipo texto simples (txt), nomeado *lista-ACDC* e depois foram retirados todos os nomes terminados em "-a" – formando um novo arquivo de lista, nomeado "*terminados em a* – e, depois, foram retirados todos os nomes terminados em "-o" – formando um terceiro arquivo de lista, nomeado *terminados em o*. Em seguida, o arquivo *lista-ACDC* com as alterações foi renomeado para *nomes M-F*. A lista foi examinada para separar manualmente aqueles que eram erros (Exemplo 1), e nomes facilmente reconhecíveis (Exemplo 2 e Exemplo 3). Aqueles que possivelmente fossem sobrenomes não foram retirados da lista. Os erros formaram um quarto arquivo chamado *lista de erros*.

(1) id="Os_Sertões_II Prosa:prosa EdC 1902 ": **Adiante** recuava o sertanejo, recuando pelos cômodos escusos.

(2) id="A_falência Prosa:romance JLdA 1901 naturalismo_realismo": indagou **D.Joana** .

(3) :id="O_gaúcho Prosa:romance JdA 1870 romantismo, regionalismo": Enfim estava **Juca** um mancebo.

Conferimos que todos os nomes no arquivo *terminados em a* eram femininos para torná-lo o arquivo dos *nomes femininos*, assim, adicionamos os outros nomes posteriomente identificados como femininos. Também conferimos todos os nomes no arquivo *terminados em o* com o mesmo intuito, depois renomeado para *nomes masculinos*.

Uma vez que, no arquivo *nomes M-F*, só restasse nomes desconhecidos ou dúvidas, retornamos ao AC/DC para procurar as passagens em que esses nomes apareciam, no intuito de encontrar algum indicador de gênero: pronome, determinante, adjetivo ou contexto (Exemplo 4 e Exemplo 5). Às vezes, só as ocorrências não eram suficientes para determinar o gênero, então esses nomes foram separados para uma pesquisa mais

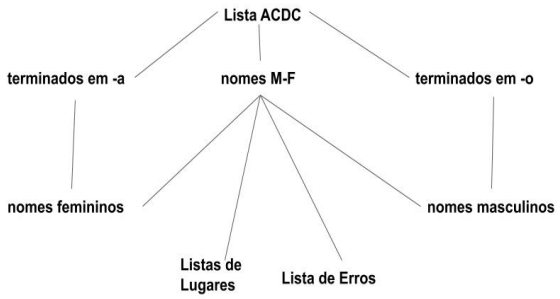

Fig. 1. Fluxograma sobre as listas de nomes próprios

detalhada a ser feita depois que o arquivo *nomes M-F* tivesse sido todo pesquisado no AC/DC. Os nomes separados foram digitados junto ao nome da obra na internet para procurar um resumo ou relação de personagem (Exemplo 6).

(4) id=”Ubijarara Prosa:romance JdA 1874 indianismo_romantismo”: Murinhém atravessou rápido a campina e apresentou-se em frente de **Canicrã**, chefe dos tapuias.

(5) id=”Macunaíma Prosa:romance MdAndrade 1928 modernismo”: Zozoiaça riu bem por causa que não podia dar **Taína-Cã** de casamento pra filha velha não.

(6) id=”Os_trabalhadores_do_mar Prosa:romance MdA 1866 ”: Nessa noite, **Clubin** ceou à mesa dos guardas das costas, e, contra o costume, saiu logo depois de cear.

Ao final, o arquivo *lista-ACDC* se repartiu em quatro arquivos de lista: *Nomes Masculinos*; *Nomes Femininos*; *Lista de Erros*; *Listas de Lugares*.

Nessa correção, optamos por deixar os pronomes de tratamento como “Vossa Excelência”, “vosmicê”, “Vossa senhoria”, “Majestade” etc. com a pos=”PROP.*hum”, pois nosso foco era nomes próprios de pessoas. No caso de nomes próprios não humanos, como o cachorro chamado Black (Exemplo 7), também não foram considerados Pessoa .

Os nomes indígenas nas obras ”Ubirajara” ”Macunaíma” e ”O ermitão do Muquém” foram interessantes de se observar, pois eram muito diferentes (Exemplo 8 e Exemplo 9). Nesse caso foi necessário uma pesquisa com contexto mais amplo e optou-se por usar a marcação pessoa ficcional.

(7) id=”Contos_fora_da_moda Prosa:conto ArtA 1894 ”: A ventura de Leandrinho tinha um único senão: havia na casa um cãozinho de raça, um bull-terrier, chamado **Black**, que latia desesperadamente sempre que farejava a presença daquele estranho.

(8) id=”Macunaíma Prosa:romance MdAndrade 1928 modernismo”: Era **Emoron-Pódole**, o Pai do Sono .

(9) id=”O_ermitão_do_Muquém Prosa:romance BG 1868 romantismo$_r$egionalismo”: – Já que Tupã, exclama ele enfurecido, obstinadamente me persegue, e me acabrunha ao peso de amarguras e infortúnios, sem que eu nada tenha feito para merecê-los, eu te invoco, ó **Anhangá**, e de hoje em diante ponho em tuas mãos o meu destino, ó manitó do mal!

1) Atribuição do sema e correções associadas: Além do gênero, o tipo semântico de cada nome próprio também foi adicionado, em uma revisão da anotação automática fornecida pelo parser PALAVRAS [9]. Marcado pela tag *sema*, a classe semântica é utilizada para indicar variadas informações semânticas. As classes semânticas podem ser várias, e em nosso caso nos interessam os tipos pessoa, lugar, obra ou religião [4]. A classe das pessoas pode ser subdividida em pessoa ficcional, usada para personagens ficcionais, e pessoa histórica, usada para pessoas que existiram. Sobre os nomes próprios referentes a religião, que foram poucos, optou-se por colocar todos com sema Religião. Assim tanto entidades e deuses quanto santos estariam com o mesmo sema. Contudo, mitologias e seres mitológicos (Exemplo 8) foram marcados com sema pessoa ficcional, [sema=Pessoa:ficc].

2) Correção de segmentação: A busca pelos nomes próprios evidenciou também muitos erros de segmentação. Encontramos diversos casos de D. ou S. que, por causa do ponto final, atrapalhavam a análise automática (Exemplo 10 e Exemplo 11), ou de nomes compostos, que eram separados, mas deveriam constituir uma unidade, apenas (Exemplo 12)

(10) id=”A_Marquesa_de_Santos Prosa:romance PS 1925 histórico”: **D.** Pedro devia conduzi-la até o salão onde estava a Imperatriz .

(11) id=”Um_dístico Prosa:conto MdA 1886 ”: Mentalmente nunca soube o que era; talvez refletia no concílio de Constantinopla, nas penas eternas ou na exortação de **S. Basílio** aos rapazes .

(12) id=”A_Marquesa_de_Santos Prosa:romance PS 1925 histórico”: **D. Ilda Mafalda de Sousa Queiroz**, a rutilante Marquesa de Valença, vestido de gorgorão negro, cadeia de ouro e mitenes de seda, corre pelos grupos o seu lorgnon de madrepérola .

B. Correção de pré-processamento

Erros de pré-processamento acontecem quando o texto não é tratado da forma correta. Os erros mais comuns eram nomes de capítulos tratados como se fizessem parte do texto do capítulo. (Exemplo 13),

(13) id=”Os_trabalhadores_do_mar Prosa:romance MdA 1866 ”: CAPíTULO II CLUBIN DESCOBRE ALGUéM

De forma semelhante, uma pesquisa com nomes próprios do tipo lugar está em andamento. Novamente, tiramos

proveito da anotação automática feita pelo PALAVRAS [9], e as expressões de busca utilizadas até o momento foram [pos="PROP.*top"] e [pos="PROP.*civ"], retornando o resultado de 3530 palavras, aproximadamente.

C. Adição de obras

As novas obras já tratadas e já adcionadas ao corpus são os cinco romances: "O Coruja", de Aluísio Azevedo, "O Bom-Crioulo", de Adolfo Caminha, "Os Dois Amores", de joaquim Manuel de Macedo, "A Carne", de Júlio Ribeiro, e "O Homem", de Aluísio Azevedo. Foram pré-processados seguindo as instruções na página do projeto. Apenas uma obra trouxe um caso não abordado nas instruções.

No romance "Os Dois Amores", um dos personagens lê uma história, de forma que está escrita para que o leitor também leia. Optou-se por não tratar esse caso como uma canção ou poema (ou seja, como "tipos especiais de texto"), mas sim como parte do capítulo a que pertence.

A pesquisa para encontrar obras escritas por mulheres, a fim de deixar o corpus mais balanceado quanto ao gênero da autoria e comparar estilos, está em andamento, porém há algumas dificuldades. As obras precisam pertencer ao domínio público e a digitalização, preferencialmente, não deve ser do tipo imagem, o que dificularia o trabalho do pré-processamento.

IV. Conclusão

Os textos estudados são fonte de várias ideias para se trabalhar com a leitura distante em português, especificamente com o corpus OBras, e para que os resultados sejam mais precisos, correções e melhorias são necessárias. Mais de mil correções envolvendo atribuições de gênero, inclusão de sema e correções de segmentação e pré-processamento foram aplicadas no corpus OBras. As correções do sema lugar serão implementadas em breve, assim como adição de obras escritas por mulheres. A melhoria e expansão do corpus sintaticamente anotado é bom para toda e qualquer pesquisa de PLN, podendo também servir de treino para modelos de aprendizado automático.

Acknowledgment

Luísa Rocha é bolsista de Iniciação Científica do Conselho Nacional de Desenvolvimento Científico e Tecnológico, no âmbito do projeto "Recursos para o 'distant reading' em português:diálogos entre PLN e Humanidades Digitais".

References

[1] F. Moretti, *A Literatura Vista de Longe*. Porto Alegre: Arquipélago, 2008.

[2] D. Santos, C. Freitas, and E. Bick, "OBras: a fully annotated and partially human-revised corpus of Brazilian literary works in the public domain." OpenCor, 2018.

[3] Linguateca. [Online]. Available: "https://www.linguateca.pt/OBRAS/OBRAS.html"

[4] D. Santos and C. Freitas, "Estudando personagens na literatura lusófona," in *Anais do STIL'2019*, Salvador, Brasil, 2019.

[5] R. Klinger, S. S. Suliya, and N. Reiter, "Automatic Emotion Detection for Quantitative Literary Studies – A case study based on Franz Kafka's "Das Schloss" and "Amerika"," in *Digital Humanities 2016: Conference Abstracts*. Kraków, Poland: Jagiellonian University and Pedagogical University, July 2016, pp. 826–828. [Online]. Available: http://dh2016.adho.org/abstracts/318

[6] S. Grayson, K. Wade, G. Meaney, J. Rothwell, M. Mulvany, and D. Greene, "Discovering Structure in Social Networks of 19th Century Fiction," in *Proceedings of the 8th ACM Conference on Web Science*, ser. WebSci '16. New York, NY, USA: ACM, 2016, pp. 325–326. [Online]. Available: http://doi.acm.org/10.1145/2908131.2908196

[7] Heuser, Ryan and Moretti, Franco and Steiner, Erik, "The emotions of london," in *Literary Lab Pamphlet 13*, 2016. [Online]. Available: https://litlab.stanford.edu/LiteraryLabPamphlet13.pdf

[8] D. Santos and E. Bick, "Providing Internet access to Portuguese corpora: the AC/DC project," in *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*. Athens, Greece: European Language Resources Association (ELRA), May 2000. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2000/pdf/85.pdf

[9] E. Bick, "The Parsing System "PALAVRAS": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework," Ph.D. dissertation, 2000.

[10] ——, "Functional Aspects in Portuguese NER," in *Proceedings of the 7th International Conference on Computational Processing of the Portuguese Language*, ser. PROPOR'06. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 80–89. [Online]. Available: http://dx.doi.org/10.1007/11751984_9