Corpus-amostra Coelho Netto: compilação, anotação e ocorrências em textos literários dos séc. XIX e XX

Francimary Macêdo Martins[1]

[1]Universidade Federal do Maranhão (UFMA)
Caixa Postal 15.064 – 91.501-970 – São Luís -MA - Brasil
francimary@dee.ufma.br

Abstract. *This paper presents doctoral thesis in progress of formation of corpus-sample Coelho Netto, a corpus compiled, annotated and analyzed morphosyntactically occurrences and opulence of verbs, adjectives and adverbs in-ly compared the texts of the same period, as well as observation the accuracy of automatic labeler Aelius (free software in Python), model Hunpos, trained at Corpus of Historical Portuguese Tycho Brahe (CHPTB) in literary texts by this author: two novels and three short stories of the period between the end of the 19th century and early 20th century. The Corpus is approximately forty-seven thousand tokens (words and punctuations).*

Resumo. *Este trabalho apresenta a tese de doutoramento em andamento da constituição do corpus-amostra Coelho Netto, um corpus compilado, anotado morfossintaticamente e analisadas as ocorrências e opulência de verbos, adjetivos e advérbios em –mente em comparação a textos do mesmo período, além de observação da acurácia do etiquetador automático Aelius (software livre em Python), modelo Hunpos, treinado no Corpus Histórico do Português Tycho Brahe (CHPTB) em textos literários desse autor: dois romances e três contos do período entre final do século XIX e início do século XX. O Corpus tem aproximadamente quarenta e sete mil tokens (palavras e pontuações).*

1. Introdução

Tendo como norteador a experiência do grupo ComPlin (Computação e Linguagem Natural) da UFC (Universidade Federal do Ceará), o CORPTEXTLIT (Corpus de Língua Portuguesa de Textos Literários do Século XIX), que anotará, morfossintaticamente, "textos de literatura brasileira do século XIX que compreenderá 40 obras do período [..] cerca de 2.500.000 *tokens*, com 10% a serem revistos manualmente" [ALENCAR, 2010b], estruturou-se o trabalho de compilação, anotação e análise de ocorrência do *corpus*-amostra Coelho Netto, doravante CACN: um *corpus*-amostra de textos literários de obras de Coelho Netto, no caso específico dois romances (*Turbilhão* e *A Conquista*) e três contos (Firmo, o vaqueiro; Mandovi e O Enterro, do livro *Sertão*) produzidos entre os anos de 1896 a 1906.

Entende-se que a disponibilização digital de textos desse autor, especificamente dois dos mais renomados romances e os principais contos de sua vertente regionalista, constitui-se em uma colaboração primordial para que o escritor possa ser lido e seus textos explorados em análises linguístico-literárias, contribuindo assim para a

Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology, pages 199–204,
Fortaleza, CE, Brazil, October 21–23, 2013.

revitalização de seus textos, que teve sua obra, às portas do Modernismo Brasileiro, intensamente negada e desqualificada pelos seus representantes, ao ponto de suas produções serem excluídas das grandes coletâneas literárias que fez com que ele fosse praticamente esquecido pelo seu país e por sua terra, não sendo lido pelas gerações futuras [NETTO, 1964]. Além de possibilitar acesso aos textos para leitura e análises pontuais, o *corpus* permite também um estudo mais completo sobre o fenômeno literário tanto do escritor quanto da época de suas produções.

A Linguística de Corpus se ressente do distanciamento que ainda persiste entre as áreas de Letras, por um lado, e Computação e Informática, por outro. Percebe-se na região Sudeste do Brasil uma crescente investigação de linguistas com o suporte da Linguística Computacional e Linguística de Corpus (LCLC). No Nordeste, as pesquisas ainda são tímidas, o que justifica que se realizem investigações utilizando a LCLC [ALENCAR, 2009]. O uso da LCLC vai possibilitar mais do que explorar um *corpus* computadorizado com possibilidades de mais precisão e maior qualidade na análise dos dados, mas também contribuir com a disponibilização de textos literários de autores não tão lidos, muito ricos em produção e poucos reconhecidos pela comunidade acadêmica, enriquecendo o banco de dados de corpora já existentes no Brasil e na Europa, como: CHPTB, CORPTEXTLIT, Linguateca, Floresta Sintática, Corpus de Araraquara, Lácio-Web/NILC, CE-DOHS dentre outros [BERBER SARDINHA, 2000; ALUÍSIO E ALMEIDA, 2006, SARDINHA, 2004].

2. O *Corpus*-Amostra Coelho Netto (CACN)

O *corpus*-amostra será denominado assim porque que se trata da compilação somente de dois romances e três contos do acervo da obra de Coelho Netto. Considerando-se o volume de sua obra publicada, em torno de dezessete romances, dezessete livros de contos, mais de mil artigos, diversos contos avulsos, comédias, peças teatrais etc. [NETTO, 1964], o *corpus* que será compilado constitui-se somente como uma amostra de tudo que foi produzido pelo escritor. O *corpus* compreende dois romances: *A conquista* (1899 – séc. XIX): cap. I, VI, XII, XVIII, XXVIII; *Turbilhão* (1906 – séc. XX): cap. I, V, X, XV, XX, XXV. E três contos do livro *Sertão* (1896 – séc. XIX): Firmo, o vaqueiro; Mandovi e O Enterro. O CACN conterá aproximadamente 47.000 *tokens*. É um *corpus*, relativamente, pequeno [BERBER SARDINHA, 2000]. O tamanho do *corpus* depende dos objetivos da pesquisa: "O *corpus* é uma amostra de uma população cuja dimensão não se conhece (a linguagem como um todo). Desse modo, não se pode estabelecer qual seria o tamanho ideal da amostra para que ela represente esta população" [BERBER SARDINHA, 2000, p. 342]. Neste caso, o tamanho do *corpus* foi delimitado conforme entendimento que para a produção de uma tese seria suficiente para o início de ações posteriores de aumento do *corpus* demandados em projetos de pesquisas destinados exclusivamente a criação e compilação de *corpora* robustos, como no caso do CORPTEXTLIT.

O *corpus* de comparação servirá como base para verificar a acurácia do anotador (Aelius, modelo Hunpos) em comparação a outros textos literários do séc. XIX e XX e também da ocorrência de verbos, adjetivos e advérbios em –mente. Esse *corpus* consiste de textos dos escritores Machado de Assis - *Esaú e Jacó* (1904): cap. 01, 04 e 08; e Camilo Castelo Branco - *Amor de Perdição* (1862): cap. 01, 02 e 04. O primeiro é do português brasileiro e o segundo do português de Portugal. Ressalta-se que estes textos

foram escolhidos para servir de comparação com os textos de Coelho Netto por esses escritores sofreram da mesma crítica pelos modernistas: de utilizarem linguagem muito rebuscada em seus escritos, pois "foram vítimas das mesmas acusações (de ser rebuscado e prolixo em seus textos), os escritores: Chateaubriand, Victor Hugo, Eça de Queiroz, Camilo Castelo Branco, Castro Alves, Álvarez de Azevedo, José de Alencar, Rui Barbosa, Euclides da Cunha, todos os românticos e os naturalistas também" [BROCA, 1958, p. 3]. Convém ressaltar que o conceito o rebuscamento adotado provinha do entendimento dos modernistas à época, como no trecho:

> Levantou-os, de novo, acima da cabeça, as mãos juntas, estrincando os dedos enclavinhados e bocejou, espichando-se nas pontas dos pés caindo depois, rijamente, sobre os tacões. Já as primeiras páginas haviam descido para a clichagem. Embaixo, martelavam pancadas crebas, como de matracas [NETTO, 1958, p. 829].

3. Ferramentas Computacionais

Para o processamento automático do CACN está sendo utilizada a linguagem de programação Python, intuitiva, de fácil assimilação, mas de altíssimo nível, orientada a objeto, de tipagem dinâmica e forte, interpretada e interativa, do tipo *open-source* (código-fonte aberto) podendo ser rodado em Windows, Linux / Unix, Mac OS X [PYTHON.ORG., 2012]. É acessado através da IDLE (*Interactive Development Environment*), traduzido como ambiente de desenvolvimento interativo. Python permite análises que buscam observar concordâncias, frequências, quantidades, repetições, diversidades, exceções, listas, extrações, sentenciação, toquenização (segmentação de um texto em unidades menores), anotações morfossintáticas dentre outros. Python utiliza a biblioteca NLTK que contém um conjunto de ferramentas necessárias à construção de programas para manipulação de dados voltados para trabalhar com o Processamento da Linguagem Natural (PLN). Tem também código-fonte livre e aberto e é uma ferramenta escrita em Python e para Python. A versão do NLTK utilizada nesse trabalho é a versão 2.0.1rc1 compatível com Python 2.7.3 instalados para o processamento do *corpus*.

O etiquetador morfossintático utilizado é o Aelius que faz parte do projeto *Aelius Brazilian Portuguese POS-Tagger,* registrado no SourceForge.net[1], e destina-se "ao pré-processamento de textos, construção de etiquetador morfossintático, anotação de corpora e auxílio de revisão humana de anotação automática" [ALENCAR, 2010a, p. 2]. O modelo utilizado é o Hunpos, treinado em textos do português do séc. XIX, do *Corpus* Histórico do Português Tycho Brahe (CHPTB), e que demonstrou melhor performance no que se refere à acurácia da anotação em textos literários desse período, de 95%, como constatou Alencar [2011].

4. Procedimentos

A compilação, anotação e análise de um *corpus* cumprem etapas distintas como referenciadas em Sardinha [2004] e Aluísio e Almeida [2006], que se configuram em etapas metodológicas para sua composição: seleção de textos; compilação e manipulação

[1] SOURCEFOGE.NET - Maior hospedagem mundial de software de código aberto. Disponível em: http://sourceforge.net/

do *corpus* que consiste primeiramente em armazenar/arquivar os textos selecionados; limpeza, edição e atualização dos textos/arquivos para o processamento computacional (retiram-se imagens, gráficos, tabelas, nº de páginas, cabeçalhos, rodapés e outras informações que não são propriamente o texto), a edição está relacionada com a comparação do texto digitado com o texto original impresso (necessária para a correta análise do anotador morfossintático), e a atualização refere-se à uniformização dos textos de acordo com a norma ortográfica vigente para que o processador possa fazer a anotação; nomeação de arquivos feita após conversão dos textos no formato "txt"; e anotação morfossintática do *corpus* compilado que consiste em: "atribuir um rótulo ou etiqueta (tag) [...] a cada palavra da língua, símbolo de pontuação, palavra estrangeira, ou fórmula matemática de acordo com o contexto em que aparecem" [AIRES, 2000, p. 8].

Após geração dos arquivos anotados, deve-se coletar uma amostra do material anotado que deverá ser corrigido manualmente, em torno de 10% do *corpus* anotado (11.700 *tokens*). A correção manual é realizada no próprio arquivo anotado que, após execução de comandos específicos, em Python, gera um outro arquivo com o texto corrigido. Somente após essa correção é que são realizadas as análises linguísticas estabelecidas a *priori* neste trabalho: frequências, quantidades, repetições, diversidade lexical (frequência de verbos, adjetivos e advérbios em *–mente).*

5. Resultados em andamento

Ainda em fase de conclusão, o trabalho apresentou alguns resultados referentes a todo o processo de compilação, anotação e análise do CACN. Concluída a fase inicial do trabalho, de seleção dos textos e digitalização, concluímos o processo de edição e atualização dos dois romances e dois contos. Feita a anotação morfossintática automática estamos em processo de correção manual do corpus para gerar arquivo para correção automática. Os processos que demandam mais precisão são os de compilação e correção manual dos textos anotados morfossintaticamente. As outras ações, relacionadas às análises linguísticas são geradas computacionalmente e automaticamente por comandos pré-estabelecidos, portanto com maior velocidade e eficiência. Convém destacar que é primordial um trabalho acurado nas etapas anteriores, para que os resultados do processamento computacional sejam efetivos. Alguns exemplos do que já foi revelado no processo de compilação:

- quanto à edição (comparação do texto digitado com o texto original impresso): *"Nas casas dos escravos,* ***às vezes****, à noite, ensaiavam as crianças."* (texto digital); *"Nas casas dos escravos,* ***as velhas****, à noite, ensaiavam as crianças."* (editado, conforme texto impresso).

- quanto à anotação morfossintática realizada pelo Aelius: *"Revistas/N-P as/D-F-P últimas/ADJ-F-P provas/N-P do/P+D conto/N de/P Aurélio/NPR Mendes/NPR o/D Anacharsis/N dos/P+D-P "/QT Idílios/NPR-P pagãos/ADJ-P "/QT ,/, Paulo/NPR Jove/NPR arredou/VB-P a/D-F cadeira/N e/CONJ pôs/VB-D -/+ se/SE de/P pé/N ,/, desabafando/VB-G ./."*

- quanto à correção manual do texto anotado: *"Revistas/N-P as/D-F-P últimas/ADJ-F-P provas/N-P do/P+D conto/N"* (texto anotado pelo Aelius); *"Revistas/N-P****@VB-AN-F-P***

as/D-F-P últimas/ADJ-F-P provas/N-P do/P+D conto/N" (correção manual para correção automática).

Este trabalho possibilitará à comunidade linguística diversas análises linguísticas, tanto em questões literárias quanto nas questões morfológicas e lexicográficas, e outras possibilidades de análises que interessem aos pesquisadores. Serão relevantes os resultados da observação quanto a diversidade lexical dos textos de CN, que levará a verificar, especificamente, a frequência de verbos, adjetivos e advérbios em *–mente* nos textos e se há ou não muitas repetições, revelando assim a diversidade lexical afirmada pelo seu filho Coelho Netto [1964] e pela crítica literária que se debruçou em analisar sua obra. Não existe, nos grandes *corpora* de textos em português, um *corpus* significativo de textos literários do português brasileiro dos séc. XIX e XX anotados morfossintaticamente para o tratamento linguístico-computacional. Espera-se que ao concluir a tese, contribuamos com a Linguística de Corpus que "têm olhado a língua, tradicionalmente, por ângulos diferentes" [BERBER SARDINHA, 2003, p. 1], no sentido de investigar os gêneros por meio de um grande número de textos, os quais formam um *corpus* eletrônico.

7. Referências

AIRES, Rachel. V. X. (2000). "Implementação, adaptação, combinação e avaliação de etiquetadores para o português do Brasil". Tese de Mestrado. Instituto de Ciências Matemáticas de São Carlos, Universidade de São Paulo. Disponível em: <http://www.linguateca.pt/Repositorio/Aires2000.ps> Acesso em: 21 jan. 2013.

ALENCAR, Leonel F. de. (2009). "Técnicas em software livre para exploração de corpora do português livremente disponíveis na WWW". UFC. In: *Veredas on-line* – linguística de *corpus* e computacional, p. 134-150. Disponível em: <http://www.ufjf.br/revistaveredas/files/2009/11/ARTIGO-Leonel-Figueiredo-de-Alencar.pdf>. Acesso em: 20 jan. 2013.

ALENCAR, Leonel F. de. (2010a). "Aelius: uma ferramenta para anotação automática de corpora usando o NLTK". IX Encontro de Linguística de *Corpus*. Porto Alegre, PUCRS, 8 e 9 de outubro. Disponível em: <http://corpuslg.org/gelc/media/blogs/elc2010/slides/Figueiredo_de_Alencar.pdf>. Acesso em: 20 jan. 2013.

ALENCAR, Leonel F. de. (2010b). "CORPTEXTLIT – Corpus de Língua Portuguesa de Textos Literários do Século XIX". Fortaleza: [s.n.], Disponível em: <http://complin.blogspot.com.br/2012/03/corpus-de-textos-historicos.html>. Acesso em: 20 jan. 2013.

ALENCAR, Leonel F. de. (2011). "Utilização de informações lexicais extraídas automaticamente de corpora na análise sintática computacional do português". In: *Rev. Est. Ling.*, Belo Horizonte, v. 19, n. 1, p. 7-85, jan./jun. Disponível em: <www.periodicos.letras.ufmg.br/index.php/relin/article/.../2505> Acesso em: 20 jan. 2013.

ALUÍSIO, S.M.; ALMEIDA, G.M.B. (2006). "O que é e como se constrói um corpus? Lições aprendidas na compilação de vários corpora para pesquisa linguística". In: *Calidoscópio* (UNISINOS). Vol. 4, n. 3, p. 155-177, set/dez. Disponível em:

<http://www.unisinos.br/publicacoes_cientificas/images/stories/pdfs_calidoscopio/vol4n3/art04_aluisio.pdf>. Acesso em: 20 jan. 2013.

BERBER SARDINHA, T. (2003). "Análise de Gênero e Linguística de Corpus: identificação das unidades internas do gênero por meio da padronização lexical". PUC: SP. Disponível em: <http://www2.lael.pucsp.br/direct/DirectPapers51.pdf>. Acesso em: 28 jan. 2013.

BERBER SARDINHA, Tony B. (2000). "Linguística de Corpus: histórico e problemática'. In: *DELTA* [online]. vol.16, n.2, pp. 323-367. Disponível em <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-44502000000200005>. Acesso em 15 jan. 2013.

BROCA, Brito. (1958). "Nota Preliminar - Coelho Netto, romancista". In: C*oelho Netto: obra seleta: romances.* V. 01. Biblioteca Luso-brasileira. Rio de Janeiro: José Aguilar. p. 1-26.

NETTO, Paulo C. (1964). "Coelho Netto: biografia para a juventude". Rio de Janeiro: Ed. Minerva.

NETTO, Coelho. Turbilhão. IN: NETTO, Coelho obra seleta: romances. V. 01. Biblioteca Luso-brasileira. Rio de Janeiro: José Aguilar, 1958. p. 827-1067.

"PYTHON.ORG". (2013) Python Programming Language – Official Website, 2013. Disponível em: <http://www.python.org/> Acesso em: 28 jan. 2013.

SARDINHA, T. B. (2004). "Linguística de Corpus". São Paulo, Manole.