UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE HUMANIDADES

DEPARTAMENTO DE LETRAS VERNÁCULAS

PROGRAMA DE PÓS-GRADUAÇÃO EM LINGUÍSTICA

CURSO DE DOUTORADO INTERINSTITUCIONAL UFC/IFMA/UFMA/SEDUC

FRANCIMARY MACÊDO MARTINS

COMPILAÇÃO, ANOTAÇÃO E ANÁLISE LINGUÍSTICO-COMPUTACIONAL DO *CORPUS* COELHO NETTO, UM *CORPUS* DE TEXTOS LITERÁRIOS DOS SÉC. XIX E XX

FORTALEZA

2014

FRANCIMARY MACÊDO MARTINS

COMPILAÇÃO, ANOTAÇÃO E ANÁLISE LINGUÍSTICO-COMPUTACIONAL DO *CORPUS* COELHO NETTO, UM *CORPUS* DE TEXTOS LITERÁRIOS DOS SÉC. XIX E XX

Tese apresentada ao curso de Doutorado em Linguística (DINTER) do Programa de Pós-Graduação em Linguística do Departamento de Letras Vernáculas da Universidade Federal do Ceará, como requisito parcial para obtenção do grau de Doutor em Linguística.

Linha de pesquisa: Linguística Aplicada

Orientador: Prof. Dr. Leonel Figueiredo de Alencar Araripe

FORTALEZA
2014

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca de Ciências Humanas

M343c Martins, Francimary Macêdo.
Compilação, anotação e análise linguístico-computacional do corpus Coelho Netto, um corpus de textos literários dos séc. XIX e XX / Francimary Macêdo Martins. – 2014.
210 f. : il. color., enc. ; 30 cm.

Tese (doutorado) – Universidade Federal do Ceará, Centro de Humanidades, Departamento de Letras Vernáculas, Programa de Pós-Graduação em Linguística, Fortaleza, 2014.
Área de Concentração: Linguística aplicada.
Orientação: Prof. Dr. Leonel Figueiredo de Alencar Araripe.
.
1. Linguística de corpus. 2.Linguística – Processamento de dados. 3.Língua portuguesa – Brasil – Morfologia. 4.Língua portuguesa – Brasil – Sintaxe. 5.Coelho Netto,1864-1934 – Crítica e interpretação. I.Título.

CDD 469.5

FRANCIMARY MACÊDO MARTINS

COMPILAÇÃO, ANOTAÇÃO E ANÁLISE LINGUÍSTICO-COMPUTACIONAL DO *CORPUS* COELHO NETTO, UM *CORPUS* DE TEXTOS LITERÁRIOS DOS SÉC. XIX E XX

Tese apresentada ao curso de Doutorado em Linguística (DINTER) do Programa de Pós-Graduação em Linguística da Universidade Federal do Ceará, como requisito parcial para obtenção do grau de Doutor em Linguística. Linha de pesquisa: Linguística Aplicada.

Aprovada em: 06 de junho de 2014.

BANCA EXAMINADORA

_____
Prof. Dr. Leonel Figueiredo de Alencar Araripe (PPGL-UFC)
(Presidente-Orientador)

_____
Profª Drª. Vera Lúcia Santiago Araújo (UECE)

_____
Profª. Drª. Luana Ferreira de Freitas (UFSC)

_____
Profª. Drª Rosemeire Selma Monteiro-Plantin (PPGL-UFC)

_____
Profª Drª Maria Elias Soares (PPGL-UFC)

AGRADECIMENTOS

Agradecimento, sobretudo, a Deus, por iluminar-me para que eu pudesse vencer mais um desafio ao qual me propus, desde a participação na seleção do doutorado até a apresentação deste trabalho. Não foram tempos fáceis.

A meu anjinho da guarda, Mariana, que sei que me acompanhará em todos os momentos da minha vida.

Aos meus pais, Francisco e Mayre, que sempre tiveram como objetivo de vida propiciar uma educação de qualidade a todos nós; por isso cheguei aqui.

Aos meus irmãos, Neta, Luzi e Júnior, sempre torcendo por meu sucesso e se orgulhando dos meus êxitos.

Ao meu marido, Marcos Sales, que esteve ao meu lado nesse período, cheio de altos e baixos, mas sempre me incentivando e dando apoio.

A todos os colegas do DINTER, por vezes mais distantes que presentes, mas mutuamente nos ajudando, que colaboraram na construção desse percurso que encerro, e certa de que poderei contar com todos nos próximos percursos que “buscarei”.

Aos amigos pessoais, e muitos deles também amigos profissionais, que sempre acreditaram no meu potencial, e incentivaram-me em todos os momentos.

Ao meu chefe, prof. Quezada, que com seu apoio contribuiu sobremaneira para que eu pudesse me tornar uma doutora.

Ao meu orientador, prof. Leonel Figueiredo de Alencar, por ajudar-me na não tão simples tarefa de enveredar pelas plagas da Linguística Computacional e Linguística de *Corpus*. Os momentos de conflitos, de *tracebacks*, de discussão contribuíram para meu “treinamento”.

Ao Hélio Leonam Barroso Silva, aluno do Curso de Letras da UFC e bolsista do Programa de Iniciação Científica do CNPq, no âmbito do projeto "Técnicas em softwares livres para a linguística de corpus" da UFC. Sua parceria nas correções manuais dos textos anotados foi imprescindível para que eu pudesse concluí-las. Foi a clareza para minha cegueira textual.

A todos os professores do DINTER, que sem dúvida nenhuma colaboraram

para meu crescimento como estudante e profissional.

E, enfim, agradecer a todos os que não citei que, ou mais longe ou mais perto, estiveram presentes comigo nessa tarefa.

Obrigada a todos. Eu consegui!

“A minha faculdade essencial é a imaginação. Vivo a sonhar, as ideias pululam no meu cérebro e sinto que são as sementes antigas que se fazem floresta” (Anselmo, alterego de Coelho Netto, no livro *A Conquista*).

RESUMO

Esta tese é a compilação, anotação morfossintática e análise linguístico-computacional de um *corpus* de textos literários dos séc. XIX e XX: o *Corpus* Coelho Netto (CCN), contendo textos dos romances *A Conquista* e *Turbilhão* e contos do livro *Sertão*. O trabalho está na interface da Linguística de *Corpus* e da Linguística Computacional (BERBER SARDINHA, 2000, 2003, 2004, 2005, 2009; BERBER SARDINHA; ALMEIDA, 2008; OLIVEIRA, 2009; BIDERMAN, 1998, 2001; ALUÍSIO; ALMEIDA, 2006; SHEPHERD, 2012; MACENERY E WILSON, 2001; LEECH, 2004; ALVES; TAGNIN, 2012; ALENCAR, 2009, 2010a, 2010b, 2011a, 2011b, 2013a, 2013b). O CCN contém 53.080 (cinquenta e três mil e oitenta) *tokens* (pontuação e palavras). A compilação consiste nas etapas de seleção, coleta de textos e manipulação; nesta são realizadas a limpeza, edição e atualização dos textos (ALUÍSIO; ALMEIDA, 2006), para depois ser submetido à anotação morfossintática e análise linguístico-computacional, com o objetivo de obter dados que comprovem ou não o uso "excessivo" de adjetivos, de verbos e de advérbios em *–mente*, demonstrando a diversidade lexical nos textos de Coelho Netto, constatando se o que a crítica modernista dizia a respeito do escritor era procedente. A anotação morfossintática foi realizada pelo etiquetador automático Aelius, modelo AeliusHunPos, um *software* livre em Python que utiliza a biblioteca *Natural Language Toolkit* – NLTK (BIRD; KLEIN; LOPER, 2009), no pré-processamento de textos, na construção de etiquetador morfossintático e na anotação de *corpora* com auxílio de revisão humana (ALENCAR, 2010a, 2013a, 2013b), e que foi treinado no *Corpus* Histórico do Português Tycho Brahe (CHPTB). A compilação e anotação do CCN envolve outras ações como a reavaliação da acurácia desse etiquetador em textos literários. Os resultados da pesquisa revelaram que: o AeliusHunpos ao anotar os textos do CCN demonstrou maior acurácia que em outros textos já anotados, de 97,9%; que o modelo AeliusHunPos mostrou um desempenho muito além ao anotar os *corpora* que com o modelo AeliusMaxEnt; e que, após a seleção e correção manual dos 10% dos *corpora* anotados e gerados arquivos padrão *gold,* sugerimos um melhoramento dos aproximados 3% de erros cometidos pelo etiquetador, visando o aumento de sua acurácia. Quanto às analises realizadas com os dados obtidos no CCN constatamos que: a diversidade lexical, especificamente

quanto a verbos, adjetivos e advérbios em –*mente*, declarada como exagerada pela crítica à Coelho Netto não procede, pois seus textos são ricos, mas quando comparados aos textos de Aluísio Azevedo e Camilo Castelo Branco, o *Corpus* de Comparação, apresentam riqueza vocabular similar ao CCN, como expostos nos resultados.

Palavras-chave: Linguística de *Corpus.* Linguística Computacional. Etiquetagem Morfossintática. AeliusHunPos. Coelho Netto.

ABSTRACT

This thesis is the compilation, morphosyntactic annotation and linguistic and computational analysis of a corpus of literary texts of 19th and 20th centuries: Corpus Coelho Netto (CCN), containing texts of the novels A Conquista and Turbilhão and short stories of the book Sertão. The work is in the Corpus Linguistics and Computational Linguistics interface (BERBER SARDINHA, 2000, 2003, 2004, 2005, 2009; BERBER SARDINHA; ALMEIDA, 2008; OLIVEIRA, 2009; BIDERMAN, 1998, 2001; ALUÍSIO; ALMEIDA, 2006; SHEPHERD, 2012; MACENERY AND WILSON, 2001; LEECH, 2004; ALVES; TAGNIN, 2012; ALENCAR, 2009, 2010a, 2010b, 2011a, 2011b, 2013a, 2013b). The CCN contains 53.080 (fifty-three thousand and eighty) tokens. The compilation consists of the steps selection, collection off texts and handling; in which cleaning, editing and updating of texts (ALUÍSIO; ALMEIDA, 2006), and then be submitted to the morphosyntactic annotation and linguistic-computational analysis, with the goal of obtaining data to show whether or not the "excessive" use of adjectives, verbs and adverbs in "*–mente*", demonstrating the lexical diversity in Coelho Netto´s texts, noting if what the modernist critics said about the writer was correct. The annotation was performed by automatic tagger Aelius, AeliusHunPos model, free software in Python that uses the Natural Language Toolkit – NLTK library (BIRD; KLEIN; LOPER, 2009), in the pre-processing of texts, in the construction of morphosyntactic tagger and the automatic annotation of corpora with the help of human review (ALENCAR, 2010a, 2013a, 2013b), and it was trained in the Historical Corpus of Tycho Brahe Portuguese (CHPTB). The compilation and annotation CCN involves other actions such as revaluation the accuracy of this tagger in literary texts. The search results indicated that: AeliusHunpos demonstrated better performance than other texts already noted (97.9 %); AeliusHunPos model showed a far beyond performance by annotating corpora with AeliusMaxEnt model; and that, after selection and manual correction of 10% annotated corpora and generated gold standard files, it is suggested an improvement of the approximate 3% of errors by the tagger, in order to increase its accuracy. Regarding the analyzes performed with the CCN, it was found that: lexical diversity - about verbs, adjectives and adverbs in "*–mente*" considered exaggerated by critics to Coelho Netto unfounded, because his texts are rich, but when compared to the texts by Aluísio

Azevedo and Camilo Castelo Branco, comparison of corpus, present vocabulary richness similar to CCN, as exposed in the results.

Keywords: Corpus Linguistics. Computational Linguistics. Morphosyntactic tagging. AeliusHunPos. Coelho Netto.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CCN – *Corpus* Coelho Netto

CHPTB – *Corpus* Histórico do Português Tycho Brahe

CN – Coelho Netto

COMPLIN – Grupo de Computação e Linguagem Natural

CORPTEXLIT – *Corpus* de Textos Literários do séc. XIX

LA – Linguística Aplicada

LCLC – Linguística Computacional e Linguística de *Corpus*

LCOMP – Linguística Computacional

LCORP – Linguística de *Corpus*

NILC – Núcleo Interinstitucional de Linguística Computacional

NLTK – Natural Language Toolkit

PALN – Processamento Automático de Linguagem Natural

PB – Português Brasileiro

PE – Português Europeu

PLN – Processamento de Linguagem Natural

WEB - World Wide Web (Rede Mundial de Computadores) ou WWW

SUMARIO

1 INTRODUÇÃO

A criação e aplicação de recursos linguísticos no âmbito da Inteligência Artificial, das Engenharias e na Linguística fez com que se tornasse possível disponibilizar robustos bancos de dados com material necessário para análises linguísticas. Biderman (2001, p. 92), já consciente dessas mudanças, compreendia que a rede mundial de computadores propiciaria o acesso a qualquer tipo de informação, tornando global a comunicação.

O computador é sem dúvida uma ferramenta indispensável no mundo moderno, e para as pesquisas na área de língua portuguesa dão uma enorme contribuição, não somente como instrumento para se obter resultados rápidos, mas, sobretudo, por “impor” um novo meio de pensar e de agir, gerando desafios específicos para a nossa língua e para a cultura dos povos que a compartilham (BERBER SARDINHA, 2005a).

Santos (1999) entende também que a única forma de garantir que a sociedade da informação do futuro privilegie a língua portuguesa é investindo em processamento computacional da nossa língua.

A relação entre a Linguística Aplicada (LA) com as relativamente novas Tecnologias de Informação e Comunicação (TICs) está vinculada ao processamento computacional da linguagem, que passou a ser uma tarefa mais fácil e bastante acessível para os interessados nessa área. Outrora, consistia em uma tarefa além de dispendiosa bastante complicada e quase inacessível aos pesquisadores, sobretudo para aqueles que trabalhavam diretamente com a linguagem, pois eram utilizados para essa tarefa grandes *mainframes* - máquinas enormes. As operações realizadas nestas máquinas requeriam uma série de operações primitivas para se chegar a algum resultado (BIDERMAN, 2001).

Os diversos recursos instalados nos computadores (*hardwares* e *softwares*) viabilizaram a manipulação dos textos digitalmente, relegando esse processo primitivo de análise quase ao ostracismo. Com o uso massivo desses equipamentos, revolucionou-se a forma de se trabalhar os textos: as análises manuais de textos escritos tornaram-se insípidas e o editor de texto passou a ser uma banalidade (BIDERMAN, 2001). É praticamente inviável, senão impossível, a manipulação de um rico acervo de textos que não seja por via digital. Porquanto, o desenvolvimento de computadores inteligentes, mais do que ampliar nosso

conhecimento, dá-nos oportunidade de transformá-lo.

Para desenvolver a nossa língua e dar a ela o destaque devido no cenário das pesquisas contemporâneas e vinculadas às tecnologias, Santos (1999) destaca alguns recursos que devem ser criados para o desenvolvimento da língua portuguesa, dentre eles: *corpora* anotados, *corpora* analisados, *corpora* alinhados, estudos de frequência, gramáticas baseadas em *corpora*.

Como base do estudo científico da linguagem a partir de uma perspectiva computacional (BERBER SARDINHA, 2005a), a Linguística Computacional vem contribuindo significativamente para novas pesquisas geradas do cruzamento entre a Linguística e as Ciências da Computação, especificamente com a Inteligência Artificial. Ao adentrar no universo da Linguística Computacional, é possível vislumbrar potencialidades de pesquisa e de resultados exponenciais até então não possíveis em pesquisas ditas manuais. A Linguística Computacional, por Othero e Menuzzi (2005) subdivide-se em Linguística de *Corpus* e o Processamento de Linguagem Natural (PLN).

Alguns fatores dificultam as pesquisas da Linguística de *Corpus*, dentre eles Alencar (2009) destaca o distanciamento que ainda persiste entre as áreas de Letras, por um lado, e Computação e Informática, por outro. Esse distanciamento também é percebido por Dias-da-Silva (2006) quando destaca que os linguistas circunscrevem-se aos limites de sua disciplina, ocupados em estudar a linguagem humana *per se,* deixando transparecer certo descaso aos estudos computacionais da linguagem e resistem a cooperar com projetos de PLN.

Existe uma significativa carência de pesquisas, ferramentas, recursos linguísticos e humanos para tratar computacionalmente a língua portuguesa, o que mais comumente temos acesso são a pesquisas e trabalhos realizados com outras línguas (VIEIRA e STRUBE DE LIMA, 2001).

Percebe-se, na região Sudeste do Brasil, uma crescente investigação/pesquisa 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 possibilita mais do que explorar um *corpus* computadorizado com possibilidades de mais precisão e maior qualidade na análise dos dados, também contribui com a disponibilização de acervos de textos de vários gêneros, enriquecendo o banco de dados de *corpora* já existentes no Brasil e na

Europa. Neste trabalho, a LCLC e a compilação de um *corpus* literário tende a enriquecer *corpora* desse gênero, como: *O Corpus* Histórico do Português Tycho Brahe, CORPTEXLIT, Linguateca, Floresta Sintática, *Corpus* de Araraquara, Lácio-*WEB*/NILC, CE-DOHS dentre outros (BERBER SARDINHA, 2000, 2004; ALUÍSIO; ALMEIDA, 2006).

O Corpus Histórico do Português Tycho Brahe[1], doravante Tycho ou CHPTB, que serve como referência para criação de diversos *corpora* literários da língua portuguesa, é um *corpus* eletrônico parcialmente anotado (anotação morfológica e anotação sintática), com livre acesso pela internet, composto inicialmente de textos em português europeu (PE), por autores nascidos entre 1380 e 1845 (GALVES, 1999). O CHPTB não contém em seu corpo uma quantidade significativa de textos do português brasileiro (PB). Dos sete textos de PB, somente dois estão anotados morfossintaticamente[2].

Constata-se na literatura sobre LCLC que os grandes bancos de dados para processamento computacional são carentes de textos literários do português brasileiro (dos séculos XIX e XX) que estejam compilados por autor, anotados morfossintaticamente e disponíveis para o tratamento linguístico-computacional. Os bancos de dados de menor tamanho e que contêm textos literários do PB, desenvolvidos por instituições e pesquisadores brasileiros, não incluem nos seus dados textos literários no formato proposto neste trabalho.

Alencar (2010a) reforça essa constatação quando destaca que apesar da disponibilidade de vários *corpora* anotados representativos do PB, a linguística histórica do português não dispõe de uma quantidade substancial de textos anotados do século XIX, apesar da cobertura do CHPTB para o período de 1500 a 1800.

O interesse em realizar esta pesquisa teve como pressuposto inicial o resgate do acervo literário do escritor Coelho Netto, de Caxias-MA, que produziu intensamente na sua época (final do séc. XIX e até metade do séc. XX). Às portas do Modernismo literário, a obra deste escritor foi intensamente negada e desqualificada (NETTO, P., 1958, 1964; LIMA, 1958; BROCA, 1958), que resultou

[1] CHPTB – disponível em: http://www.tycho.iel.unicamp.br/~tycho/*corpus*/en/index.html

[2] Os textos do PB disponíveis no Tycho são: Atas dos Brasileiros, Cartas Brasileiras, Atas dos Africanos; Cartas para Cícero Dantas Martins (Barão de Jeremoabo), Cartas para vários destinatários, Maria ou a menina roubada, e Iracema: destes só os dois primeiros estão anotados morfossintaticamente.

em sua exclusão das grandes coletâneas literárias, fazendo com que fossem praticamente esquecidos seus feitos literários e sua representatividade na literatura brasileira em sua época. Além disso, linguisticamente, nos instigou investigar os fatos linguísticos destacados pelos modernistas ao criticar a obra de Coelho Netto.

Entendemos que a disponibilização digital de textos do autor constitui-se uma colaboração primordial para que o escritor possa ser lido e seus textos explorados em análises linguístico-literárias com viés computacional. Além de possibilitar acesso aos seus textos e leitura destes, permitirá também um estudo mais completo sobre o fenômeno literário tanto do escritor quanto da época de suas produções, observando-se, sobretudo, dados da variedade do português escrito.

Vislumbrando as possibilidades que as tecnologias permitem de se trabalhar com textos diversos de uma forma mais abrangente, com uma precisão de coleta de dados não possível de forma manual, e também inspirados na experiência do grupo de estudos CompLin (Computação e Linguagem Natural)[3] da UFC e do CORPTEXLIT (*Corpus* de Textos Literários do séc. XIX), estruturamos este trabalho de compilação do *Corpus* Coelho Netto, doravante CCN, um *corpus* de textos literários de algumas obras de Coelho Netto: os romances *Turbilhão* e *A Conquista,* e três contos (*Firmo, o vaqueiro*; *Mandovi* e *O Enterro)* do livro *Sertão*, produzidos entre os anos de 1896 e 1906, portanto séculos XIX e XX. 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, P., 1964), o *Corpus* compilado constitui-se somente como amostra do que foi produzido pelo escritor.

Paralelamente à compilação do CCN, compilamos um *corpus* de comparação (*Corpus* de Contraste) composto de obras de escritores com textos do mesmo período: Aluízio Azevedo e Camilo Castelo Branco.

Os *Corpora* deste trabalho são considerados pequenos (BERBER SARDINHA, 2000), pois contêm menos de 80 mil *tokens*[4]: o CCN contém 53.080 *tokens* e o *Corpus* de Contraste 53.971 *tokens*.

O objetivo geral traçado para a realização deste trabalho consiste em compilar, anotar morfossintaticamente e realizar análise linguístico-computacional de

[3] Homepage do Grupo CompLin: http://complin.blogspot.com.br/
[4] *Tokens* compreendem palavras e pontuações.

um *corpus* de textos literários do final do séc. XIX e início do séc. XX da obra de Coelho Netto, constituindo assim o *Corpus* Coelho Netto, utilizando o etiquetador automático Aelius, modelo AeliusHunpos, treinado no *Corpus* Histórico do Português Tycho Brahe (CHPTB).

Todos os *corpora* foram anotados morfossintaticamente (CCN e *Corpus* de Contraste), sendo que em 10% de cada *corpus* realizamos correção manual da anotação, por dois anotadores, com vistas, sobretudo, a gerar dados para as posteriores correções do etiquetador Aelius e aumentar sua precisão ao anotar textos literários. A intenção da correção é reavaliar o desempenho linguístico do etiquetador com o CCN e com o *Corpus* de Contraste. Para as análises linguístico-computacionais serão utilizados os *Corpora* anotados em sua totalidade.

Todo o trabalho é norteado por objetivos específicos para consecução do objetivo geral, como:

a) compilar textos literários do escritor Coelho Netto, do final do séc. XIX e início do séc. XX;
b) compilar o *Corpus* de Contraste com textos literários de autores contemporâneos de CN, especificamente Aluísio Azevedo e Camilo Castelo Branco;
c) anotar morfossintaticamente o CCN e o *Corpus* de Contraste, utilizando o etiquetador automático Aelius, modelo AeliusHunPos;
d) corrigir manualmente 10% do CCN anotado e do *Corpus* de Contraste, para verificação e comparação de acurácia do etiquetador utilizado;
e) construir, a partir do texto corrigido manualmente, um padrão *gold* dos C*orpora;*
f) avaliar a acurácia do etiquetador AeliusHunPos nos *Corpora*, comparando com o modelo AeliusMaxEnt, todos treinados pelo Tycho;
g) relacionar a ocorrência de verbos, adjetivos e advérbios em *-mente* no CCN, em contraste com outros textos literários da mesma época, nesse caso, os textos do *Corpus* de Contraste;
h) disponibilizar estes *corpora* em repositórios de corpus computadorizados acessíveis pela *WEB.*

Convém referenciar que a realização e a conclusão deste trabalho, que consiste em compilar *corpora* literários, embutem diversas ações que contribuem para o melhoramento das ferramentas computacionais utilizadas na Linguística de *Corpus*, como é o caso do etiquetador Aelius. Sua reavaliação e aprimoramento são etapas inerentes e consequentes ao processo de anotação morfossintática. Também, a ação de comparação dos textos digitais com os textos impressos, no processo de compilação do corpus, consiste em uma ação importante para a validação de recursos computacionais disponíveis na WEB que fazem o tratamento/conversão de textos escaneados em textos digitais: caso do programa OCR (*Optical Character Recognition*), recurso automático de reconhecimentos de caracteres, muito utilizado nessa conversão.

Delineados os objetivos específicos, e considerando-se a literatura sobre LCLC e sobre Coelho Netto, emergem alguns questionamentos abrangentes em relação ao objeto da pesquisa:

- Qual a acurácia do etiquetador Aelius, modelo AeliusHunPos, ao anotar o CNN, considerando-se os resultados já obtidos com outros *corpora*?
- Qual o desempenho do AeliusHunPos ao anotar o *Corpus* de Contraste?
- Há nas obras de Coelho Netto uma diversidade lexical muito grande, como explorado na literatura sobre o escritor, com destaque para os verbos, adjetivos e advérbios em *-mente*?
- No CCN há uma maior frequência de verbos, adjetivos, e advérbios em *– mente*, em contraste a textos de escritores contemporâneos a CN, confirmando o que declarava a crítica sobre os textos do autor?

Os resultados encontrados contribuem significativamente para áreas distintas, mas convergentes, como:

- Linguística Computacional: avaliação da acurácia para melhoria no desempenho do etiquetador morfossintático, o Aelius, modelo AeliusHunpos, subsidiando-o com dados para que possa ser aprimorada sua anotação, e gerar melhores resultados, cometendo o menor número de erros possíveis no processo de análise morfossintática automática;

- Processamento da Linguagem Natural: utilização do computador e seus programas para a realização de análises automáticas linguísticas, com programas apropriados para essas ações, como: Python, NLTK, Aelius dentre outros. Todos

são testados durante as ações de análises, validando ou não sua precisão, e fazendo as devidas adaptações melhorando sua qualidade;

- Linguística de *Corpus*: compilação, anotação, análise e disponibilização de *corpora* de textos literários, enriquecendo acervos linguístico-computacionais do PB, pois sua análise se volta para a busca de específicos fenômenos linguísticos determinados nesta pesquisa;

- Literatura (estudos de acervos literários): análise computacional da linguagem literária, buscando eventos linguísticos que possam contribuir para a compreensão do movimento literário vivenciado à época, nesse caso específico rebater ou corroborar o que a crítica modernista atribuía à produção do escritor Coelho Netto.

Concluído o trabalho, submeteremos os *Corpora* à equipe do Linguateca para apreciação, adequação e autorização de publicação em seu Repositório de Recursos, ficando disponível para acesso livre no link http://dinis.linguateca.pt/Repositorio/, como outros quatorze já existentes, dentre eles somente quatro são do Brasil.

Para tanto, esta Tese está estruturada em seis capítulos, a partir deste capítulo introdutório (capítulo primeiro), em que referenciamos e contextualizamos a proposta deste trabalho, demonstrando sua importância e em que bases iniciais se fundamenta, bem como destacando os objetivos da pesquisa e as intenções dela.

No capítulo dois consta toda a fundamentação teórica sobre os temas que permeiam o trabalho, buscando, à luz da literatura específica, as justificativas e confirmações teóricas para o porquê dessa pesquisa. Especificamente, com sessões que tratam sobre o escritor Coelho Netto, cujos textos serão utilizados para compilação, anotação e análise linguístico-computacional; sobre Linguística Computacional, com os conceitos atribuídos à área e suas principais características; sobre Linguística de *Corpus*, expondo seus conceitos e suas características específicas.

O terceiro capítulo é destinado à descrição da metodologia deste trabalho. Apresentamos os dois *corpora* da pesquisa: o CCN e o *Corpus* de Contraste, com as obras escolhidas; os instrumentos para tratamento dos dados, detalhando os vários recursos tecnológicos utilizados; e finalmente todas as etapas do processo de criação do *Corpus* Coelho Netto e do *Corpus* de Contraste, relatando as ações e as dificuldades e ou limitações encontradas nesse percurso.

No capítulo quatro expomos os resultados das anotações e análises dos dados obtidos dos *Corpora*.

Em seguida, finalizando a tese, faremos considerações que se não são conclusivas, mas são informativas no sentido de retomar as partes mais relevantes no processo de compilação, anotação e análise linguístico-computacional do CCN, enfatizando sua importância, e apresentando algumas perspectivas de investigações futuras, senão de continuidade da pesquisa desta tese.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, revisamos os aspectos teóricos que embasam nosso trabalho. Inicialmente fazendo um breve histórico sobre Coelho Netto, de modo que seja possível compreendermos melhor sua produção literária e, sobretudo, o processo crítico a que foi submetido. Destacamos as obras de grande destaque do escritor e as que foram escolhidas para comporem o *Corpus* Coelho Netto.

Em sequência, detalhamos os fundamentos que sustentam a Linguística Computacional e a Linguística de Corpus, mostrando os conceitos atribuídos a essas áreas e suas principais características. Damos destaque às ferramentas computacionais utilizadas neste trabalho e as etapas necessárias à compilação e anotação de um *Corpus.*

2.1 Coelho Netto

Considerando-se a tão pouca exposição sobre os fatos pessoais e históricos sobre o escritor, convém descrever uma sucinta biografia de sua vida. Afora alguns dados sobre CN serem de domínio público, destacamos outros relevantes que se configuram em uma fonte de referência para conhecimento de sua vida literária, especialmente quanto à crítica feita à sua obra.

Henrique Maximiano Coelho Netto, escritor maranhense, do município de Caxias, mesmo berço de Gonçalves Dias, foi o escritor nacional mais opulento em obras e em vocábulos do séc. XIX. Foi um dos raríssimos escritores que vivia da pena e para a pena, e produziu incessantemente, como exigia o seu tempo e suas necessidades (NETTO, P., 1964).

Nascido aos 21 dias de fevereiro de 1864, dia de São Maximiano, era filho da miscigenação entre português (seu pai, o negociante Antonio de Jesus Coelho), e índio (sua mãe, a amazonense civilizada Ana Silvestre Coelho). Foi criado em meio à austeridade e aos rígidos princípios morais do pai e à ternura da mãe, e teve como “mãe-de-berço” a sua ama, D. Rosária, que desde cedo o envolveu com estórias simples e encantadoras, que iriam lhe servir, no futuro, de inspiração em diversas de suas obras (NETTO, P., 1964). Mudou-se para o Rio de Janeiro com seis anos de idade, e a convivência com escritores notáveis da época, inclusive maranhenses, contribuiu para que desenvolvesse suas preferências tendenciais para a literatura e

o jornalismo. Aos 35 anos fez um retorno breve à sua terra natal, e na ocasião recebeu uma grande manifestação jamais vista no Maranhão. No seu transcurso pelo estado, sua influência foi decisiva para o renascimento da literatura maranhense, como pronunciou Antônio Lobo (NETTO, P., 1964, p. 48)**:**

> Data, com efeito, da passagem de Coelho Netto pelo Maranhão, o início da vigorosa e promissora renascença literária a que de presente assistimos. [...] preparando surdamente, em todos, os cérebros aptos à prática das letras, o belíssimo movimento literário que ora se nos depara nas Atenas Brasileira (*sic*).

Morreu aos 70 anos, acometido de intoxicação urêmica, em 28 de dezembro de 1934; "já não via, não ouvia, não sentia" (NETTO, P., 1964, p. 124). Foi o fundador da cadeira nº 2 da Academia Brasileira de Letras (ABL), e Patrono da cadeira nº 24 da Academia Maranhense de Letras (AML).

2.1.1 Sua Obra

Conforme biografia feita por seu filho (Paulo Coelho Netto), Coelho Netto*,* no período de 1891 a 1934, publicou 250 edições de suas obras, com um total de 600.000 volumes, além de 8.000 artigos para jornais do país e do estrangeiro e 17 romances traduzidos para 11 idiomas: foram 112 volumes publicados, 5 inéditos, 4 inconclusos e 9 considerados perdidos (MORAES, 1976). Produção tão numerosa assim, Lima (1958) compara somente com a de Balzac e Camilo Castelo Branco.

Em comemoração ao Centenário de nascimento de Coelho Netto, a Biblioteca Nacional realizou a Exposição Coelho Neto, em 1964. Por ocasião da exposição, foi lançado um documento denominado Exposição Coelho Neto (BIBLIOTECA NACIONAL DO RJ, 1964) onde consta todo o trabalho intelectual de Coelho Netto, perfazendo um total de 257 itens, entre autógrafos, livros, discursos, crônicas, fotografias e objetos pessoais.

Apesar de sua maior produção literária ter ocorrido no período de influência do Realismo, Simbolismo e às portas do Modernismo, classificar Coelho Netto em um estilo literário específico é praticamente impossível. O fato de não se prender a um só gênero rendeu a CN a crítica de ser um fabricante de romances (PINHO, 2009).

Sobre essa discussão, Broca (1958, p. 26) ressalta que CN "[...] conservando o fundo romântico, conseguiu, entretanto, algumas adaptações felizes ao realismo, e por aí devemos julgá-lo e situá-lo na literatura brasileira".

Paulo Coelho Netto (1964) resume assim toda a produção de CN:

> ele abordou o romance, o conto, a novela, o drama, a comédia, o sainete, a crítica, a história, a poesia, o episódio lírico, o poema dramático, a pastoral, a alegoria, a conferência, a fábula, o apólogo, a crônica, a paráfrase, a narrativa de viagem, a reportagem, a nota humorística, o artigo doutrinário de jornal, o artigo de polêmica e o discurso parlamentar (p. 131).

Péricles Morais, escritor amazonense, citado em Lima (1958), que consagrou um estudo mais acurado da obra de Coelho Netto, quanto ao fato de tentarem fixar CN em uma específica escola literária, disse:

> Naturalista impregnado de romantismo, romântico de transição, ou melhor, realista com ilustrações românticas, atraído pelo naturalismo [...] a sua individualidade literária resulta da combinação de influência inteiramente divergentes que se exerceram no escritor pela mesma forma e com a mesma intensidade. A sua literatura confunde todos os gêneros e, paradoxalmente, conserva a unidade em tão vasta multiplicidade (p. 93).

Havia uma forte pressão por parte dos críticos, sobretudo os modernistas e em especial Lima Barreto, para que CN aderisse a uma escola estilística, e sobre isso Bezerra (1982) destaca o argumento de CN:

> Querem que eu modifique o meu vocabulário, e que escreva como fulano ou sicrano; mas, se o meu estilo é este, se foi nele que escrevi a minha obra; se for ele que me dá uma individualidade, como se pode compreender que eu o repudie, adotando outro? Camilo tinha o seu modo de escrever; Euclides o seu. Eu tenho o meu. Estou no meu direito (p.10).

Sua primeira produção literária foi a poesia abolicionista *No Deserto,* publicada no "Jornal do Comércio"*,* aos 17 anos.

Publicou seu primeiro romance utilizando o pseudônimo Anselmo Ribas, chamado de *A Capital Federal* (1893): o romance é uma sátira aos contrastes da civilização diante da vida interiorana e da segurança tranquila da vida provinciana: CN oscila entre o apego às origens maranhenses e à aculturação carioca. A partir de deste romance, a crítica da época começou a observá-lo, e acusavam-no de ser altamente influenciado por Eça de Queiroz, comparando *A Capital Federal* com *A Cidade e as Serras,* afirmando que as semelhanças entre os romances eram

notórias. É importante ressaltar que o romance de Eça só foi escrito oito anos após o de CN.

Em 1894 surgiu um romance considerado, pelos que o criticavam, como uma obra aceitável: *Miragem.* Neste livro não se observa o preciosismo que, segundo a crítica, diziam que o autor utilizava em demasia. O romance reflete impressões pessoais, com tipos reais; é realista, tendo um fundo romântico de contornos poéticos. Obra aceitável, mas não menos criticada, pois diziam ser este romance pura invenção do escritor, em se tratando do local onde foi, em sua maior parte, ambientado. Coelho Netto, em protesto, relatou (NETTO, P., 1958, p. 40):

> Este romance [...] foi apreciado e combatido. Críticos tabernários, que não o leram, zurziram com muito solecismo embebido em aguardente. Os pendengos [...] diziam que tudo em tal livro era pura invenção, que eu, de Vassouras, conhecia apenas as de varrer. [...] do episódio inicial fui testemunha. Conheci Maria Augusta e Tadeu; deste vi, mais de uma vez, o sangue das hemoptises. Minha é a parte relativa à vida militar. Fi-la para aproveitar o que sabia dos pródromos da República e para fixar as minhas impressões da manhã histórica [...] de 15 de novembro. O ferreiro Nazário é uma das tristes verdades da minha obra. Conversei muita vez esse grande infeliz (*sic*).

Ressaltando uma produção estritamente qualitativa, além da quantitativa já mencionada, destacam-se *Turbilhão, Treva, Sertão, A Conquista, O Morto* como significativas o bastaste para fazer o nome perdurável de qualquer ficcionista (LIMA, 1958).

Em função de uma “intransigente condenação à sua obra novelística” (ADONIAS FILHO, 1964, s.p.), Coelho Netto teve sua produção literária esquecida por durante praticamente quarenta anos desde sua morte. Sobre essa “condenação” fazemos, a seguir, um apanhado desse momento.

2.1.2 A crítica

A vida literária de Coelho Netto foi recheada de glórias e de dissabores, teve grande êxito e consagrou-se como Príncipe dos Prosadores Brasileiros[5]. Foi alvo da maior campanha de descrédito por parte dos artistas de 22, e em razão dessa campanha, seu nome entrou num franco declínio e o escritor admirado, à sua

[5] Eleito três vezes consecutivas, por unanimidade, por concurso realizado pelas revistas Fênix, Fon-Fon e O Malho, como Príncipe dos Prosadores Brasileiros, fazendo par com Olavo Bilac, o Príncipe dos Poetas Brasileiros (NETTO, P., 1958).

época, de repente viu-se questionado, renegado, o que contribuiu para não ser enquadrado nas coletâneas nacionais (NETTO, P., 1964).

Gurgel (2012, p. 1) em referência à condenação imposta ao escritor pelos modernistas, compreende que CN foi vítima de preconceito

> [...] de grande parcela da academia e da crítica literária, satisfeita no seu exercício de papaguear o que aprendeu neste ou naquele manual, mas raramente disposta a ler, com espírito despojado de ideologias, a produção dos autores.

Mendes e Ignácio (2010) ressaltam que a prosa de Coelho Netto apresenta características parnasianas, com um refinamento estilístico que busca sempre expor novos vocábulos: não costumava usar a mesma palavra em trechos pequenos. Destacam também “um anseio pela adjetivação numerosa, rara e erudita” (p. 2), sobretudo na obra *Turbilhão*. Essa postura acirrou ainda mais a crítica negativa sobre seu trabalho, pois os modernistas que defendiam e valorizavam uma linguagem mais simples, mais direta, acabaram por reprovar o que, para Coelho Netto, era uma prática muito comum.

Broca (1958, p. 23) sobre essa questão relata:

> A pobreza de expressão de grande parte dos nossos escritores principalmente romancistas, como vocabulário e sintaxe restrita, forjando com dificuldade seu instrumento verbal, explica a repulsa que o estilo de Coelho Netto, opulento e luxuriante, passou a despertar nestas últimas décadas.

Acusavam-no de “rebuscador de palavras difíceis, eternamente debruçado sobre os dicionários” (NETTO, P., 1958, p. XCIX). Seu filho, Paulo Coelho Netto, redarguiu sobre essas críticas dizendo:

> Queriam que um homem dotado de extraordinária cultura e assombrosa imaginação, possuindo o mais rico vocabulário da língua portuguesa, calculado em 20.000 palavras, escrevesse como alguns “gênios” lançados pelas catapultas modernistas, que tinham como mestres, na técnica de repetição, papagaios aliteratados (NETTO, P., 1958, p. XCIX-C).

Ainda sobre a verborragia apregoada a CN, seu filho (NETTO, P., 1964) narra um episódio no mínimo curioso. Em uma de suas conversas com Euclides da Cunha sobre verbos luminosos e ardentes da nossa língua, Coelho Netto começou a citar alguns exemplos: abrilhantar, aureolar, arcoirisar, acender, aclarar, adurir, assoleimar, aurifulgir, afuzilar, acalorar, abrasar, alumiar; e dezenas e dezenas de

outros. Ao final, anotaram nada menos que duzentos e dezoito vocábulos. Moraes (1976, p. 149) chamava-o de “malabarista do verbo”.

Coelho Netto viveu em uma época em que se produzia a literatura chamada de “sorriso da sociedade” (termo cunhado por Afrânio Peixoto), servilmente submissa a modelos europeus, desde os aspectos socioculturais até os intelectuais. Acabou por isso a aderir ao “modismo dominante das frases de efeito, da oratória vibrante e arrebatadora, dos torneios verbais onde muito acima do que se dizia estava a forma por que se dizia” (MORAES, 1976, p. 148).

Lima (1958, p. XXXV) afirma que CN era possuidor de um vernáculo portentoso, “ressuscitando um verbalismo inatual”, procurando usar sempre o termo preciso, exato: passou pela *belle-époque*, que foi o império do dicionário. Também chamou atenção ao fato de Coelho Netto mostrar em seus textos a sua “submissão ao culto do adjetivo”. Quanto ao uso demasiado dos adjetivos, foi considerado como um “libidinoso da adjetivação”, tamanha importância dada aos dicionários (MORAES, 1976, p. 149).

A impressão causada em muitos pela denegação da obra de CN, vale para lembrar que foram vítimas das mesmas acusações 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) lembra que os modernistas, que se apresentaram como renovadores da nossa cultura, ignoraram heroicamente tudo de grande que a literatura ocidental produziu em 1922.

Acusado de ser prolixo, CN assumiu essa característica ao dizer “precisar desbastar parte de sua obra da demasia de palavras” (BROCA, 1958, p. 24), mas os romances *Turbilhão*, *Miragem* e *O Morto* não careciam deste desbaste.

Niskier (2010, s.p.) comentando sobre a enfática campanha modernista a Coelho Netto, destaca que “o modernismo condenou-o como representante do passadismo, acusado de afetação, palavreado rebuscado e enfático, abuso de termos incomuns, prolixidade e helenismo”.

Duramente atingido pelas investidas críticas da geração de 22, “falava-se muito em descoelhonetizar o Brasil [...] anos depois, alguns se lançaram a uma aventura que bem poderia ser chamada de recoelhonetização” (MORAES, 1976, p. 149). A crítica a Coelho Netto foi mais benévola em Portugal, onde “ninguém se

espanta com a riqueza léxica de um Camilo, a exuberância de um Fialho, o preciosismo de um Aquilino Ribeiro" (BROCA, 1958, p. 24).

No término dos anos 30 e nos anos 40 é esboçada uma reação em favor de um julgamento mais objetivo da obra de Coelho Netto, em que se empenharam, mais particularmente, Otavio de Faria e Brito Broca. Um dos poucos e mais dedicados estudiosos que consagrou um estudo mais acurado da imensa trajetória do seu espólio literário foi Péricles Morais - ensaísta amazônico que nutria uma grande adoração pelo escritor, chegando aos extremos de uma admiração incondicional que só lhe permitia ver esplendores, mesmo onde houvesse o que censurar (LIMA, 1958).

Constituindo-se como um dos seus defensores, senão admiradores, Machado de Assis teceu elogios quase entusiásticos que é mister ressaltar. Referindo-se à *Miragem* disse: "Coelho Netto tem o dom da invenção, da composição, da descrição e da vida [...] é dos nossos primeiros romancistas e, geralmente falando, dos nossos primeiros escritores" (ASSIS, 1895, s.p.).

Brito Broca, que se preocupou com o Coelho Netto romancista, iniciando as primeiras linhas de revisão de sua obra de ficção, pode ser considerado o mais bem informado sobre o autor de *Treva*, sem extremismo, compreensivo e, acima de tudo, conclusivo em tudo o que se refere a ele. O próprio Broca (1958, p. 3) explicita: "A hostilidade que Coelho Netto vem encontrando nas gerações novas, de 1922 para cá, resulta, em grande parte, do fato de elas lhe desconhecerem a obra ou conhecerem-na de maneira bastante falha e superficial".

O crítico Wilson Martins e o escritor Guimarães Rosa (ocupante da cadeira de CN na ABL) rebateram as críticas dizendo que CN foi um grande injustiçado: "A vitória do modernismo se fez como se houvesse necessidade de abater um grande inimigo, no caso, Coelho Neto" (NISKIER, 2010, s.p.).

Campos (1945, p. 287) consagrou a Coelho Netto um estudo sobre sua obra. É com palavras de admiração e respeito que ele declara:

> O Sr. Coelho Netto não é, em verdade, apenas um escritor: é uma literatura. O estilista maneiroso, alinhador de períodos elegantes, é, não raro, fruto da paciência, do estudo, da tenacidade; o espírito criador, que tira do caos um mundo, esse, não se inventa, não se imita.

Durante 50 anos de atividades literárias e jornalísticas, CN exerceu influência na grande maioria dos escritores de seu tempo, sendo também, motivo de

orgulho para seu estado de nascimento, o Maranhão. Fazendo parte do mote de ilustres escritores da Atenas do Brasil (São Luís) que pairaram no cenário nacional, apesar de ter saído de sua terra, ainda em tenra idade, desta não se esqueceu, pois é comum em suas obras referências à sua terra, tanto no léxico quanto nas referências a pessoas, como no caso específico em *Firmo, o Vaqueiro*, quando menciona "o cafuzo, um codoense de fama" (NETTO, C., 1926c, p 127), ou então quando utiliza termos característicos da região, como expressões cristalizadas na fala do maranhense como: "hen... hen", "quá", consideradas como uma marca polissêmica desse falar (NERES, 2010).

Coelho Netto é pouco lido na atualidade. As novas gerações não o conhecem bem, em consequência, talvez, da terrível campanha de marginalização de sua obra posta em prática pelos artistas de 22. Broca (1958, p. 3) já destacava o fato de não fazerem jus ao talento e obra de CN ao dizer que:

> A produção demasiado vultosa desse escritor, num ambiente como o nosso, em que se lê tão pouco, não tem permitido, geralmente, aos mais curiosos, percorrer-lhe senão um pequeno número de livros, e sendo essa obra irregular, havendo nela do bom, do muito bom, do medíocre e do ruim, acontece, muitas vezes, o conhecimento incidir na parcela mais fraca, de onde os juízos levianos e falhos, embora quase sempre dogmáticos.

Toda a obra de Coelho Netto foi publicada pela Editora Lello, na cidade do Porto (Portugal), o que talvez tenha dificultado o acesso à sua obra pelo público brasileiro. Mas nem mesmo o fato de ela ter sido republicada pela Editora Aguilar (Ediouro) em 1958, nas décadas seguintes à morte do autor, em uma *Obra Seleta*, em três volumes, não o ajudou a ser lembrado, e também não colaborou para que ele figurasse nas antologias que, sequentemente, surgiram homenageando os célebres escritores brasileiros.

No catálogo atual da Ediouro Publicações, somente se encontram as obras *Turbilhão* e *Rei Negro* para venda. Fato que comprova que nem as principais obras do escritor estão disponíveis. A Biblioteca Virtual de Literatura[6] e o Domínio Público[7] disponibilizam alguns romances, contos, artigos, crônicas, críticas literárias e teatro. Estes *sites* disponibilizam praticamente as mesmas obras. O *Wikisource*, biblioteca digital coletiva de textos de domínio público, apesar de listar nove livros do

[6] Biblioteca Virtual de Literatura: ver textos disponíveis em http://www.biblio.com.br/
[7] Domínio Público: ver textos disponíveis em http://www.dominiopublico.gov.br/pesquisa/ResultadoPesquisaObraForm.do;jsessionid=7C2CD7C97DD8F9732FE2A33C4C563681

escritor, tem atualmente somente três livros com textos disponíveis, sendo dois romances (*A Conquista* e *Turbilhão*) e, mais recentemente, por contribuição nossa e como parte efetiva deste trabalho, os contos do livro *Sertão* escolhidos para compor o CCN.

Este é um dos fatos que justifica a compilação de textos de Coelho Netto, no sentido de também resgatar algumas de suas produções de mais destaque e que não estão disponíveis na WEB, como os contos de *Sertão*. Considerando-se o cenário atual de acesso às informações propagadas pelas tecnologias digitais, a disponibilização desses textos contribui sobremaneira para que eles possam ser acessados, lidos, analisados e apreciados, pois atualmente sua obra é praticamente ignorada pelo público brasileiro.

Sobre essa compilação, Alencar (2013c, p. 1) reforça, em sua avaliação, que:

> a análise linguística computacional dos textos de Coelho Netto, utilizando ferramentas disponíveis apenas recentemente para o português, permite capturar automaticamente certas propriedades formais a partir de um grande número de dados, o que dificilmente se poderia fazer manualmente.

Hoje, há uma movimentação intensa nos meios literários intentando resgatar Coelho Netto, para de fato discuti-lo e não condená-lo, como assim fizeram os modernistas. Respaldados nos avanços tecnológicos que permitem que ambientes como as redes sociais veiculem e socializem informações com rapidez e alcance imensuráveis, uma *timeline* do *Facebook*[8] serve, atualmente, como ponto de referência de dados sobre Coelho Netto. Criada recentemente (agosto de 2013) por um bisneto de Coelho Netto em função dos 150 anos de morte do escritor, neste espaço são divulgados documentos, artigos, fotografias e curiosidades da família (várias gerações), arquivos sobre a crítica e suas obras, e ilustrações de raridades/curiosidades como o sinete com seu brasão e um bilhete da loteria federal com a foto do escritor.

A compilação do corpus Coelho Netto é fundada nos aspectos destacados nessa seção, sobretudo pela vultosa produção literária e a crítica feita à sua obra.

[8] Timeline Coelho Netto no Facebook: https://www.facebook.com/familiacoelhonetto?fref=ts

2.1.3 Obras que compõem o Corpus Coelho Netto

2.1.3.1 A Conquista

Este romance-documento surgiu em 1899. Com forte carga autobiográfica, é a memória da juventude boêmia de Coelho Netto, que coincidiu com as lutas finais da Abolição e da Proclamação da República. É uma obra valiosa como documentário da *belle-époque* carioca – período de proeminência de pluralidade de tendências filosóficas, científicas, sociais e literárias. É enredado num discreto realismo, e considerado o romance mais lido do escritor (NETTO, P., 1958).

Lima (1958, p. LXVII) diz ser este romance “um livro precioso para quem quiser reconstituir a vida literária dos fins do século XIX”. O livro teve quatro edições (MORAES, 1976).

Mendes e Ignácio (2010, p. 3) reforçam a análise desta obra:

> *A Conquista* é predominantemente imaginativa – excetuando-se sua trilogia, escrita nos moldes de uma obra memorialista e, por isso, baseando-se nas experiências do autor na condição de ser humano engajado na causa republicana e abolicionista, na imprensa carioca ou, mesmo, como homem de letras que viveu no período compreendido pela *Belle Époque.*

A obra expõe aspectos da sociedade brasileira do final do século XIX, mas é principalmente a personificação da vida do autor: a visão intelectual, o trabalho e a remuneração do escritor, e a militância política. Os personagens dos livros são alusões a elementos reais da cena carioca, para os quais o escritor criou vários pseudônimos. José do Patrocínio aparece no romance como ele próprio.

É a história de um grupo de rapazes boêmios ligados à literatura que viviam como nômades, morando em ruas diferentes, comendo em distintos restaurantes, adequando-se às condições adversas da falta de dinheiro. Assim, o principal tema do livro é o questionamento da situação econômica do produtor cultural naquela época. O protagonista Anselmo, estudante e candidato a escritor (Coelho Netto), e Ruy Vaz romancista naturalista (Aluísio Azevedo) se encontram com outros tipos importantes para o romance: o polêmico poeta Paulo Neiva (Paula Ney); Victorino Motta; Duarte, o romântico; Octávio Bivar (Olavo Bilac), Luís Moraes (Luís Murat) e Lins (PINHO, 2009).

Em um dos capítulos de *A Conquista*, Coelho Netto dá destaque à capoeira, marginalizada na época e proibida por lei de 1890 a 1937: era uma grande paixão do escritor. Era um exímio capoeirista (JORGE, 1999), e a esse esporte o escritor dedicou uma crônica exclusiva ao tema, no seu livro Bazar, em 1922: *Nosso Jogo*[9].

2.1.3.2 Turbilhão

Utilizando um estilo depurado e correntio, Coelho Netto publicou *Turbilhão* criando figuras humanas, daquelas que se incorporam em definitivo tanto à literatura quanto à memória do leitor: romance com forte apelo sensual. Neste, o escritor conseguiu enlaçar a fantasia com a realidade, ou o Romantismo com o Realismo, tendo-se tornado uma de suas obras mais bem realizadas (LIMA, 1958).

Composto de tipos marginais, como o estudante Paulo, a irmã Violante, a velha mãe Júlia e o mulato Mamede, a obra assim é resumida:

> [...] está centrada na trajetória de Paulo, órfão de pai, estudante de medicina, de baixa condição financeira. Sendo, como já se afirmou, órfão de pai, e morando com a única irmã, Violante, e a mãe, D. Júlia, esta sempre às voltas com achaques e doenças, o moço fica chocado, indignado logo no começo da narrativa, tão logo toma conhecimento do fato de que a irmã havia fugido de casa. Para os padrões morais e sociais da época, isso significava o fim da reputação de uma donzela, já que tal ato, sem uma imediata reparação – leia-se: casamento – faria com que a moça fujona ficasse relegada, fatalmente, à condição de uma meretriz (MENDES; IGNÁCIO, 2010, p. 5).

Tem como pano de fundo e alude ficcionalmente “às mudanças que se verificavam no modo de vida da população do Rio de Janeiro na primeira década do século XX” (MENDES; IGNÁCIO, 2010, p. 4).

A crítica de Broca (1958, p. 19) sobre a obra, elogia: “*O Turbilhão*, publicado em 1906, assinala o ponto culminante dessa carreira tão cheia de altos e baixos [...] Só esse livro, parece-me, bastaria para dar a Coelho Netto um lugar de destaque no ficcionismo brasileiro”. E que é “indiscutivelmente uma das obras mais felizes, se não a mais feliz, do fecundo escritor” (p. 1). Lúcia Miguel Pereira, crítica

[9] Nessa crônica o escritor dá significados a vários termos para os movimentos que até hoje são utilizados na capoeira, fazendo parte da Nomenclatura Histórica de Movimentos. Texto disponível em: http://www.capoeira.jex.com.br/lit+classica/capoeira+-+o+nosso+jogo+

ferrenha da obra de CN, sequer fez referência à essa obra. Foi editado três vezes (MORAES, 1976).

Mesmo com todo o sucesso que esse romance teve quando de sua publicação, como outras obras de menos relevância nas produções de Coelho Netto, *Turbilhão* caiu no ostracismo. A Fundação Casa de Rui Barbosa, e seu setor de Filologia, propôs-se a resgatar obras que, na atualidade, encontram-se esquecidas do grande público, mas “ainda não se debruçaram, até o presente momento, sobre esse romance de Coelho Netto” (MENDES; IGNÁCIO, 2010, p. 4).

2.1.3.3 Sertão

Com sangue indígena nas veias e nascido no interior do Maranhão, conhecendo de perto a selva amazônica e outros recantos sertanejos do Brasil, CN decidiu pôr no papel toda essa experiência. O livro de contos *Sertão* tem histórias regionalistas que lhe conferem um lugar de destaque na evolução desse estilo de narrativa. O autor ressuscita nesta obra situações que fazem parte das crenças cultivadas no seio da mata brasileira. Como o próprio contista referencia em uma entrevista dada a João do Rio (JOÃO DO RIO, 1904, p. 18):

> Para a minha formação literária [...] não contribuíram autores, contribuíram pessoas. Até hoje sofro a influência do primeiro período da minha vida no sertão. Foram as histórias, as lendas, os contos ouvidos em criança [...] Nunca mais essa mistura de ideias e de raças deixou de predominar, e até hoje se faz sentir no meu ecletismo. A minha fantasia é o resultado da alma dos negros, dos caboclos e dos brancos.

Sertão, publicado em 1896, contém sete contos: *Praga; O enterro; A tapera; Firmo, o Vaqueiro; Cega, Mandovi* e *Os Velhos*. É uma coletânea dedicada a Paulo Prado, que foi um importante patrocinador e incentivador da Semana de Arte Moderna. O livro teve seis edições (MORAES, 1976).

Os contos deste livro são em geral quadros sertanejos idealizados. Ninguém, ao lê-los, pode dizer exatamente onde se desenvolvem. Mas apesar de os críticos da época dizerem que nada do que Coelho Netto escrevia era produzido de imagens reais, seu filho o defende dizendo que na época da publicação de *Sertão*, Coelho Netto passara dois meses em um longínquo vilarejo com grande selva em frente, restabelecendo-se de uma doença de infância, e isso contribui para inspirá-lo em sua produção (NETTO, P., 1958).

O livro destaca o amor de Coelho Netto pelo sertão, que retrata em cada conto o sertanejo, a natureza, seus costumes, seus medos e seus falares. Como disse Assis (1897, p. 1), é próprio do autor maranhense o senso da vida exterior:

> Coelho Neto ama o sertão, como já amou o Oriente, e tem na palheta as cores próprias de cada paisagem. Possui o senso da vida exterior. Dá-nos a floresta, com os seus rumores e silêncios, com os seus bichos e rios, e pinta-nos um caboclo que, por menos que os olhos estejam acostumados a ele, reconhecerão que é um caboclo.

A escolha por esse livro de contos leva em consideração, sobretudo, a sua popularidade, pois inicialmente foram publicados em jornais. Essa popularidade levou a ser um dos livros com mais edições de CN.

Do livro *Sertão* foram escolhidos para compor o *Corpus* Coelho Netto os contos: *O enterro; Firmo, o vaqueiro* e *Mandovi*.

O enterro é um conto que descreve o momento do enterro da velha cabocla Teçai, de 70 anos, descendente dos índios goitacazes. Velha temida e respeitada do lugar chamado Taba de Itamina, pelas suas pragas e malefícios e pelo terror da lenda que se criara ao seu redor: Teçai era a alma pagã de Tagiira, índia morta por Tupã ao trocar seu primeiro beijo e prestes a entregar sua virgindade a um aventureiro branco (NETTO, C., 1926a).

Firmo, o Vaqueiro é considerado uma das obras-primas do gênero na evolução da Literatura Brasileira, como afirma BROCA (1958, p. 6) ao dizer que o conto “é uma página indispensável a qualquer antologia de contos brasileiros”. Tem como personagem principal o velho Firmo, caboclo de 80 anos, vaqueiro, musculoso e rijo, grandes olhos negros e de cabelos longos e cacheados (NETTO, 1926b). A história é narrada pelo “patrãosinho” (não nominado no livro) que tinha Firmo como seu companheiro quando ia passar as férias na roça. Faz um breve retrato da história do vaqueiro e suas aventuras quando novo. Aos 80 anos, o vaqueiro vivia de recordações e abatido pelo reumatismo. No Natal o velho adoece, com muita febre, mal podendo se mover e passava os dias na rede. Para alentá-lo, o patrãozinho vai visitá-lo, levando consigo uma visita: Raimundinho, o sobrinho do vaqueiro. Firmo passa então seus últimos momentos da vida fazendo versos e cantando com o sobrinho e o patrãozinho. No outro dia, às 4 da manhã, o vaqueiro já não responde ao chamado do sobrinho que “atira-lhe um verso”, e Firmo não responde. Estava

morto. “Tio Firmo, mesmo velho e doente, não era homem de deixar um verso no chão” (NETTO, C., 1926c, p. 129), declara o sobrinho ao patrãozinho.

O conto *Mandovi* tem como personagens: Mandovi “caboclo de peito largo, com uma barba crespa, negra e densa... gosava (*sic*) fama de valente e ninguém ousava enfrentar elle (*sic*)” (NETTO, C., 1926c, p. 211); Tigre, o cachorro preto; e os vaqueiros companheiros de lida de Mandovi. A história se inicia na venda de Manuel Monte, com uma reunião dos vaqueiros. O conto todo gira em torno do retorno de Mandovi à sua casa (no Serrinha) acompanhado de seu cachorro Tigre, atravessando a mata em uma noite de lua cheia. No percurso escuta um grito chamando por seu nome e acredita ser assombração, ficam com medo tanto ele quanto Tigre. Mais adiante, encontra outros vaqueiros, e pergunta do que se tratava o grito. Na verdade, era uma ave da mata de grito estridente, e que o vulto branco era uma folha velha de palmeira despencada. No entanto, um dos vaqueiros lembra-se de um italiano que apareceu morto na beira do rio. Por conta disso, Mandovi não mais se convence que o vulto branco não era a folha de palmeira, mas sim que era o italiano e chega em casa apressando e acreditando nessa versão.

A compilação do *Corpus* Coelho Netto tem como fundamentação as teorias da Linguística Computacional e da Linguística de Corpus, detalhadas nas seções a seguir.

2.2 Linguística Computacional

Outrora, a Linguística Computacional (LCOMP) era denominada por uma variedade de termos: além de Linguística Computacional, era conhecida também como Linguística Matemática, Linguística Estatística, e Mecanolinguística, disciplinas que realizavam pesquisas iguais ou de tipos conexos, sendo que o único elemento comum a essas áreas era o uso das técnicas algorítmicas do computador (TOSH, 1972).

Até meados de 1960, a Linguística Computacional[10] centrou-se exclusivamente nos estudos das linguagens formais e das linguagens de programação. Com o estímulo da Linguística e a influência da Filosofia da Linguagem e da Psicologia, a LCOMP passou a abordar outras matrizes: morfologia,

[10] O nome foi cunhado por David Hays, em 1967 (DIAS-DA-SILVA, 2006).

sintaxe, semântica, pragmática, discurso, texto, aquisição de linguagem, entre outros (DIAS-DA-SILVA, 2006).

A LCOMP é a área que investiga o tratamento computacional da linguagem e das línguas naturais, subdividida, conforme Othero e Menuzzi (2005) em duas subáreas: a Linguística de *Corpus* e o Processamento de Linguagem Natural (PLN).

Em linhas gerais, a Linguística de *Corpus* é o estudo da língua por meio da exploração e análise de *corpora* eletrônicos (robustos bancos de dados que contém amostras de linguagem natural) dos mais variados tipos: *corpora* de linguagem falada, *corpora* de linguagem literária, *corpora* com textos de jornal, *corpora* compostos exclusivamente por falas de crianças em estágio de desenvolvimento linguístico etc, e vem se dedicando também “à descrição, à formalização e à emulação computacional das habilidades linguísticas dos falantes” (OTHERO; MARTINS, 2011, p. 100).

Já o PLN estuda a linguagem voltando-se para a construção e desenvolvimento de ferramentas computacionais (DOMINGUES; FAVERO; MEDEIROS, 2008), em especial os tradutores automáticos, *chatterbots (*programas que interagem com humanos através de diálogos em linguagem natural, na modalidade escrita*), parsers* (geradores automáticos, principalmente, de resumos, e corretores ortográficos e gramaticais)*,* reconhecedores automáticos de resumos dentre outros (OTHERO; MENUZZI, 2005). O planejamento, a construção e o desenvolvimento dessas ferramentas computacionais, também chamadas de “recursos linguístico-computacionais” por Di Felippo e Dias-da-Silva (2009, p. 187) constituem-se em tarefas nada triviais. Por vezes, a LCOMP é considerada sinônima do PLN (BERBER SARDINHA, 2005, DI FELIPPO; DIAS-DA-SILVA, 2009).

Sobre o PLN, Di Felippo e Dias-da-Silva (2009, p. 187) dizem que:

> As pesquisas nessa área, ao mesmo tempo em que se beneficiam com os estudos provenientes da Linguística, têm propiciado não só desenvolvimento de tecnologias ou recursos aplicáveis a várias atividades, mas também o próprio desenvolvimento da Linguística e da Ciência da Computação, duas das várias disciplinas matrizes do PLN.

Do ponto de vista linguístico, as pesquisas do PLN estão concentradas em cinco níveis de análise: a) fonético ou fonológico; b) morfológico; c) sintático; d) semântico ou e) pragmático (VIEIRA; LOPES, 2010). O dinamismo da língua

constitui-se um enorme desafio para o PLN, que é um passo fundamental para aproximar a comunicação entre humanos e computadores.

A *Association for Computational Linguistics* (ACL)[11], uma sociedade científica internacional e profissional que envolve pessoas que pesquisam problemas que envolvem a linguagem natural e computação, define a LCOMP como um estudo científico da linguagem a partir de uma perspectiva computacional. Motivada por estudos científicos, tenta explicar um fenômeno linguístico ou psicolinguístico por meio computacional. Em outros momentos, tem uma motivação puramente tecnológica, buscando proporcionar um ambiente de trabalho componente de um discurso ou sistema de linguagem natural. A LCOMP está incorporada em vários sistemas operantes do processamento da linguagem como: reconhecimento de fala, sintetizadores de fala, sistemas de resposta de voz, motores de busca da *WEB*, editores de textos, tradutores *online*, banco de dados de ensino de línguas, dentre outros.

A LCOMP comporta "trabalhos que privilegiem o estudo da linguagem humana, escrita e falada, por meio de criação de programas de computador específicos" (BERBER SARDINHA, 2005a, p. 25). Do ponto de vista de pesquisador, o linguista pode se preocupar em ver a linguística como aplicação de teorias linguísticas em máquinas (MCENERY; WILSON, 2001).

Considerando-se a forma bastante polivalente das acepções teóricas sobre a atuação da LCOMP, Alencar (2009, p. 135) resume algumas delas:

> · LC1: *linguística computacional* como subdisciplina linguística voltada para os aspectos algorítmicos das línguas naturais e do processamento da linguagem natural;
> · LC2: *linguística computacional* como área de investigação direcionada ao desenvolvimento de ferramentas computacionais para a pesquisa linguística e para o processamento de dados linguísticos [...];
> · LC3: *linguística computacional* como implementação de fenômenos da linguagem no computador, área conhecida também como processamento da linguagem natural (PLN), intimamente relacionada às áreas da inteligência artificial e da ciência da cognição de modo geral;
> · LC4: *linguística computacional* como ciência aplicada, voltada para o desenvolvimento de aplicativos para tradução automática, correção ortográfica e gramatical etc., constituindo um ramo da engenharia de software, como sugerem as designações em inglês *grammar engineering* ou *language technology*, respectivamente engenharia da gramática e tecnologia da linguagem.

O advento da LCOMP está vinculado à evolução nas pesquisas

[11] ACL - http://www.aclweb.org/

linguísticas e, sobretudo ao avanço das tecnologias, quando o computador passou a ser uma ferramenta essencial para estudos das áreas de humanas e da linguística. O desenvolvimento de recursos computacionais com contribuições da Inteligência Artificial fez com que a Linguística Computacional se concretizasse enquanto uma área que se ocupa da tecnologia linguística necessária para o processo computacional da linguagem, sendo utilizada para várias aplicações.

Na década de 50, os estruturalistas norte-americanos constituíram o primeiro grande *corpus* objetivado à análise linguística: o *Brown*, contendo um milhão de palavras do inglês americano (OLIVEIRA, 2009; BERBER SARDINHA, 2000). Sua importância dá-se pelo fato primordial de que nessa época não existiam textos escritos em formato digital. Os *corpora* não computadorizados inspiraram os *corpora* atuais, caso do SEU[12] (*Survey of English Usage*) que foi compilado por Randolf Quirk e sua equipe, em Londres, a partir de 1953, e foi planejado para conter um milhão de palavras: o *corpus* Brown usou o SEU como referência (BERBER SARDINHA, 2000, 2004).

Dentre os *corpora* que servem como marco de referência histórica temos: *Brown*, *Birminghan*, e BNC: o primeiro em inglês americano e os dois últimos em inglês britânico, contendo respectivamente 1 milhão, 20 milhões e 100 milhões de palavras. Na Grã-Bretanha, um dos centros mais desenvolvidos na área da LCOMP, diversas universidades dedicam-se à pesquisa baseada em *corpus* para a descrição de vários aspectos da linguagem, tanto para a teorização quanto para a criação de *corpora* e de materiais de apoio a diversas áreas (BERBER SARDINHA, 2000).

Biderman (2001) considera que o pioneirismo da LCOMP deu-se nas décadas de sessenta e setenta, com seus processos primários vinculados à Tradução Automática, e para isso alguns linguistas e pequenos centros nos EUA e na Europa se apoiaram nos estudos do PLN via computador. De ação pretensamente utópica, considerando-se o cenário informático da época, a tradução automática converteu-se em realidade, mesmo que em passos lentos e investida de bastante complexidade.

Nos primórdios da LCOMP os dados mais procurados em um *corpus*

[12] O SEU "foi organizado em fichas de papel, cada uma contendo uma palavra do *corpus* inserida em 17 linhas de texto. As palavras foram analisadas gramaticalmente, com cada ficha recebendo uma categoria gramatical. O conjunto de categorias resultante serviu de base para o desenvolvimento dos etiquetadores computadorizados contemporâneos, que fazem a identificação de traços gramaticais automaticamente" (BERBER SARDINHA, 2000, p. 326).

eram as concordâncias de texto ou KWIC (*Key-Word-In-Context*). Para encontrá-las nos textos, primordialmente nos estudos bíblicos clássicos, a busca era feita manualmente por meticulosos estudiosos: o computador então mudou essa situação. Surgiu o primeiro *software* representativo das atividades da LCOMP: O FOLIO VIEWS, que além de fornecer as concordâncias de texto também mostrava a frequência das palavras (BIDERMAN, 2001).

Os resultados obtidos pelos linguistas computacionais avançam na velocidade dos avanços tecnológicos, pois cada vez mais se torna factível a criação de produtos tecnológicos capazes de dominar a linguagem natural de um ser humano (OTHERO; MENUZZI, 2005).

Esse início, mundialmente, constituiu-se em um divisor de águas nos estudos e pesquisa em PLN, subsidiando exponencialmente a atuação da LCOMP nos estudos linguísticos. A morfologia, a sintaxe, a lexicografia, a terminologia e a tradução se beneficiam do processamento computacional para seus estudos e análises (BIDERMAN, 2001). Criaram-se enormes bancos de dados, instituídos como *corpus* que se destinavam a testar hipóteses e/ou fornecer evidências na pesquisa linguística.

É importante destacar que a Linguística de *Corpus* foi sistematizada enquanto área de estudo da Linguística quando do surgimento do *corpus* computadorizado.

Os *corpora* linguísticos têm por objetivo além de produzir *softwares* e desenvolvimento e aperfeiçoamento de programas (ação específica da LCOPM), também serve para o estudo dos fenômenos linguísticos (especificidade da Linguística de *Corpus*).

Considerando-se a similaridade de ações entre LCOMP e PLN, a Linguística de *Corpus* está mais afinada com a proposta deste trabalho, pois se trata de compilação de um *corpus*.

Detalharemos, portanto, a Linguística de *Corpus*, expondo os conceitos que os diversos teóricos dão a essa área, seu histórico e as principais características e apontando as diversas áreas de aplicação.

2.3 Linguística de Corpus

A Linguística de *Corpus,* doravante também LCORP, requer um maior detalhamento teórico nesse trabalho, pois em aspectos científicos é a área que mais desenvolveu pesquisas vinculadas à Linguística, produzindo e veiculando resultados que modificaram diversos aspectos no trabalhar com a linguística, tanto constituindo *corpora* quanto mostrando resultados encontrados com as análises desses recursos.

Noam Chomsky, em 1957, lançou o *Syntactic Structures*, obra que influenciaria sobremaneira uma mudança de paradigma na linguística: o empirismo saía de cena e surgia a sustentação para os trabalhos baseados em *corpora,* mudando o paradigma linguístico vigente (BERBER SARDINHA, 2000).

> [...] a partir desta obra de Chomsky, os dados necessários para o linguista estavam em sua mente e eram acessíveis por meio da introspecção. Não havia necessidade de coletar-se dados abundantes de terceiros. Estes serviriam apenas para o estudo do desempenho, quando todos sabiam que o que interessava era a investigação da competência linguística (BERBER SARDINHA, 2000, p. 324).

Nessa época, o processamento manual de *corpora* realizado para a compilação do *Corpus* Brown era motivo de várias críticas, sendo a mais contundente o fato de se criar *corpora* gigantescos, como o criado por Thordinke e Gates, nos anos 20, contendo 18 milhões de palavras retiradas de textos escolares (ESCUDERO, 1983 *apud* ARRUFAT, 1997), todo feito de forma manual e, por isso não sendo totalmente confiável, considerando-se sua extensão.

Outro *corpus,* este sobre a ortografia do alemão, organizados pelo alemão J. Käding, consumiu a mão de obra de 5.000 analistas e continha 11 milhões de palavras. Os críticos diziam que “o ser humano não é talhado para tarefas deste tipo” (BERBER SARDINHA, 2000, p. 327), ficava claro, portanto, que esforços como esses precisavam ser incrementados com instrumentos que permitissem a análise de grandes quantidades de dados.

A Linguística de *Corpus* é uma área interdisciplinar que teve seu início na década de 60, quando surgiu o interesse em aplicar os estudos de *corpus* nas ciências da linguagem. O termo em si só surgiu na década de 80, por Geoffrey Leech (ele dizia que o termo “*corpus linguistics*” só aparecia vez por outra) ao

publicar, em 1992, o livro *Corpus Linguistics: Recent developments in the use of computer corpora in English language research* (ZYNGIER, 2004).

Os anos de 1960 foram importantes, porque, nessa época, surgiram os computadores *mainframes* (BIDERMAN, 2001; BERBER SARDINHA, 2004) que davam suporte informático a centros de pesquisas universitários, e passaram então a serem aproveitados para a pesquisa em linguagem. Os anos 80, com a chegada dos microcomputadores pessoais (PCs), foi o período decisivo para o reaparecimento e fortalecimento das pesquisas baseadas em *corpus* (BERBER SARDINHA, 2000).

No Brasil, nos anos 1990, e com a proliferação dos computadores pessoais, dos textos em formato digital e do desenvolvimento e acessibilidade de ferramentas que analisam *corpora*, como o *Word SmithTools*, foi que a LCORP pôde finalmente, começar a se desenvolver de fato (FRANKENBERG-GARCIA, 2012).

Definindo a LCORP, Berber Sardinha (2004, p. 03) diz:

> [...] ocupa-se da coleta e da exploração de *corpora*, ou conjuntos de dados linguísticos textuais que foram coletados criteriosamente, com o propósito de servirem para a pesquisa de uma língua ou variedade linguística. Como tal, dedica-se à exploração da linguagem por meio de evidências empíricas, extraídas por computador.

Oliveira (2009, p. 52) entende que:

> A Linguística de *Corpus* situa-se na interdisciplinaridade e na complementaridade, relacionando-se com outras áreas do conhecimento, teorias ou abordagens linguísticas, que ao somarem conhecimentos, poderão contribuir para um melhor conhecimento do seu objeto comum de estudo que é a linguagem.

A LCORP apresenta-se como uma nova metodologia, no sentido de utilizar textos naturais e ferramentas informáticas para descrever a língua, mas também “como uma nova disciplina, que dá uma nova abordagem à descrição linguística” (FRANKENBERG-GARCIA, 2012, p. 12). Portanto, está condicionada diretamente à tecnologia, pois sua história está vinculada à disponibilidade de ferramentas computacionais para análise de *corpus* (BERBER SARDINHA, 2000).

Leech (2004) diz que a LCORP é uma ciência empírica, na qual o investigador procura identificar padrões de comportamento linguístico por inspeção e análise de amostras de linguagem natural.

Dentre os campos de investigação que a Linguística de *Corpus* atua, citamos: análise textual; ensino-aprendizagem de línguas; a lexicografia; a análise do discurso histórico, político e jornalístico; os estudos literários; os estudos de tradução; a sociolinguística, e principalmente o desenvolvimento de novas ferramentas de PLN. Estas ferramentas, Frankenberg-Garcia (2012) entende que possibilitaram uma leitura vertical do texto com uma consequente visão de padrões de uso da língua sem precedentes, pondo em questão certos pressupostos linguísticos nunca antes contestados, como bem enfatiza Tognini Bonelli (2010, p. 17):

> [...] a quantidade de dados processados graças à ajuda do computador, acabou por ser uma revolução teórica e qualitativa, na medida em que oferece percepções sobre linguagem que abalaram os pressupostos subjacentes a muitas posições teóricas estabelecidas neste campo (tradução nossa).[13]

A Linguística de *Corpus* surge como referência para o estudo dos grandes *corpora* que surgiram*:* "o *corpus* veio possibilitar um confronto entre a teoria e os dados empíricos da língua" (BIDERMAN, 2001, p. 78). A LCORP segue uma abordagem empirista da linguagem, que significa dar "primazia aos dados, em geral reunidos, sob a forma de um *corpus*" (BERBER SARDINHA, 2004, p. 30), e que compreende um sistema probabilístico (probabilidade dos sistemas linguísticos em descrever dados de acordo com os contextos em que os falantes os empregam), cuja linha de pesquisa está vinculada a Halliday (BERBER SARDINHA, 2000) e a Sinclair (1995, 2005), destacando que uns traços são mais frequentes que os outros e assim encaixando-se nos preceitos teóricos da LCORP.

Bennett (1998) diz que mais do que entender o que é LCORP é preciso entender o que não é LCORP:

- Ela não é capaz de fornecer evidência negativa (um *corpus* não pode dizer o que é possível ou correto, ou impossível ou correto. Só pode nos fornecer o que tem ou não no *corpus*);
- Ela não é capaz de explicar o porquê (não explica porque os dados do *corpus* aparecem daquele jeito, só nos fornece os dados);

[13] "(…) the quantity of data processed thanks to the aid of the computer) has turned out to be a theoretical and qualitative revolution in that it has offered insights into language that have shaken the underlying assumptions behind many well-established theoretical positions in the field".

- Ela não é capaz de fornecer todas as informações possíveis ao mesmo tempo, de todas as linguagens (o tamanho, a representatividade, e o objetivo do *corpus* é quem define isso).

Porquanto, as pesquisas que utilizam a Linguística de *Corpus* compartilham algumas características: são empíricas e analisam os padrões reais de uso dos textos naturais; utilizam coletâneas grandes[14] e criteriosas de textos naturais; fazem uso extensivo de computadores para a realização das análises; e dependem de técnicas quantitativas e qualitativas (BERBER SARDINHA, 2000).

Resumindo a atuação da LCORP, ela envereda pelas áreas da Lexicografia, de Tradução, da Linguística Aplicada e do Processamento de Língua Natural, e as principais ênfases são para: a) Compilação de *corpora*; b) Desenvolvimento de ferramentas para análise de *corpora*; c) Descrição de linguagem; d) Exploração do uso de descrições baseadas em *corpora* para várias aplicações tal como ensino-aprendizagem de línguas, processamento de linguagem natural por máquinas, reconhecimento de voz e tradução (KENNEDY, 1998).

A convergência de pesquisas entre a Linguística de *Corpus* e a Linguística Aplicada (LA) vem se tornando cada vez mais rica, pois a LCORP está progressivamente colaborando para responder a muitas questões que ocupam as discussões dos pesquisadores da LA. E essa relação vem sendo enfatizada de maneira recorrente por linguistas aplicados (OLIVEIRA, 2009, p. 53), considerando que a LA, que é uma linguística baseada por ocorrências, está sempre em busca de novos dados.

Este trabalho tem, portanto, uma interface vinculada à área de Linguística Aplicada, mas também à área da Descrição Linguística baseada em *corpora,* como corrobora Rocha (2000, p. 239): “as investigações baseadas em *corpus* recai sobre a descrição linguística”, sobretudo ao se pesquisar sobre o léxico e a gramática a partir de análises descritivas de um *corpus* (BERBER SARDINHA, 2000)

No campo da Literatura, a disponibilização de grandes quantidades de dados compilados pelos linguistas de *corpus* evita que muitas das generalizações, que antes eram feitas por suposições intuitivas, sejam mais conclusivas. Isso acabou com uma das maiores críticas a estudos baseados em dados reais, que era a falta de

[14] Quanto a esse item, é motivo de discussão de HUNSTON (2002), SARMENTO (2008) e SHEPHERD (2012) que têm uma percepção diferenciada quanto à importância do tamanho do *corpus*.

confiança na análise manual. Quanto a essa influência positiva, ZYNGIER (2011, p. 102) ressalta:

> A Linguística de *Corpus*, portanto, pode fornecer ao crítico literário as ferramentas e os princípios necessários para que a literatura seja analisada sob uma nova perspectiva, mais objetiva do que a que tradicionalmente caracteriza a análise de base hermenêutica.

Sobre a Linguística de *Corpus* e o uso de textos literários, Zyngier (2004, p. 43) explica:

> A linguística de *corpus* pode mostrar o contexto de certos itens lexicais utilizados por um determinado autor e, desta forma, permitir afirmações fundamentadas sobre possíveis interpretações. Pode-se, também, comparar o uso que um autor faz de certas palavras a um conjunto de textos contemporâneos demonstrando, assim, o grau de criatividade. Textos do século XX não oferecem problemas nessa direção, já que se tem acesso a uma variedade de materiais escritos.

Alguns estudos como os de Abreu (2007), de Pellegrini e Petkoff (2009), e de Freitas (2007) tendo como investigação textos de Machado de Assis utilizando a Linguística de *Corpus*, reforçam a importância de se utilizar textos literários tratados computacionalmente para estudos linguísticos.

No caso do estudo do léxico, podemos diferenciar as palavras de maior e menor frequência (*hapax legomena*) e fazer comparações com *corpus* especializado de outros escritores do mesmo período histórico, ou então com textos contemporâneos. Em qualquer *corpus*, as formas de frequência de *hapaxes* são a maioria, porque “a maioria das palavras de uma língua é composta de palavras de ocorrem poucas vezes” (BERBER SARDINHA, 2000, p. 344).

Portanto, dentre as particularidades da LCORP é que ela se constitui em “uma área essencialmente aplicada”, destinando-se a coletar *corpora*, usando programas já existentes e aperfeiçoando-os. Ressaltando que os *corpora* são “a própria mola propulsora da área” (BERBER SARDINHA, 2005a, p.27-28).

Há um forte debate entre os linguistas tradicionais e os linguistas de *corpus* quando da compreensão de que esta área contribua significativamente para a linguística em geral. Exemplo desse debate, quiçá embate, tem-se nos EUA, que como observa Berber Sardinha (2000), a força da linguística gerativo-transformacional nos departamentos de linguística conflita naturalmente com a LCORP, e por isso esta não é tão desenvolvida fora da Europa. Outra referência é o

fato de que um dos maiores expoentes da LCORP mundial, o americano Douglas Biber, é integrante de um departamento de inglês.

2.3.1 Corpus na LCORP

A LCORP tem como sua base de pesquisa os bancos de dados gerados com textos acessíveis para o trato computacional: denominados de *corpus.* A concepção de *corpus* até então adotada na Linguística mudou com o desenvolvimento da LCORP: o universo linguístico-computacional passou a ter outra representação O termo passa a ter definições distintas, uma da Linguística e outra da Linguística de *Corpus*.

De forma bastante generalizada, *corpus* para a Linguística sempre foi utilizado como um recurso de análise e "não pode ser considerado como constituindo a língua, mas somente como uma amostra da língua. O *corpus* deve ser representativo, i. é., deve ilustrar toda a gama das características estruturais" (DUBOIS *et al*, 1993, p. 158-159). É, pois, um conjunto de textos escritos ou orais disponível para análise.

Para a LCORP, um *corpus* é uma coleção de partes de textos disponíveis em formato eletrônico, selecionados de acordo com critérios externos representativos, na medida do possível, de uma variedade de língua ou linguagem como fonte de dados para a investigação linguística (SINCLAIR, 2005). Sinclair é considerado o maior linguista de *corpus* da história pelo seu pioneirismo ao trabalhar na área do léxico com o dicionário COBUILD, de 1987, compilado a partir de um *corpus* computadorizado (ALUÍSIO; ALMEIDA, 2006).

Para Biderman (2001, p. 79), um *corpus* linguístico computadorizado é "uma coletânea de textos selecionados segundo critérios linguísticos, codificados de modo padronizado e homogêneo [...] um *corpus* constitui um conjunto homogêneo de amostras da língua de qualquer tipo (orais, escritos, literários, coloquiais etc.)". Ademais, o *corpus* pode ser homogêneo ou heterogêneo, desde que propiciem ampliar o conhecimento das estruturas linguísticas da língua representada por elas.

Santos (2008) compreende que um *corpus* é uma coleção de objetos para uso em Processamento de Linguagem Natural/Linguística Computacional/Linguística para medição, teste ou avaliação. Esses objetos são:

> textos, frases, palavras, entrevistas, erros ortográficos, entradas de dicionário, citações, pareceres jurídicos, filmes, imagens com legendas, traduções, correções (de textos de alunos de língua ou de tradução), telefonemas [...] usa-se um corpo para medir um dado fenômeno" (p. 45-51).

Para Shepherd (2012, p. 27):

> As parcerias e interfaces da Linguística de *Corpus* com a área de Programação de Linguagem Natural têm sido inúmeras e já produzem tecnologias para dicionários, analisadores sintáticos e morfológicos, corretores ortográficos, interfaces em língua natural e ferramentas voltadas ao ensino.

Porquanto, o *corpus* linguístico da LCORP está associado à informática, pois é quase improvável, atualmente, se criar um banco imenso de dados que seja manipulável para estudos, se não por meio de um computador.

Percebe-se então que o que diferencia as concepções de *corpus* nas duas áreas de estudos é no que se refere ao seu formato: o *corpus* da LCORP deve estar em formato eletrônico (ALUÍSIO; ALMEIDA, 2006). Doravante, o termo *corpus* neste trabalho se referirá a *corpus* eletrônico.

Um *corpus* tem especificações quanto ao tipo de textos que ele compõe, considerando-se para isso o tamanho, propósito e forma como foram compilados. Dentre os tipos de *corpus* definidos por Sinclair (1995), Hunston (2002), Tagnin (2010), Berber Sardinha (2000), caracterizam-se pelo:

- Modo: falado e escrito;
- Tempo: sincrônico, diacrônico, contemporâneo, histórico;
- Seleção: de amostragem ou geral, monitor, comparável, paralelo, dinâmico ou estático, equilibrado;
- Conteúdo: especializado, regional ou dialetal, multilíngue;
- Autoria: de aprendiz, de língua nativa;
- Finalidade: de estudo, de referência, de treinamento ou teste.

Além da tipologia, importam também algumas características fundamentais como as destacadas por McEnery e Wilson (2001):

- amostragem e representatividade: deve ter uma amostragem suficiente de dados que se quer analisar para ser representativo no estudo da língua. Tem o "objetivo de reunir uma ampla amostra de texto de

diferentes autores e gêneros que tomados em conjunto representam uma proporção relevante para o estudo de uma língua" (p. 30);

- tamanho finito: excetuando o *corpus*-monitor[15], todos os *corpus* devem ter tamanho definido (finito);
- formato "legível em máquina"/eletrônico: os textos devem estar disponíveis em meio eletrônico, pois essa forma proporciona vantagem nas análises linguísticas (podem ser pesquisados e manipulados, e facilitam, agilizam a análise e possibilitam o enriquecimento/alimentação dos *corpus* com informações extras);
- padrão de referência: constitui uma referência padrão para a variedade da língua que representa, pois deve estar disponível para outros pesquisadores. Novos resultados sobre o tema em análise podem ser comparados com outros resultados já publicados, sem a necessidade de refação ou reuso do *corpus*.

Berber Sardinha (2000) elenca quatro pré-requisitos necessários para a formação de um *corpus*:

1º. o *corpus* deve ser composto de textos autênticos, em linguagem natural: são textos naturais que não foram criados com o propósito de figurarem no *corpus*;

2º. deve ser autêntico no sentido de ser escrito por falantes nativos;

3º. o conteúdo do *corpus* deve ser escolhido criteriosamente, considerando-se principalmente as condições de naturalidade e autenticidade já mencionadas, obedecendo a um conjunto de regras estabelecidas para esse fim;

4º. deve ser representativo, e está relacionado à extensão que deve ter o *corpus* de acordo com os seus objetivos: esse pré-requisito é bastante problemático de se estabelecer, pois se tornou relativo na formação de um *corpus* (a importância do tamanho), como é discutido por Berber Sardinha (2004), Hunston (2002), Sarmento (2008) e Shepherd (2012).

Quanto à representatividade, portanto, o tamanho do *corpus* depende dos

[15]Este pode receber novos textos enriquecendo o seu banco de dados.

objetivos da pesquisa, e como Berber Sardinha (2000, p. 342-343) destaca:

> 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 [...] Não há critérios objetivos para a determinação da representatividade.

Hunston (2002) e Sarmento (2008), em referência ao tamanho do *corpus*, defendem que se o pesquisador preferir compilar um *corpus* menor, este pode ser considerado tão confiável quanto um *corpus* maior, pois o tamanho influenciará sobremaneira na velocidade e a eficiência do *software* de acesso a esse *corpus*, assim como a capacidade do computador de processá-lo.

A prática do linguista de *corpus*, ao longo de todos esses anos, determinou as mudanças encaradas até então como “regras”. Muitos aspectos hoje são vistos de forma diferente, especialmente sobre o tamanho e a representatividade dos *corpora*. “O maior é melhor” já não é senso comum, pois é “possível formular interessantes hipóteses com *corpora* pequenos, como é geralmente o caso dos *corpora* de aprendizes, ou os *corpora* de textos literários de um determinado autor, compilados para os estudos de estatística de *corpus*” (SHEPHERD, 2012, p. 16).

Nesse universo, os linguistas de *corpus* e outros pesquisadores necessitam saber distinguir o que é um *corpus*. Diante de várias características já citadas sobre o que é um *corpus*, convém ressaltar o que não é considerado como um *corpus* para a LCORP (ATKINS; CLEAR; OSTLER, 1991, p. 1):

> - Arquivo: depósito de textos sem organização prévia;
> - Biblioteca eletrônica: coleção que segue alguns critérios de seleção;
> - *Corpus*: uma parte da biblioteca eletrônica, construído a partir de um desenho explícito, com objetivos específicos;
> - *Subcorpus*: uma parte de um pode ser fixa ou mutável, dinâmica e flexível durante a análise.

Souza e Gatti (2002, p, 241) resumem os *corpora,* similar à tipologia adotada por Sinclair (1995), Berber Sardinha (2000), Hunston (2002), Tagnin (2010), em cinco tipos:

> - Monolíngues, bilíngues ou multilíngues;
> - Anotados (contendo anotação/ classificação morfossintática das palavras dos textos) ou não-anotados;

- Paralelos (contendo textos de uma determinada língua e traduções destes textos em uma ou mais línguas) ou comparáveis (contendo originais de determinados tipos de texto em duas ou mais línguas);
- Sincrônicos ou diacrônicos;
- Abertos (aos quais continuamente são acrescentados textos) ou fechados.

Destaque para essa tipologia é quanto à inserção do tipo de *corpus* anotado ou não anotado.

Apesar da tipologia adotada por Souza e Gatti (2002), e considerando-se a facilidade que a *WEB* permite quanto à acessibilidade livre de dados sem imposição de licenciamentos ou permissões, o fato de um *corpus* estar aberto ou fechado significa dizer que ele ou está livre para acesso via *WEB* ou só pode ser utilizado por meio de permissão solicitada.

As concordâncias de textos eram os dados mais procurados nos *corpus* nos primeiros estudos da LCORP*.* Antes do advento do processamento computacional, essa base textual era organizada de forma minuciosa e consistia em uma tarefa laboriosa, haja vista que era feita manualmente: a verificação das ocorrências de uma palavra em um texto ou coletânea de textos tornou-se mais fácil com o uso do computador e os aplicativos específicos para esse fim.

Somente na década de 90, a computação passou a fazer parte do cotidiano acadêmico, e a partir de então os textos em formato digital e as ferramentas disponibilizadas para análises de *corpora* contribuíram para o desenvolvimento da LCORP, no caso o *WordSmith Tools*. Muitas instituições e pesquisadores utilizam essa ferramenta, mas pelo fato de ela necessitar de aquisição de licença de uso, outras ferramentas de código aberto (*open source*) estão cada vez mais em evidência, possibilitando as mesmas tarefas sem custo nenhum.

Os computadores passaram a interferir senão mudar a concepção do termo *corpus*, sobretudo no formato de armazenamento e exploração dos dados. A gama de dados que podem ser processados por meio eletrônico, antes inimagináveis de serem processados de forma manual, não digital, permitiu que “muitas hipóteses sobre determinados fenômenos linguísticos pudessem ser testadas rápida e eficientemente” (ALUÍSIO; ALMEIDA, 2006, p. 158), permitindo assim uma minuciosa observação e descrição de fenômenos linguísticos até então não possíveis de ser analisados manualmente.

A Linguística de *Corpus* colabora, portanto, para a redefinição dos rumos da Linguística. Na área de compilação de *corpora* eletrônicos, pesquisadores como Rocha (2005), Maciel (2005) e Tagnin (2005) coletaram *corpora* específicos para seus propósitos, pois a aplicação de ferramentas simples e gerais (como contadores de frequência, ocorrências e concordanciadores) contribui para a redefinição dos rumos da linguística em geral, com a manipulação de *corpora* capazes de mostrar "facetas inéditas da linguagem" (BERBER SARDINHA, 2005a, p. 27). E também, "tem 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.

Como exemplo dessas investigações, podemos citar o uso de lista de frequências que mostra as palavras que ocorrem em um *corpus*, juntamente com o número de vezes que cada palavra aparece. A comparação de listas de frequência pode fornecer informações interessantes sobre diferentes tipos de textos ou em comparação com textos da mesma época, especialmente os *corpora* especializados, pois os "textos são formatados por textos anteriores, através de repetições ou através de rotinas e convenções [...] os textos são historicamente herdados" (SARMENTO, 2008, p. 33).

Apesar de comum nas investigações baseadas em *corpus*, a frequência de ocorrências não pode ser considerada uma trivialidade, "pois é através do conhecimento da frequência atestada que se pode estimar a probabilidade teórica" (BERBER SARDINHA, 2000, p. 352).

Bacelar do Nascimento (2002, p. 1) reafirma a importância dos *corpora* para a Linguística, quando diz que eles

> favorecem essencialmente uma Linguística descritiva, fortemente apoiada pelas novas tecnologias, e permitem tomar como ponto de partida da descrição, a análise de quantidades significativas de dados autênticos, à semelhança do que se faz noutros domínios científicos. O uso de *corpora* permite a realização de descrições linguísticas de base empírica e promove, com isso, a discussão de questões teóricas solidamente fundamentadas.

Nosso objetivo neste trabalho é a compilação de um corpus de textos literários, que serve a objetivos específicos, como destaca Sarmento (2008, p. 28-29) quanto à criação de *corpus* dessa natureza:

> Esse tipo de *corpus* tem por objetivo ser representativo de certo tipo de texto, ou linguagem. É comumente compilado pelo próprio pesquisador para

> refletir o tipo de linguagem que quer investigar. Não há limite para o grau de especialização envolvido, mas há parâmetros para limitar o tipo de texto incluído.

Os textos literários podem conter mais ocorrências de criatividade lexical que outros textos, pois tendem a liberar todo o potencial criativo da língua; outras categorias textuais se baseiam num potencial linguístico reduzido (ABREU, 2007).

2.3.2 A LCORP no Brasil

A história da LCORP no Brasil ficou sedimentada a partir dos Encontros de Linguística de *Corpus*, organizados desde 1999. Já está na sua 12ª versão. Os primeiros encontros foram organizados por Stella Tagnin da USP, e primordialmente objetivavam discutir os princípios básicos para a constituição de um *corpus* (BERBER SARDINHA; ALMEIDA, 2008). Mas, como Berber Sardinha (2003) lembra, o pioneirismo no processamento da linguagem por computador no país pode ser atribuído a Maria Teresa Biderman, Zilda Maria Zampparoli já desde os anos de 1980.

No terceiro Encontro de LCORP foram socializadas as experiências dos doze projetos de *corpora* em desenvolvimento no Brasil. Daí então, a cada encontro, novas experiências foram mostradas, bem como discutidas questões importantes sobre os trabalhos desenvolvidos como: o intercâmbio entre as instituições, padronização de recursos e especialmente a formação de parcerias e de grupos de pesquisadores. O mapeamento do que estava sendo desenvolvido no Brasil em LCORP era necessário nesse momento inicial, pois quanto mais congregassem experiências e pesquisadores, mais fortalecida ficava a Linguística de *Corpus*. Instituições como USP, UNICAMP, UFSCar, UNESP, UERJ, PUC-RS e UFMG têm se revezado em sediar estes eventos.

O primeiro livro publicado da área no Brasil foi produzido por Berber Sardinha em 2004, com o título “Linguística de *Corpus*”: dez anos então de produção impressa na área. Portanto, a LCORP é bastante recente, considerando-se que as outras áreas de pesquisa têm décadas de existência e já bastante cristalizadas no meio acadêmico. Os principais centros de pesquisa da LCORP concentram-se nas regiões onde são sediados os encontros.

Mais recentemente outros centros vêm se destacando na produção de trabalhos de LCORP, como: UFC, UFJF, PUC/SP, UFPEL, UFU, UFRGS e outras, como é possível observar no site da Plataforma Lattes quando se busca os grupos de pesquisas dessa área.

Além disso, nos últimos anos, diversos eventos aconteceram em todo o território nacional, como: "simpósios, comunicações, mesas redondas, seminários, palestras e pôsteres que acontecem em eventos de Linguística, Linguística Aplicada, Tradução, Letras, entre outros" (BERBER SARDINHA; ALMEIDA, 2008, p. 33). Geralmente esses eventos envolvem a comunidade linguística internacional, sobretudo os europeus, pois a Grã-Bretanha é um dos centros mais desenvolvidos em "pesquisas baseadas em *corpus* para a descrição dos mais variados aspectos da linguagem" (BERBER SARDINHA, 2000, p. 328). Fora da Europa, o desenvolvimento da LCORP é muito pequeno. Mesmo os EUA, que têm grandes centros de pesquisas avançadas especialmente nas áreas tecnológicas, têm uma presença mais modesta no avanço das pesquisas em Linguística de *Corpus*.

Do ano 2000 para os dias de hoje o desenvolvimento da LCORP no Brasil ganhou outras nuances. Berber Sardinha e Almeida (2008, p. 35) ratificam essa situação ao afirmar que, de 1999 a 2008,

> houve um crescimento qualitativo e quantitativo das pesquisas realizadas e a crescente formação em recursos humanos, uma vez que já se observa a realização de iniciações científicas, mestrados e doutorados que têm como tema central a Linguística de *Corpus*.

Esse crescimento legitima cada vez mais a Linguística de *Corpus*. Dezenas de projetos estão sendo executados no âmbito acadêmico, sedimentando assim o tema nas universidades.

Eventos como PROPOR[16] (*International Conference on Computational Processing of Portuguese Language*), STIL[17] (*Brazilian Symposium in Information and Human Language Technology*), ELC[18] (Encontro de Linguística de *Corpus*) e LIPRAL[19] (Colóquio de Linguística para o Processamento Automático de Linguagem Natural), todos de padrão internacional, fomentam e sedimentam cada vez mais as pesquisas e produtos desenvolvidos pelos linguistas computacionais e pelos

[16] PROPOR - http://nilc.icmc.usp.br/cgpropor/
[17] STIL 2013 - http://www2.unifor.br/bracis2013/stil/
[18] XI ELC - http://www.nilc.icmc.usp.br/elc-ebralc2012/index.php/pt/
[19] 1º LIPRAL - http://eventos.ufes.br/index.php/lipral/LiPrAL2012

linguistas de *corpus*. A convergência das áreas de Engenharia de Computação, Engenharia Elétrica, Inteligência Artificial, Literatura e Linguística, nesses eventos, inauguram um cenário senão diferente, mas inovador em todas essas áreas, abrindo frente para novas linhas de pesquisa.

Boa parte das pesquisas e leituras sobre LCORP estão disponíveis na *WEB*. A LCORP tem poucas publicações (livros e periódicos) disponíveis para consulta pelos prováveis pesquisadores que queiram enveredar por essa área. As produções de 2004 para cá têm aumentado consideravelmente, mas ainda estão concentradas na região Sul e Sudeste.

Berber Sardinha e Almeida (2008) em um levantamento realizado sobre pesquisadores, grupos de pesquisa, teses e dissertações, *corpora* e ferramentas, consultando o portal da CAPES e outra fontes, observaram que apesar de, à época, as produções publicadas serem tímidas, em pequeno número, perceberam também que os dados encontrados não refletem a realidade, visto que há mais dissertações e teses defendidas em LCORP que não são socializadas nos canais de divulgação acessíveis via internet. Constatamos que essa realidade pouco mudou.

O levantamento recente, realizado por Alves e Tagnin (2012) para analisar a visibilidade da LCORP em oitenta e quatro trabalhos acadêmicos, confirmou que a LCORP está presente principalmente em cinco áreas: ensino, tradução, terminologia, descrição da linguagem e PLN, por ordem de preferência e interesse.

Alves e Tagnin (2012) se ressentem da dificuldade em buscar informações sobre os trabalhos acadêmicos produzidos na área de LCORP, relatam que faltam referências claras nos títulos, resumos e palavras-chaves que auxiliariam na busca e recuperação desses trabalhos por outros estudiosos, tendo assim a LCORP uma maior visibilidade “fazendo jus a sua relevância, tanto como objeto de pesquisa em si, quanto como abordagem em várias outras áreas” (p.33). Alinhados a essa sugestão, estruturamos o título deste trabalho com o intuito de dar maior visibilidade à esta pesquisa, e facilitar a busca por trabalhos acadêmicos nestas áreas.

Buscamos como referência de indexadores de trabalhos científicos o Domínio Público e BDTD (Biblioteca Digital Brasileira de Teses e Dissertações) - esta é recomendação da CAPES para que todas as IES publiquem suas produções.

O Domínio Público contava, até o momento da pesquisa, com 4 teses e 6 dissertações com o assunto Linguística de Corpus, e 2 teses e 4 dissertações com o assunto Linguística Computacional. Já na BDTD encontramos disponíveis 12 teses e 43 dissertações com o assunto Linguística de Corpus, e 4 teses e 6 dissertações com o assunto Linguística Computacional.

Buscamos outros indexadores de publicações acadêmicas (teses e dissertações) e encontramos quantidade e variedade de publicações diferentes das publicadas na BDTD, mas esta é a que contém maior volume de publicações disponíveis.

Considerando-se a disponibilização dessas publicações, por vezes repetidas em alguns diretórios, fizemos uma busca rápida para visualizar as publicações das Instituições. Ao fazer a busca avançada na BDTD, observamos que as publicações da UFC não estão disponíveis, portanto buscamos as publicações no diretório de publicações da UFC/TEDE, dentre as quais constam 5 dissertações com os assuntos Linguística Computacional e Linguística de Corpus. O portal Domínio Público disponibiliza poucas produções acadêmicas nessas áreas (oito no total), e percebe-se que das que estão publicadas neste domínio, somente uma está disponível no BDTD. Ocorre que ainda não há um banco de teses e dissertações vinculadas a todas as IES brasileiras, cada instituição tem seu banco de publicação e que nem sempre facilitam a pesquisa avançada, por assunto. Ou por não dar acesso livre, ou pelas dificuldades relatadas por Alves e Tagnin (2012) ao se buscar trabalhos nessas áreas na *WEB*.

Em uma busca rápida, grosso modo, na Plataforma Lattes, utilizando a expressão “Linguística de Corpus”, encontramos 1.431 ocorrências de pesquisadores que atuam na área da LCORP, entre brasileiros e estrangeiros atuando no Brasil. Colocando a expressão “Linguística Computacional” foram encontradas 557 ocorrências de pesquisadores. Não foi possível fazer o cruzamento dos dados para precisar se os pesquisadores coincidem nas duas áreas, mas somente com esses dados é possível depreender o quantitativo de pesquisadores da LCORP.

Com o aumento de pesquisadores consequentemente o número de produções aumentou, comparando-se com os dados coletados em 2008, por Berber Sardinha e Almeida (2008), que constataram 132 ocorrências de pesquisadores que mencionavam “Linguística de *Corpus*” em seus currículos. Portanto, um avanço

significativo de 947,8%. Do levantamento feito, nem todos são linguistas de *corpus* ou computacional, basta somente que em seu currículo venha mencionado o assunto para constar na relação. Uma triagem mais detalhada demandaria um trabalho muito minucioso, visto que o portal não dispõe de um buscador mais avançado, e para isso teria que ser analisado currículo por currículo, o que não é pretensão deste estado da arte.

Também, comparando a busca realizada ainda por estes autores, que encontraram na Plataforma Lattes 12 grupos de pesquisas cadastrados, que têm como linha de pesquisa a “Linguística de Corpus”, realizamos uma busca similar no Diretório dos Grupos de Pesquisa no Brasil/CNPQ e foram encontrados 41 grupos de pesquisa (ver Apêndice I), sendo que dos 12 grupos relacionados pelos autores citados, somente um não continuou sua atividades, em compensação outros pesquisadores criaram vinte e nove novos grupos de pesquisa. Um aumento exponencial em torno de 241%. Com isso, constatamos um grande crescimento das pesquisas nesta área, especialmente se levarmos em consideração que a diferença de uma busca para outra foi de somente sete anos.

Utilizando o termo “Linguística Computacional” no Diretório foram encontrados 41 grupos de pesquisa (ver Apêndice J). Fazendo um cruzamento dos dados, observamos que dos 41 grupos de pesquisa de LCORP somente 11 são coincidentes com os grupos de pesquisa da LCOMP. Totalizando os grupos de pesquisa de LCLC teríamos então 71 grupos.

Estes dados mostram que a cada dia as pesquisas de LCLC avançam, mas se se comparássemos com outras áreas da Linguística, perceberíamos que as pesquisas ainda são pequenas. Nessa comparação, vale ressaltar o tempo de atuação de cada área no ambiente acadêmico.

2.3.3 Os Corpora do PB e PE

Os diversos *corpora* do português brasileiro constituídos desde o desenvolvimento da Linguística de *Corpus* no Brasil, basicamente década de 90, constituem-se resultados exitosos das diversas instituições e pesquisadores que se debruçaram sobre o tema para destrinçá-lo e desenvolvê-lo. Quanto ao uso que se faz desse material é impossível de se mensurar, pois cada *corpus*, considerando-se

sua especificidade, “passa a ter apenas uma pequena parte do total de amostras potenciais da língua” (OLIVEIRA; DIAS, 2009, p. 194).

Estes projetos tiveram como inspiração o *corpus* Brown (*Brown University Standard Corpus of Present-Day American English*)[20], de 1964, considerado o primeiro *corpus* linguístico eletrônico e continha à época um milhão de palavras, contendo quinze categorias de textos compilados a partir de trabalhos publicados nos Estados Unidos em 1961 (BERBER SARDINHA, 2000).

Os corpora do PB, também, foram construídos tendo como base diversos *corpora* da Língua Portuguesa instituídos em grandes centros acadêmicos europeus que mesclam textos do português brasileiro e do português europeu.

Configurando-se como um dos maiores centros de recursos distribuídos para o processamento computacional da língua portuguesa, a Linguateca[21] incorpora diversos *corpora* da língua portuguesa. Todos os trabalhos desenvolvidos nesse centro são públicos, de livre acesso pela internet. Tem como objetivo primordial “criar condições para a existência de recursos bons e gratuitos para a língua portuguesa” (LINGUATECA, 2012). Agrega diversos corpora tanto do PB quanto do PE.

A Linguateca provém da continuação do projeto “Processamento Computacional do Português” lançado pelo Ministério da Ciência e da Tecnologia (MCT) de Portugal entre maio de 1998 e maio de 2000. Este centro de recursos é composto de pesquisadores da Universidade Técnica de Lisboa, do Instituto Superior de Línguas e Administração de Lisboa, da Universidade de Coimbra, da Pontifícia Universidade Católica do Rio de Janeiro e da USP.

Com base no levantamento feito por Berber Sardinha e Almeida (2008); e Berber Sardinha (2000), acrescidos de dados recentes, destacamos nos apêndices A e B alguns *corpora* criados e mantidos no Brasil e na Europa.

[20] *Corpus* Brown – http://khnt.hit.uib.no/icame/manuals/brown/INDEX.HTM
[21] No site do Linguateca estão relacionados os diversos *corpora* tanto do PB quanto do PE - http://www.linguateca.pt/

2.4 Ferramentas Computacionais na LCLC

A LCLC utiliza *corpus* digitais e ferramentas computacionais para sua análises e tarefas linguísticas.

Dentre as ferramentas utilizadas na análise automática de textos para extrair dados e ou validar hipóteses linguísticas, o mais utilizado é o programa *WordSmith Tools*, um "quase absoluto" (ALENCAR, 2009, p. 136), aplicado em alguns trabalhos realizados nessa área (ver BERBER SARDINHA, 2000, 2003, 2005b; SHEPHERD; BERBER SARDINHA; PINTO, 2012; TAGNIN; VALE, 2008; SARMENTO, 2008).

O *WordSmith Tools* é um conjunto de programas integrados destinado à análise linguística:

> Mais especificamente, esse software permite fazer análises baseadas na frequência e na co-ocorrência de palavras em *corpora*. Além disso, ele permite pré-processar os arquivos do *corpus* (retirar partes indesejadas de cada texto, organizar o conjunto de arquivos, inserir e remover etiquetas, etc.), antes da análise propriamente dita (BERBER SARDINHA, 2009, p. 6, *sic*).

O programa é de fácil uso, deveria ser uma ferramenta que fosse acessível a todos, o que ocorre é que como considera Alencar (2009, p. 136), o programa tem uma série de desvantagens: "tem código proprietário, não permitindo a introdução de aperfeiçoamento por outros pesquisadores; exige licença que, para o usuário individual, custa 50 libras esterlinas; não é multiplataforma (só funciona em ambiente Windows)".

Python[22] é outra ferramenta computacional utilizada na LCLC. Criada em 1991, por Guido Van Rossum, é uma linguagem de programação intuitiva, de fácil assimilação, mas de altíssimo nível, orientada a objeto, de tipagem dinâmica e forte, interpretada e interativa e, originalmente, tinha foco em usuários como físicos e engenheiros. O nome Python é originário do programa da TV britânica *Monty Python Flying Circus* (BORGES, 2010; PYTHON.ORG, 2014).

Essa linguagem de programação se configura como uma linguagem dinâmica, por utilizar uma metodologia ágil e ter nascido do "desenvolvimento de *software* de código aberto e defendem um enfoque mais pragmático no processo de

[22] Python – site oficial: https://www.python.org/

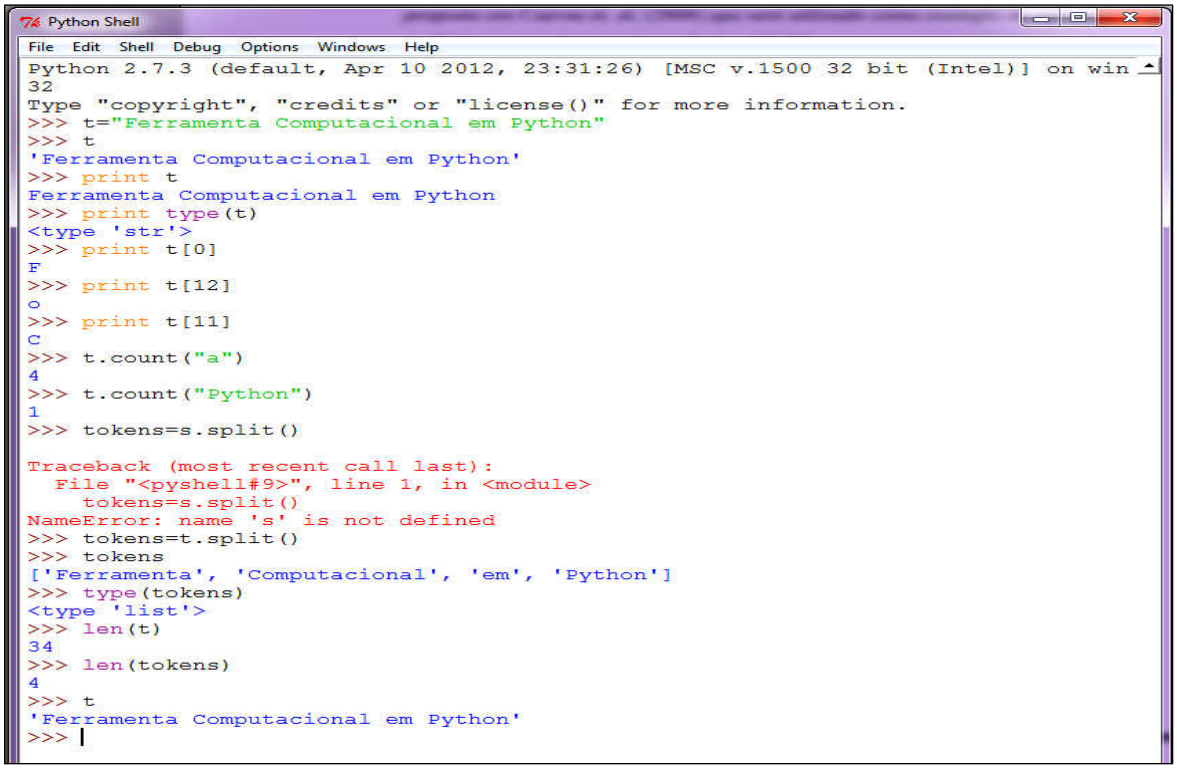
criação e manutenção” (BORGES, 2010, p. 11). Contém uma vasta coleção de módulos prontos para uso, além de *frameworks* de terceiros que podem ser adicionados. Python é uma das mais populares e poderosas linguagens de programação e é considerada uma excelente ferramenta para o processamento linguístico de dados.

Seu uso é livre, portanto com código-fonte aberto não necessitando de aquisição de licença, e pode ser rodado em Windows, Linux / Unix, Mac OS X. Essa linguagem é comumente comparada ao Tcl, Perl, Ruby, Scheme ou Java (linguagens de programação). Conta com uma vasta biblioteca de funcionalidades para processamento de XML e HTML, banco de dados e muito mais (PYTHON.ORG, 2014).

Na defesa do uso de Python, Bird, Klein e Loper (2009, p. 363) defendem o programa como primeira linguagem de programação a ser aprendida para a análise automática de textos, sobretudo pela transparência, e a simplicidade do código dessa linguagem em comparação com Perl, Prolog e Java, entre outras linguagens.

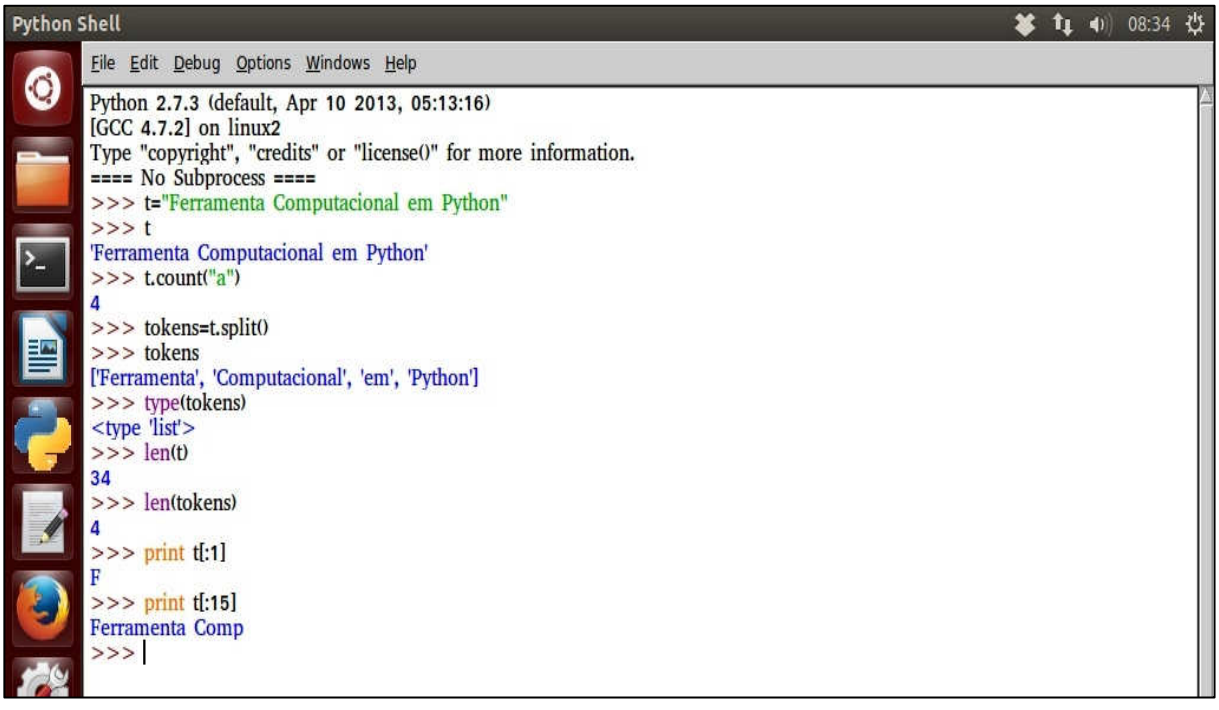
Comungando com o entendimento destes pesquisadores, optamos por trabalhar com Pyhton. Por ter uma interface gráfica simples, Python pode ser acessada através da IDLE (*Interactive Development Environment*), traduzido como ambiente de desenvolvimento interativo, como mostrados nas figuras 1 e 2.

Figura 1 – IDLE/Shell de Python no Windows: interface de execução de programas em Python.



Fonte: Elaborada pela autora

Figura 2 – IDLE/Shell de Python no Linux: interface de execução de programas em Python.



Fonte: Elaborada pela autora

Possibilitando um acesso a diferentes *corpora*, Python permite a manipulação dos dados como melhor convier ao utilizador, buscando diversos aspectos da linguagem. Basta que para isso o utilizador faça o *download* do programa[23] e instale os pacotes necessários para a consecução das análises. De forma mais delimitada, permite também buscar concordâncias, frequências, quantidades, repetições, diversidades, exceções, listas, extrações, sentenciação, toquenização (segmentação de um texto em unidades menores), anotações morfossintáticas dentre outras.

Atualmente, a *Python Software Foundation* detém os direitos autorais sobre as versões de Python. É uma organização independente sem fins lucrativos que tem a missão de promover e divulgar tecnologia de código aberto relacionado com a linguagem de Python. A última versão de Python é a 3.4.2 de 2014.

O *Natural Language Toolkit (NLTK)* é um conjunto (kit) de ferramentas utilizado na construção de programas de Python para manipulação de dados voltados para trabalhar com o Processamento da Linguagem Natural, que necessita manipular grandes volumes de textos com alta performance e capacidade de processamento.

Teve seu desenvolvimento iniciado em 2001 por Steven Bird, Ewan Klein e Edward Loper, como parte de uma disciplina em linguística computacional da Universidade da Pensilvânia (NLTK.ORG, 2012). Desde então, o NLTK tem sido utilizado em mais de 60 cursos universitários e em mais de 20 países. Sua evolução pode ser avaliada com base nas experiências e *feedbacks* dados por professores e alunos que participaram de vários cursos ministrados com e sobre NLTK (BIRD et al., 2008).

Contém um conjunto de bibliotecas de processamento de texto para geração de *tokens*, classificação e análise, dispondo de ferramentas e recursos para diversas etapas de análise automática de textos, como: toquenização ao *parsing*[24] sintático e semântico; etiquetagem morfossintática; o *chunking*[25]; o reconhecimento

[23]PYTHON - http://www.python.org/

[24] O mesmo que análise sintática automática ou semiautomática das sentenças das línguas naturais (OTHERO; MARTINS, 2011).

[25] Na Linguística Computacional é compreendido como um método para reconhecer segmentos de texto que são sintagmas nominais, dividindo-as em pedaços, fazendo uma análise sintática parcial das sentenças (ALENCAR, 2011b): “utiliza autômatos finitos e expressões regulares para formar o agrupamento de palavras dependentes entre si, através da análise de sua classe gramatical. Através deste processo, são identificados grupos linguísticos denominados *chunks*” (PADILHA; LACERDA, 2012, p. 3).

de entidades nomeadas e a classificação de textos (BIRD, KLEIN, LOPER, 2009; ALENCAR, 2013a).

Seguindo a mesma linha de linguagem dinâmica de Python (código-fonte livre e aberta), o NLTK é adequado “para linguistas, engenheiros, estudantes, educadores, pesquisadores [...] está disponível para Windows, Mac OS X e Linux” (NLTK.ORG, 2012). Em suma, é uma ferramenta escrita em Python e para Python. A flexibilidade do NLTK permite aos seus usuários, mesmo que iniciantes em programação, desenvolver seus próprios programas em Python (ALENCAR, 2011a).

O kit de pacotes NLTK contém muitas tarefas básicas de PLN com uma extensa documentação: “é uma gigantesca biblioteca com centenas (senão milhões) de funções para o processamento computacional da linguagem” (ALENCAR, 2011a, p. 9). E é considerado por Alencar (2011a, p. 11) como o “mais didático, amigável e abrangente projeto de *software* livre na área de PLN”.

Os desenvolvedores do NLTK[26] disponibilizam no *site* pacotes, módulos e links com versões atualizadas. Além disso, tem um link que dá acesso livre ao conteúdo do livro *Natural Language Processing with Python*[27] (BIRD; KLEIN; LOPER, 2009) que orienta sobre o uso do NLTK contendo centenas de exemplos e está direcionado, sobretudo, para pessoas que querem aprender a utilizar programas que analisam a linguagem escrita, sem que para isso tenham muita experiência em programação. A biblioteca do NLTK conta atualmente somente com dois *corpora* anotados do português, e que podem ser manipulados: O Floresta Sintá(c)tica e o MAC-Morpho (ALENCAR, 2013a). No Floresta, uma de suas partes, a Selva, contém amostras de textos literários do PB, como detalhado no item 5.1. Já o MAC-Morpho é composto somente de textos jornalísticos.

O NLTK é uma biblioteca voltada especialmente para o ensino e aprendizagem da linguística computacional, do processamento automático da linguagem natural (PALN) e da linguística de corpus (ALENCAR, 2013a, p. 9).

Dentre os módulos mais importantes do NLTK, temos:

[26] NLTK - http://www.nltk.org
[27] Book Natural Language Processing with Python - http://nltk.org/book/

Quadro 1 - Módulos do NLTK

Módulos NLTK	Tarefa de processamento
nltk.corpus	Acessa corpora
nltk.tokenize, nltk.stem	Processamento de strings (linhas)
nltk.collocations	Busca de colocações
nltk.tag	Part-of-speech tagging (etiquetagem)
nltk.classify, nltk.cluster	Classificação
nltk.chunk	Chunking (reconhecimento de segmentos de textos)
nltk.parse	Parsing (análise)
nltk.sem, nltk.inference	Interpretação semântica
nltk.metrics	Métricas de avaliação
nltk.probability	Probabilidade e estimativas
nltk.app, nltk.chat	Aplicações
nltk.toolbox	Trabalho de campo linguístico

Fonte: Adaptado de BIRD, KLEIN, LOPER, 2009.

Importante observar para melhor entendimento, que todos esses módulos, dependendo dos dados pretendidos, com auxílio de Python, são mesclados a outros comandos mais regulares para geração dos dados finais, como se observa no exemplo do Quadro 3:

Quadro 2 - Exemplo de grupo de comandos para extração de dados

```
>>> import os, nltk, Aelius
>>> from Aelius import Extras, Toqueniza, AnotaCorpus
>>> corpus=AnotaCorpus.extrai_corpus("ocortico06.pdt.txt")
>>> tokens=corpus.words()
>>> len(tokens)
1154
```

Fonte: Elaborado pela autora

Neste exemplo, os comandos geraram a quantidade de *tokens* que contém o arquivo "*ocortico06.pdt.txt".* Para um entendimento básico de como funciona um simples processamento de comandos, temos:

- ***os, nltk, Aelius*** são módulos do NLTK
- ***Extras, Toqueniza, AnotaCorpus*** são funções importadas do Aelius

- ***corpus*** = nome dado ao resultado a ser gerado
- ***AnotaCorpus.extrai_corpus*** função que extrai dados do arquivo
- ***"ocortico06.pdt.txt"*** é a denominação do arquivo explorado
- ***corpus_word()*** é o método de extração de palavras
- ***len(tokens)*** faz leitura do quantitativo de *tokens*

Uma ferramenta muito utilizada na LCLC é o etiquetador automático, que foi desenvolvido com base nos estudos da Linguística Computacional, envolvendo as subáreas de PLN e Linguística de *Corpus*, e tem como tarefa "identificar as categorias gramaticais das palavras em sentenças" (DOMINGUES; FAVERO; MEDEIROS, 2008, p. 269) através da etiquetagem morfossintática, que tem uma tarefa básica e importante na Linguística de *Corpus* ao anotar *corpus*, extrair e recuperar informações, correções gramaticais, traduções automáticas dentre outras.

Os etiquetadores automáticos são estruturados e baseados em abordagens de acordo com o tipo de conhecimento que usam para representar o modelo linguístico: a) baseada em regras ou simbólica, caracterizada pelo uso de regras codificadas manualmente e depois pela aplicação de abordagens semi-autorizadas; b) probabilística, que aplica métodos de etiquetagem estáticos (baseados em *n-gramas* nos modelos de Markov); c) híbrida, que surgiu porque alguns etiquetadores empregavam o conhecimento baseado em regras e o probabilístico simultaneamente; sua característica é a etiquetagem de textos com regras aprendidas automaticamente de *corpora* anotados (DOMINGUES; FAVERO; MEDEIROS, 2008; ALENCAR, 2010a).

Alencar (2011b, p. 75) diz que o termo etiquetagem morfossintática (ou *POS-tagging*) "é empregado de forma lata, abrangendo a etiquetagem morfológica". Já o termo etiquetagem morfológica é utilizado com menos frequência como hiperônimo do primeiro. A diferenciação é feita por Xiao (2010 *apud* ALENCAR, 2011b, p. 75) ao dizer que "a anotação morfológica tem como escopo unidades sublexicais como prefixos, sufixos e raízes, ao passo que a etiquetagem morfossintática incide sobre unidades lexicais".

Neste trabalho, utilizaremos o termo anotação morfossintática ou anotação automática e não etiquetagem morfossintática, considerando-se que a ferramenta computacional, o Etiquetador, tem com uma de suas tarefas a de anotar

textos (CARVALHO; VASCONCELOS; ALENCAR, 2012; ALENCAR, 2010a; DUARTE, 2009). Diante das leituras realizadas para composição desse trabalho e que tratam do assunto, percebe-se que não há uma uniformidade quanto ao uso desses termos: por vezes são considerados como sinônimos.

A anotação morfossintática “é uma tarefa intermediária que tem como objetivo principal analisar e entender a língua natural” (CARVALHO; VASCONCELOS; ALENCAR, 2012, p. 125). O processo de anotação consiste na colocação automática de um código de etiquetas morfossintáticas (*tags*) e utiliza um sistema de etiquetas divididas em dois grupos: etiquetas categoriais (classificam o item lexical em uma classe gramatical) e etiquetas flexionais (indicam designadores de informações modo-temporais, ou não-verbal, indicadoras de traços flexionais de gênero e número) (GALVES; BRITTO, 1999; SARMENTO, 2008). Apesar de fundamentalmente morfológicas, o formalismo dessas análises contempla a possibilidade de se investigar, *a posteriori*, informações sintáticas e semânticas: é, portanto, o primeiro passo para os processos de anotação (NAMIUTI, 2004). Cada etiquetador adota etiquetas próprias, e nem sempre segue o mesmo padrão, por exemplo: DET F S e D-F-S são etiquetas utilizadas para anotar a palavra “a” por etiquetadores diferentes, significam “determinantes”.

Alguns etiquetadores foram adaptados para o processamento de etiquetagem de textos em PB (DOMINGUES; FAVERO; MEDEIROS, 2008; OLIVEIRA; FREITAS, 2006), os mais conhecidos são:

- TreeTagger, 1994, abordagem probabilística;
- Brill TBL, 1995, abordagem híbrida;
- MXPOST, 1996, abordagem probabilística;
- PALAVRAS, 2000, abordagem baseada em regras;
- FreeLing, 2003, abordagem híbrida;
- CoGroo, 2006, abordagem híbrida.

Domingues, Favero e Medeiros (2008) fazem uma análise da acurácia de alguns desses etiquetadores e, a par dos resultados obtidos, entendem que estes não devem ser comparados, pois cada um “utiliza recursos, estratégias e abordagens diferentes para a etiquetagem” (p. 271). Mas, algo que se pode perceber é que os resultados demonstraram que as abordagens híbridas funcionam melhor ao invés de se utilizar uma só abordagem.

Alencar (2011a) construiu diversos etiquetadores morfossintáticos, por meio do NLTK e utilizando e aplicando as técnicas de aprendizado de máquina dessa biblioteca no *Corpus* Tycho Brahe. Esses etiquetadores integram o Aelius, que é "um conjunto de módulos em Python para treino e aplicação de etiquetadores na anotação de textos em diversos formatos" (ALENCAR, 2011b, p. 14).

O Aelius que faz parte do projeto *Aelius Brazilian Portuguese POS-Tagger*[28] (ALENCAR, 2013b), é registrado no SourceForge.net[29] e adota uma abordagem híbrida, que se configura em mesclar as abordagens baseadas em regras e em *n-gramas*. Este projeto intenta preencher

> lacunas na linguística de *corpus* do português de orientação diacrônica, ao mesmo tempo em que contribui para sanar a escassez de *corpora* e *language models* do português no NLTK e acrescenta a essa biblioteca algumas funções bastante úteis para o desenvolvimento de etiquetadores mais eficazes (ALENCAR, 2010a, p, 2).

Executa, para tanto, as seguintes tarefas:

> a) pré-processamento de *corpora*;
> b) construção de *language models* e etiquetadores com base num *corpus* anotado;
> c) avaliação do desempenho de um etiquetador;
> d) comparação entre diferentes anotações de um texto (ALENCAR, 2010a, p. 2).

Além dessas, o Aelius realiza anotação de *corpora* e auxilia na revisão humana de anotação automática (ALENCAR, 2010a).

Dentre os etiquetadores construídos por Alencar (2011a), desde sua formatação, o Aelius adotou vários modelos, como o: AeliusRUBT, AeliusBRUBT, AeliusMaxEnt e AeliusHunPos que foram treinados a partir do *Corpus* Tycho Brahe. Outro modelos como o AeliusHunPosMM e Aelius StandfordMM foram treinados a partir do *corpus* de treino Mac-Morpho (um dos seis *corpora* constituintes do Projeto Lácio-*WEB* do NILC-São Carlos)[30].

O etiquetador Aelius também usa o etiquetador LX-Tagger[31], que por sua vez foi desenvolvido a partir do aplicativo Java MXPOST, que "não é *software* livre,

---

[28] AELIUS BRAZILIAN PORTUGUESE POS-TAGGER - http://sourceforge.net/projects/aelius/files/
[29] SOURCEFOGE.NET - Maior hospedagem mundial de *software* de código aberto. http://sourceforge.net/
[30] **Projeto NILC-São Carlos -** http://www.nilc.icmc.usp.br/lacioweb/
[31] **LX-TAGGER** – http://lxcenter.di.fc.ul.pt/tools/pt/conteudo/LXTagger.html#lxtagger

mas pode ser gratuitamente baixado da Internet e utilizado com base no Aelius" (ALENCAR, 2011b, p. 14). Mas seu uso não permite a criação de produtos derivados, ou seja, as anotações realizadas pelo LX-Tagger não podem ser disponibilizadas gratuitamente, pois necessita de uma licença para uso livre.

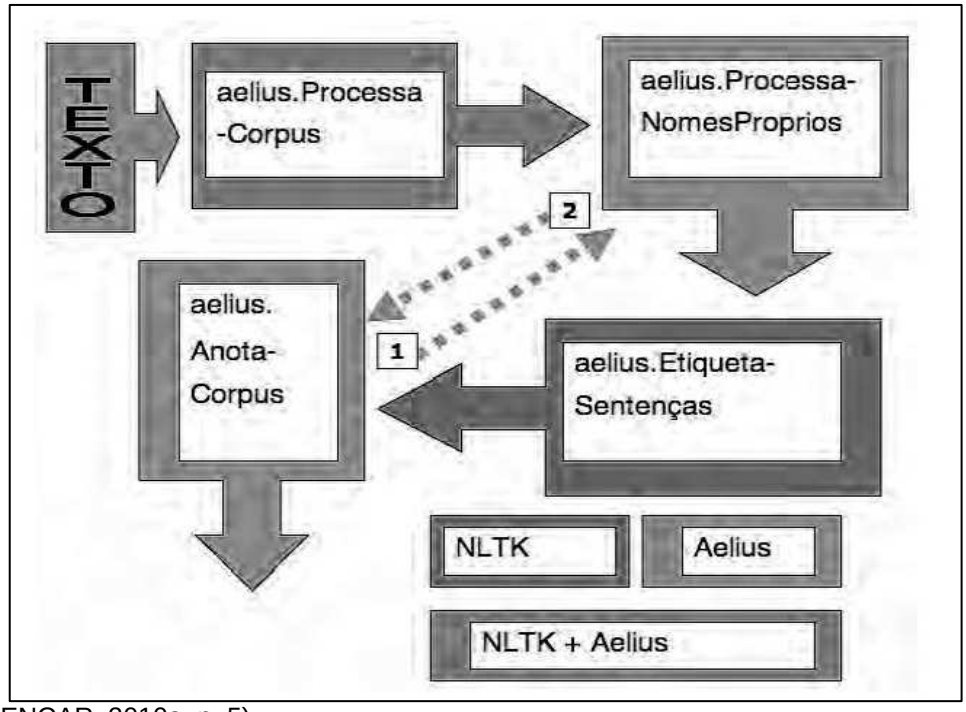
O Aelius teve sua primeira versão apresentada no IX Encontro de Linguística de Corpus na PUCRS em 2010. O nome é uma homenagem ao gramático latino Aelius Donatus, que classificou as palavras em nome, verbos, adjetivos, preposições etc. O desenvolvimento da ferramenta teve como propulsor inicial o fato de se disponibilizar um recurso computacional de "análise automática de textos por meio de diversas bibliotecas livremente disponíveis" (ALENCAR, 2013a, p. 7).

A anotação morfossintática no Aelius ocorre da seguinte forma:

> No *input*, entra-se com um texto não etiquetado. Enquanto etiqueta o texto, o sistema verifica se a palavra pertence a classe dos nomes próprios, nos casos afirmativos, o sistema etiqueta a palavra com a *tag* NPR. Depois, no *output*, o Aelius retorna um texto etiquetado com o *tagset* do *Corpus* Histórico do Português Tycho Brahe (CARVALHO; VASCONCELOS, ALENCAR, 2012, p. 126).

Essa ação é demonstrada na figura abaixo:

Figura 3 - Anotação Morfossintática no Aelius



Fonte: (ALENCAR, 2010a, p. 5).

Essa ferramenta antes do momento da anotação morfossintática “realiza o processo de toquenização, ou seja, o interior do texto é dividido em *tokens* para outra análise como: marcas de pontuação, palavras compostas unidas por hifens etc.” (CARVALHO; VASCONCELOS, ALENCAR, 2012, p. 126).

Por isso trabalhamos com *tokens, que* correspondem ao número total de itens ou palavras, incluindo as repetições de um mesmo item ou palavra, e não com *types*, que não consideram as repetições. Interessa-nos as repetições das palavras, para verificarmos as frequências dos termos.

O processo de anotação é todo automático, gerado por algumas funções/ comandos específicos, como mostrados na figura anterior.

Outra necessidade mais premente para o desenvolvimento do Aelius surgiu em função de se utilizar um etiquetador morfossintático de alta acurácia para textos literários do século XIX. Como os etiquetadores têm a função auxiliar na desambiguação na anotação morfossintática, o Aelius então contribui para isso. Como exemplo de desambiguação, Alencar (2010a) destaca a detecção da grande quantidade de erros relacionados à etiqueta NPR (nome próprio). Inicialmente, o Aelius foi treinado no CHPTB, e fazia uma confusão no momento de anotação por não conter um algoritmo que distinguisse nomes próprios de outras palavras de inicial maiúscula. Isso contribuía muito para que o etiquetador cometesse muitos equívocos diminuindo a acurácia do Aelius.

O Aelius etiquetou inicialmente textos literários usando a arquitetura BRUBT, que alcançou uma acurácia de 95% em uma amostra de textos literários (ALENCAR, 2010a, 2011a). Com a melhoria no desempenho do etiquetador, sobretudo na desambiguação, o AeliusHunPos demonstrou melhor performance, acima de 95%, ao anotar textos literários do séc. XIX, como constatou Alencar (2011a). Portanto, o Aelius “é um sistema com grande desempenho para a etiquetagem do português brasileiro, especialmente, para textos literários do século XIX” (CARVALHO; VASCONCELOS; ALENCAR, 2012, p. 128).

Os modelos utilizados para etiquetagem e treinamento do Aelius, estão demonstrados no Quadro 3:

Quadro 3 - Modelos do Etiquetador Aelius

MODELO	ACURÁCIA	INTERNO	*CORPUS* DE TREINO	ARQUITETURA DE APRENDIZAGEM DE MÁQUINA/LINGUAGEM	NATIVO
AeliusRUBT.pkl	95,29%	sim	CHPTB-M	nltk.TrigrammTagger/ Python	sim
AelillsBRUBT.pkl	95,30%	sim	CHPTB-M	nltk.FastBrillTaggerTra/ Python	sim
AeliusHunPos	96,35%	sim	CHPTB-M	HunPos / OCaml	não
AeliusMaxEnt	95,81%	sim	CHPTB-M	MXPOST / Java	não
AeliusStanfordMM	92,60%	sim	MAC-Morpho	StanfordTagger / Java	não
AeliusHunPosMM	97,17%	sim	MAC-Morpho	HunPos / OCaml	não
LX-Tagger	97,71%	sim	CINTIL	MXPOST / Java	não

Fonte: ALENCAR, 2013a, p. 14.

Observa-se nesse quadro que o CHPTB foi treinado por quatro modelos do Aelius, sendo que a melhor acurácia foi encontrada usando o HunPos (96,35%), o que justifica o uso deste no *Corpus* Coelho Netto, sobretudo pelo fato de o *Corpus* Tycho Brahe ser considerado o maior e mais famoso *corpora* de textos literários da língua portuguesa (PE e PB). O nível de precisão obtido quando da anotação pelo Aelius é bem maior que dos outros etiquetadores similares disponíveis e voltados para a análise do português, sobretudo o português brasileiro.

O CORPTEXLIT, inicialmente, foi anotado morfossintaticamente utilizando o Aelius modelo AeliusBRUBT, e alcançou um índice de acurácia de 96.84% (ALENCAR, 2010a, 2013b). Utilizando o HunPos, obteve uma acurácia de 97,47%. Esses dados também reforçam o uso deste modelo em textos literários.

Independente do tamanho do *corpus*, o alto índice de acurácia obtido por um tipo de texto não garante desempenho igual em textos de natureza muito diferente daqueles do *corpus* de treino:

> [...] um teste desses modelos com a etiquetagem dos dois primeiros parágrafos de *Luzia-Homem* [...] parece corroborar a expectativa de que um modelo treinado em textos jornalísticos da década de 1990 apresenta queda de acurácia quando aplicado a um texto literário de época bastante anterior (ALENCAR, 2013a, p. 14).

Portanto, para cada tipo de texto deve-se utilizar um modelo já treinado no *corpus* específico.

Consideramos que este trabalho contribuirá com o projeto CORPTEXLIT, aumentando o número de textos que vai constituir os *corpora* deste projeto, e também para o aumento da acurácia do Aelius, que terá a oportunidade de etiquetar outros textos literários do séc. XIX e incluindo alguns do início do séc. XX, corroborando o que pretende Alencar (2010a, p. 8) quando diz que: “o Aelius, aplicado por outros pesquisadores na anotação de textos de gêneros e épocas diversas, fornecerá novos materiais para estudos diacrônicos, dialetológicos ou literários baseados em *corpora*”.

Além do mais, gera outra “possibilidade de aumentar a acurácia do etiquetador ao diminuir erros que envolvem contextos sintáticos claramente identificáveis, por meio da inclusão de regras de reetiquetagem elaboradas manualmente” (CARVALHO; VASCONCELOS, ALENCAR, 2012, p. 133).

Como exemplo de anotação morfossintática mostramos o seguinte trecho do capítulo 01 de *A Conquista,* realizado pelo AeliusHunPos.

Exemplo 1

*A/**D-F** manhã/**N** tépida/**ADJ-F** ,/**,** rosada/**VB-AN-F** e/**CONJ** ressoante/**ADJ-G.***

As *tags* (em negrito, grifo nosso) correspondem a: D-F (artigo definido feminino); N (substantivo); ADJ-F (adjetivo feminino); VB-AN-F (particípio passivo feminino); CONJ (conjunção); ADJ-G (adjetivo singular, gênero comum de dois). As *tags* utilizadas tem como referência o manual de etiquetas do Tycho.

2.5 Compilação de um *Corpus*

A compilação do *corpus* consiste na criação em si do *corpus*, e está relacionado à “escolha dos textos do banco de dados que o comporão” (BERBER SARDINHA, 2004, p. 146).

Existem fases distintas no processo de compilação de um *corpus*, que se configuram em etapas metodológicas para sua compilação. Quando se trata de compilar um *corpus* próprio, que não um já compilado anteriormente, cumprem-se as seguintes etapas (ALUÍSIO; ALMEIDA, 2006):

a) seleção de textos, considerando-se os requisitos apresentados para a compilação de um *corpus*. Nesse caso observa-se o tipo de texto inerente à pesquisa proposta (para este trabalho, selecionamos textos literários), tamanho e modalidade (se oral ou escrito);

b) compilação e manipulação do *corpus,* que consiste primeiramente em armazenar/arquivar os textos selecionados. Em algumas situações é possível encontrar os textos já digitalizados e disponíveis na *WEB*. No caso de textos impressos, é necessária a sua digitalização. Neste trabalho serão utilizados tanto textos da *WEB* quanto textos digitados, tendo como referência o original do autor. Já a manipulação consiste na tarefa de converter manual e ou automaticamente o formato dos textos em “doc”, “html”, “pdf” para “txt”. A etapa seguinte é a limpeza e formatação 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). O texto limpo permite que os processadores/ferramentas computacionais realizem as ações necessárias ao uso do *corpus*;

c) nomeação de arquivos feitas após conversão dos textos no formato “txt”. Escolhe-se um padrão de nomes de arquivos que facilitem a identificação dos processadores no momento de utilização dos arquivos para análise;

d) pedidos de permissão de uso – essa é a parte da constituição de um *corpus* que não é técnica. Demanda muito tempo para se obter os direitos de uso de material (com autores e editores), que é livre: alguns *corpora* não estão disponíveis para uso público justamente porque esses processos às vezes se arrastam por anos. O material utilizado em nosso trabalho já é de domínio público: portanto, de livre acesso e uso.

A compilação de um *corpus* na Linguística de *Corpus* envolve tarefas complexas e minuciosas para que o resultado seja qualitativo e real.

Quanto à importância da criação/compilação de um *corpus*, Biderman (2005 *apud* ALUÍSIO, 2007), quando da criação do Projeto do Dicionário Histórico do Português do Brasil (DHPB), constituído de textos do séc. XVI a XVIII, relata que: “a criação do *corpus* informatizado que estamos gerando e construindo tem uma

importância vital para as pesquisas sobre o Português do Brasil e quase tão grande quanto o próprio dicionário que vamos produzir", corroborando ainda mais a importância da compilação de *corpora* informatizados.

A compilação e a manipulação, necessariamente dita, do *corpus* consiste ao final em gerar arquivos de cada texto selecionado. Esse momento é de suma importância, considerando-se que o arquivo organizado em um padrão coerente permitirá recuperar as informações de forma mais rápida e eficaz quando do momento de execução dos comandos em Python. Todos deverão ser arquivados em formato "txt", pois as ferramentas de exploração de *corpora* suportam como entrada apenas documentos em texto puro (SILVEIRA, 2008). Os arquivos devem ser criados em letras minúsculas, pois Python faz distinção entre maiúsculas e minúsculas. E ao se trabalhar com maiúsculas e minúsculas, simultaneamente, dificulta-se a indexação e execução dos comandos.

Dentro da etapa de manipulação dos *Corpora*, estes devem passar também pelas etapas de edição e atualização, que consistem primeiramente em comparar o texto digitado com o texto original impresso. Considerando-se que são textos literários antigos, é importante que se proceda a essa etapa em conformidade com os procedimentos realizados com *corpora* dessa mesma natureza, como no caso do Tycho.

Quanto à edição, pesquisas atuais apontam as edições semidiplomáticas (que tem um grau de interferência maior) como aceitáveis, pois guardam a maior parte das características linguísticas do original, como as lexicais, morfológicas, e sintáticas, isso no sentido de gerar uma leitura facilitada no que se refere aos aspectos grafemáticos do leitor do texto, seja ele humano ou automático.

Em algumas situações, e dependendo da antiguidade dos textos, as características gráficas e grafemáticas dificultam o processamento automático, como no caso da anotação morfossintática. Sendo assim, precisariam de um grau de interferência mais elevado do que o feito em uma edição semidiplomática (PAIXÃO DE SOUSA; KLEPER; FARIA, 2012, p. 193): "os estudos linguísticos baseados em textos antigos dependem, antes de tudo, da garantia da fidelidade às formas originais dos textos".

No exemplo 2, mostramos um pequeno recorte representando a necessidade de edição do texto:

Exemplo 2:

1 - "E não só história dos livros, outras sabia que eu jamais em letras vira: a que descrevia a ***vara*** *branca seduzindo o remador do Itapicuru e o conto do sucupira" [...] Nas casas dos escravos,* ***às vezes****, à noite, ensaiavam as crianças." (*trecho do texto disponível na *WEB* de *Firmo, o vaqueiro* – grifo nosso*).*

2 - *"E não só história dos livros, outras sabia que eu jamais em letras vira: a que descrevia a* ***iara*** *branca seduzindo o remador do Itapicuru e o conto do sucupira" [...] Nas casas dos escravos* ***as velhas****, à noite, ensaiavam as crianças."* (trecho editado conforme original impresso de 1926 - grifo nosso).

Esse exemplo demonstra a necessidade de edição dos textos disponíveis na *WEB*, pois dependendo do processo de digitalização ou de reconhecimento óptico de caracteres pelo OCR, muitos caracteres podem ficar corrompidos, gerando outra palavra, expressão ou pontuação, interferindo inicialmente na quantidade de *tokens* e, posteriormente, na categorização das etiquetas e nos resultados das análises.

Palavras digitadas erroneamente, supressão de palavras e de expressões, substituição de maiúsculas por minúsculas, troca de palavras são algumas das deficiências encontradas nos textos digitados e livremente disponíveis na *WEB.* É importante que seja feita essa edição porque o resultado influenciará sobremaneira no momento de tratamento computacional do *corpus,* no sentido de não se obter os dados solicitados.

Quanto à atualização dos textos limpos e editados, está relacionada com a uniformização dos textos de acordo com a norma ortográfica vigente. Deve-se considerar a ortografia das palavras e não suas especificidades linguísticas ou arcaísmos, preservando ao máximo suas características originais. Estas etapas devem ser todas feitas manualmente, para depois gerar os arquivos específicos. Vejamos um exemplo de atualização:

Exemplo 3:

1 - "Outubro. O sol, em pleno, alargava por todo o campo uma luz fixa e caustica. Não havia sombra – tudo resplandecia e um tedio morno, pesado, de preguiça, parecia apoderar-se das ***proprias*** *coisas, prendendo-as em* ***immobilidade****, de onde nem mesmo o bulir das folhas tirava o* ***dôce murmurio****, tão* ***agradavel*** *ao ouvido de quem trabalha sob a rude prancha da soalheira estiva"* (trecho do conto *O Enterro*, do livro Sertão – texto original do escritor – grifo nosso).

2 - "*Outubro. O sol, em pleno, alargava por todo o campo uma luz fixa e caustica. Não havia sombra – tudo resplandecia e um tedio morno, pesado, de preguiça, parecia apoderar-se das* ***próprias*** *coisas, prendendo-as em* ***imobilidade****, de onde nem mesmo o bulir das folhas tirava o* ***doce murmúrio****, tão* ***agradável*** *ao ouvido de quem trabalha sob a rude prancha da soalheira estiva*" (trecho do conto *O Enterro*, do livro Sertão – texto atualizado – grifo nosso).

Após o processo de limpeza, edição e atualização, os textos/arquivos estão preparados para o processo de anotação morfossintática ou outro tratamento computacional que importa ao pesquisador.

2.6 Anotação e análise de um *Corpus*

Em uma espécie de "apelo" à comunidade linguística para que se criem ferramentas computacionais que permitam aos linguistas trabalhar com grandes quantidades de dados, Galves (2003), destacando a importante interface da linguística com a computação, diz que:

> Não adianta ter vários milhões de palavras e não ter como extrair a informação desses milhões de palavras, o que à mão se torna totalmente impossível. Então, precisamos ter uma anotação, e precisamos de ferramentas automáticas de anotação, porque anotar à mão milhões de palavras é uma tarefa inglória.

A anotação permite que um *corpus* seja muito mais útil, podendo ser usado para pesquisa tanto por parte do pesquisador que realizou a anotação, mas também para outros que possam torná-lo útil para seus propósitos. Anotar é dar "mais-valia" para o *corpus* (LEECH, 2004), agregando valor a um *corpus* à medida que amplia consideravelmente o leque de investigação e perguntas que um *corpus* anotado pode dar facilmente as respostas.

Porquanto, a anotação pode ajudar em muitos tipos de processamento automático e análises; em gerar listas de frequências com a classificação, com vistas a permitir desambiguação de muitos termos iguais, mas com sentidos diferentes; a análise sintática em si; e para auxiliar no caso do processamento sintético de voz (LEECH, 2004).

Um *corpus* pode ser passível de anotação estrutural e anotação linguística (ALUÍSIO; ALMEIDA, 2006). A anotação estrutural se refere à "marcação de dados externos e internos dos textos" (p. 161). São dados que não compõem o

texto propriamente dito, como: autoria, tamanho, arquivo, capítulos, parágrafos, gráficos, notas etc. Já a anotação linguística "se divide em anotação morfológica ou gramatical, anotação sintática, semântica, e de discurso" (VIEIRA; STRUBE DE LIMA, 2001, p. 35).

O nível morfológico ou gramatical cumpre a função de:

> 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. Para as palavras da língua utiliza-se uma etiqueta referente à sua categoria morfossintática ou gramatical (adjetivo, verbo, substantivo, etc.); para os símbolos de pontuação aceitos no discurso (por exemplo, vírgula, ponto final, exclamação, interrogação, etc.) geralmente utiliza-se o próprio símbolo (AIRES, 2000, p. 8).

Leech (2004) relaciona seis tipos de anotações realizadas pelos etiquetadores: a) anotação fonética; b) anotação semântica; c) anotação pragmática; d) anotação do discurso; e) anotação estilística; f) anotação lexical. Ele compreende também que "é possível pensar em incontáveis tipos de anotações que podem ser úteis para tipos específicos de pesquisa"[32].

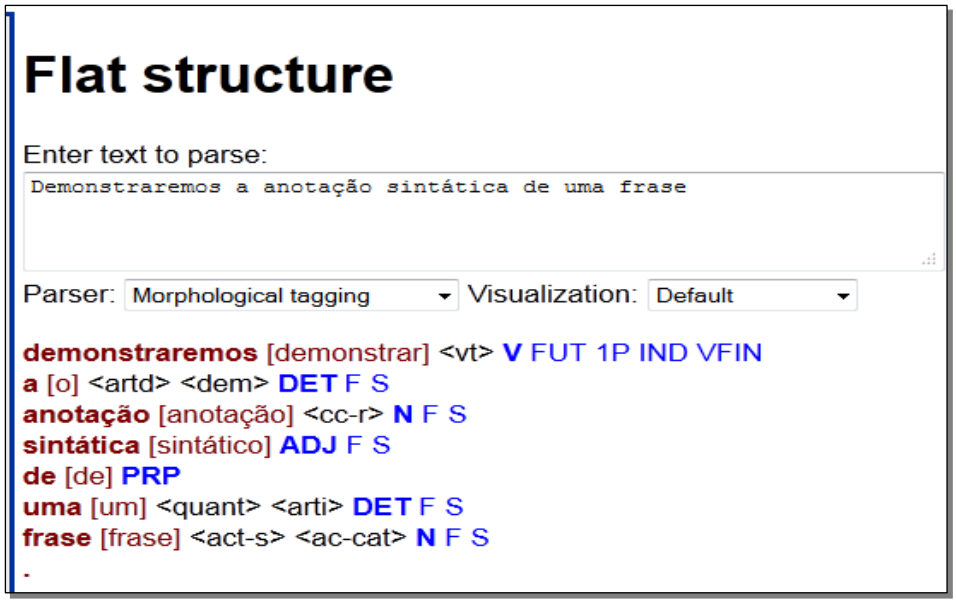
Para Berber Sardinha (2004) a anotação do *corpus* consiste na inserção de cabeçalhos informativos nos arquivos e na sua etiquetagem, seja morfossintática, sintática ou semântica.

Como exemplo de etiquetador destacamos o VISL, que é atualmente a interface livre na *WEB* mais eficiente para etiquetagens morfológicas, sintáticas e semânticas.

Para a demonstração de anotação realizada pelo VISL, segue o exemplo com a seguinte frase: "Demonstraremos a anotação sintática de uma frase".

[32] "In fact, it is possible to think up untold kinds of annotation that might be useful for specific kinds of research" (tradução nossa).

Figura 4 – Interface de anotação morfológica com base no VISL – modelo *Flat structure.*



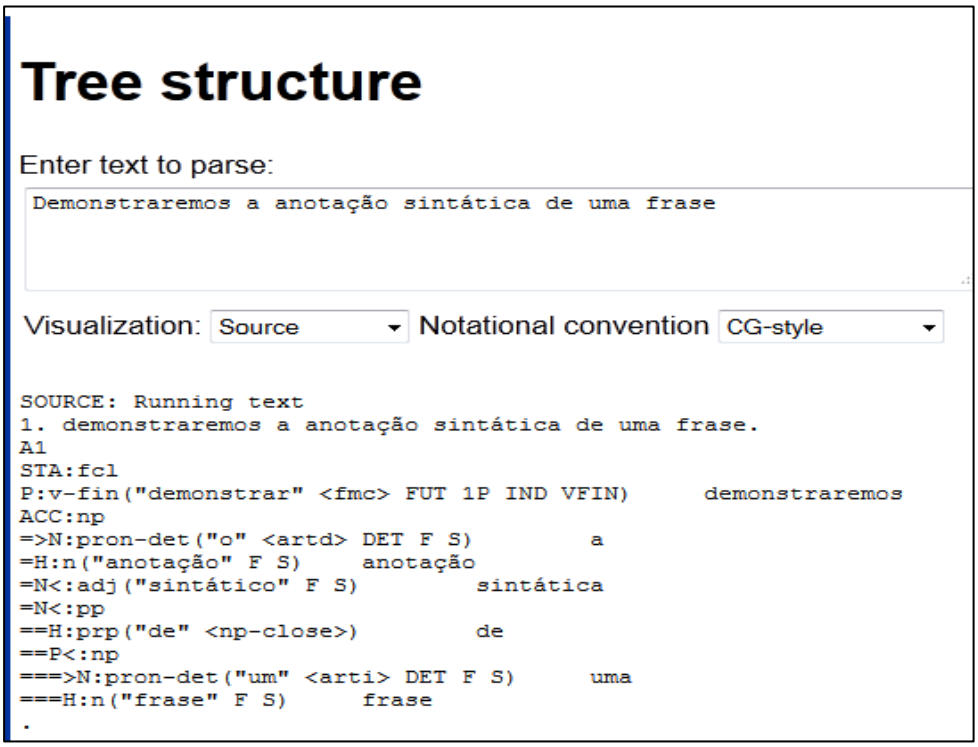
Fonte: Elaborada pela autora.

As etiquetas (*tags*) que representam as anotações morfológicas deste exemplo, conforme o VISL[33] são as seguintes: V FUT (verbo finito futuro), DET F S (artigo feminino singular), N F S (nome feminino singular), ADJ F S (adjetivo feminino singular), PRP (preposição). As classes de palavras estão destacadas na figura em azul “negrito”, as outras informações são referentes às flexões das palavras.

A mesma frase foi utilizada para a anotação, neste caso, sintática, na estrutura de uma árvore, visto na figura 5.

[33] *Tags* morfológicas - http://beta.visl.sdu.dk/visl/pt/info/portsymbol.html#syntags

Figura 5 – Interface de anotação sintática com base no VISL – modelo *Tree structure*.



Fonte: Elaborada pela autora.

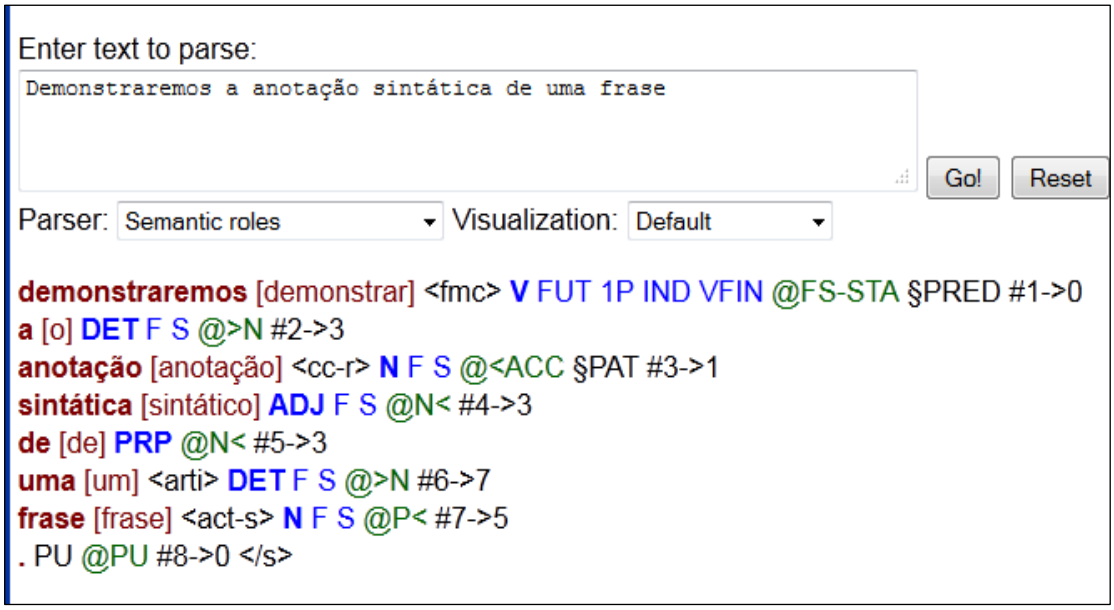
As etiquetas que representam as anotações sintáticas deste exemplo, conforme as *tags* sintáticas do VISL[34] são: STA (enunciado declarativo), P (predicador), ACC (objeto direto acusativo), >N (adjunto pré-nominal), H (núcleo), N< (adjunto pós-nominal). Estas etiquetas correspondem às que estão em maiúsculas na árvore acima.

A anotação semântica considera "o significado das expressões linguísticas de maneira independente de quem as usa ou de como são usadas" (VIEIRA; STRUBE DE LIMA, 2001, p. 2). Nesse nível de anotação "são considerados somente os substantivos, entidades nomeadas e alguns adjetivos, para os quais é possível atribuir um valor semântico" (TOMAZELA; BARROS; RINO, 2008, p. 32).

Na figura 6, demonstramos um exemplo de anotação semântica, tendo a mesma frase para análise.

[34] *Tags* sintáticas - http://beta.visl.sdu.dk/visl/pt/info/portsymbol.html#syntags

Figura 6 – Interface de anotação semântica com base no VISL – modelo *Flat structure*.



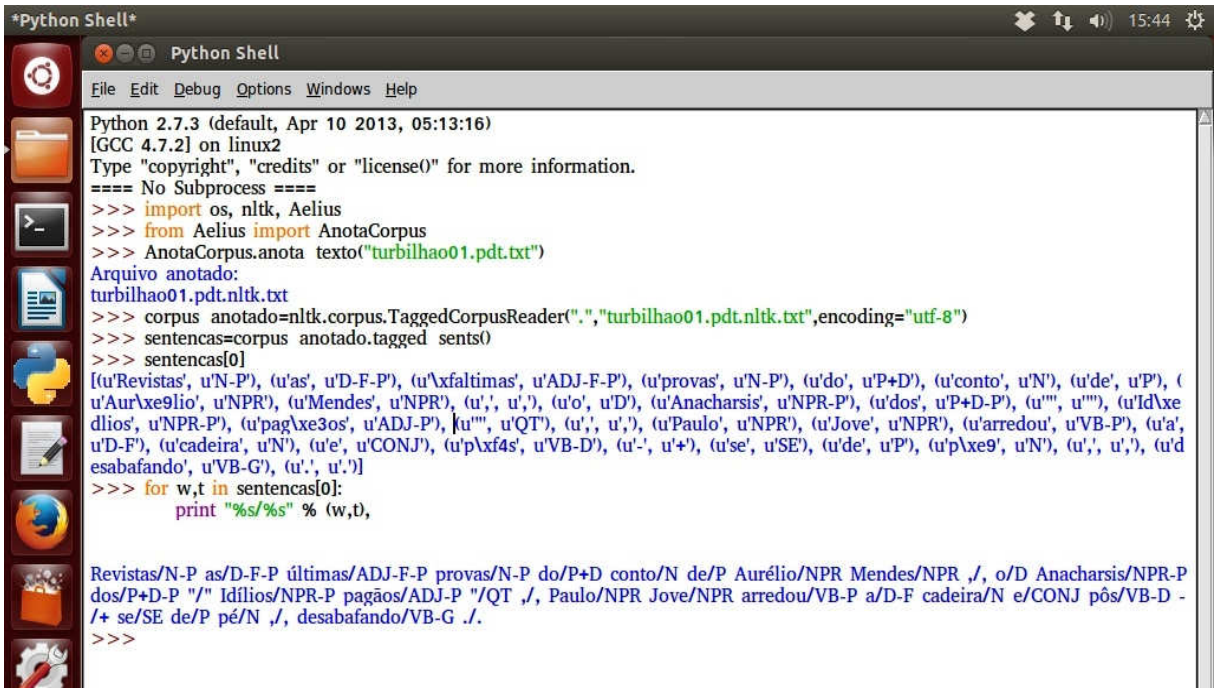
Fonte: Elaborada pela autora

Nesse caso, o VISL só atribuiu duas etiquetas, que representam: § PRED (predicador), § PAT (paciente). Essas etiquetas correspondem às que estão em preto depois do símbolo §. Esse modelo, *Flat structure,* faz ao mesmo tempo todas as anotações (morfológica – em azul, sintática – em verde, e semântica – em preto).

As figuras 7, 8 e 9 mostram exemplos de anotação morfossintática realizada por alguns etiquetadores já citados no item 2.4, utilizando o mesmo texto do livro *Turbilhão*, capítulo 1: “Revistas as últimas provas do conto de Aurélio Mendes o Anacharsis dos ‘Idílios pagãos’, Paulo Jove arredou a cadeira e pôs-se de pé, desabafando” (NETTO, C., 1958, p. 829).

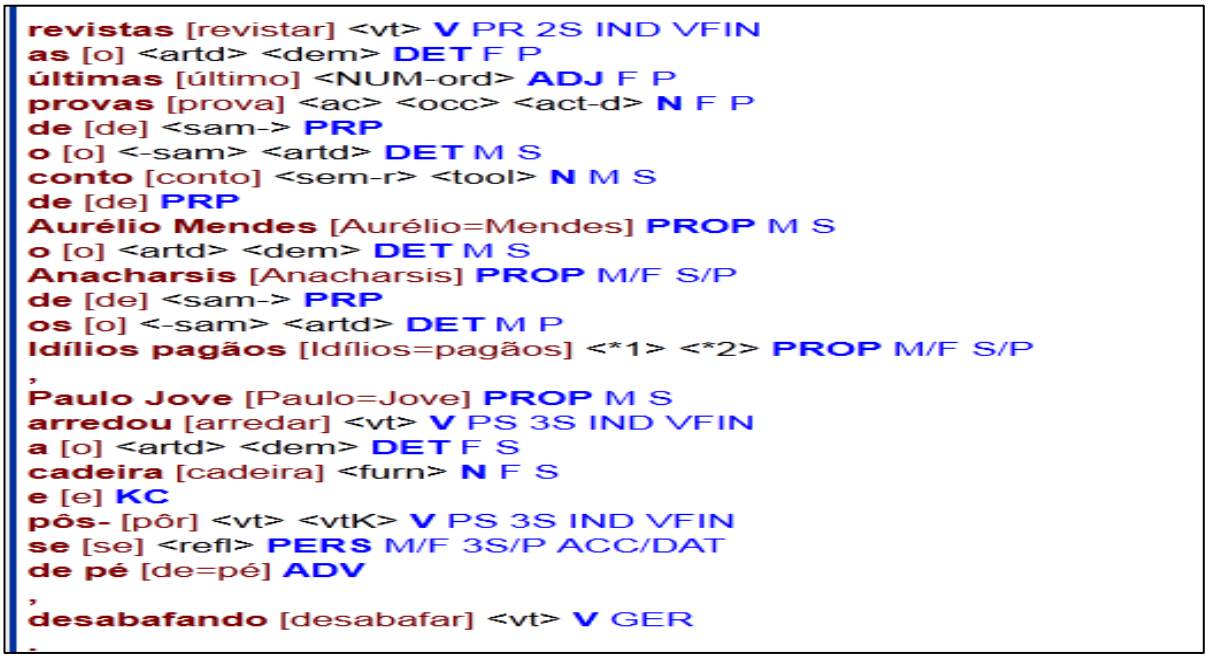
A demonstração de anotação desses etiquetadores tem o intuito de exemplificarmos as possibilidades que temos no que se refere a etiquetadores *opensource.* Optamos por utilizar nesse trabalho *o* etiquetador Aelius, porque tem uma interface mais amigável e trabalha somente com análise morfossintática, e não utiliza a CG como base para suas anotações.

Figura 7 - Trecho anotado pelo etiquetador morfossintático Aelius.



Fonte: Elaborada pela autora

Figura 8 - Trecho anotado pelo etiquetador do VISL.



Fonte: http://beta.visl.sdu.dk/visl/pt/parsing/automatic/parse.php

Figura 9 - Trecho anotado pelo etiquetador do LX-Center.

<s> Revistas/REVER,REVISTA/PPA#fp as/DA#fp últimas/úLTIMO/ADJ#fp provas/PROVA/CN#fp de_/PREP o/DA#ms conto/CONTAR/V#pi-1s de/PREP Aurélio/PNM Mendes/PNM o/DA#ms Anacharsis/PNM de_/PREP os/DA#mp *"/PNM Idílios/PNM pagãos/PAGãO/CN#mp "/"/ADJ#mp ,*//PNT Paulo/PNM Jove/PNM arredou/ARREDAR/V#ppi-3s a/DA#fs cadeira/CADEIRA/CN#fs e/CJ pôs/PÔR/V#ppi-3s -se/CL#gn3 de/PREP pé/Pé/CN#ms ,*//PNT desabafando/DESABAFAR/V#GER </s>

Fonte: http://lxcenter.di.fc.ul.pt/services/pt/LXServicesSuitePT.html

No exemplo da anotação na figura 7, podemos observar uma das principais dificuldades enfrentadas pelo etiquetador - o problema de ambiguidade relacionado ao contexto (polissêmico) em que se encontra a palavra (DOMINGUES; FAVERO; MEDEIROS, 2008). A ambiguidade "ocorre devido às diferentes categorias gramaticais que um item léxico pode receber em diversos contextos" (DOMINGUES, 2011, p. 15), busca-se, portanto solucionar o problema, em contexto específico, quando o etiquetador atribui a categoria gramatical apropriada.

Nunes *et al* (2005, p. 37) entendem por léxico como "um conjunto bastante abrangente e heterogêneo de unidades lexicais, contendo inclusive nomes próprios, siglas, abreviaturas", todas associadas a uma série de traços morfossintáticos, considerados na anotação morfossintática automática dos *Corpora.*

O exemplo 4 detalha melhor a situação de ocorrência de ambiguidade no momento da anotação:

Exemplo 4:

*F1 = "**Revistas** velhas estão espalhadas pelo chão" (frase criada aleatoriamente).*

*"**Revistas/N-P** velhas/ADJ-F-P estão/ET-P espalhadas/VB-AN-F-P pelo/P+D chão/N" (texto anotado automaticamente).*

*F2 = "**Revistas** as últimas provas do conto de Aurélio Mendes" (trecho extraído do romance Turbilhão)*

***"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" (texto anotado automaticamente, figura 7).*

Nesse caso, o etiquetador no modelo AeliusHunPos, não distingue corretamente a classe de palavras contextualizada na sentença F2 (como mostrado na Figura 8). Ele anota a palavra (destaque acima em negrito) como Substantivo (N),

cometendo ambiguidade ao interpretar a palavra como outra categoria morfológica: a anotação correta seria Verbo (VB).

Em F1 a palavra é categorizada como substantivo (Nome), fazendo uma correta análise morfológica. Situações como essa, de ambiguidade, são resolvidas com correção manual e alimentação de dados no Aelius para posterior correção da ambiguidade: implementa-se em Python um algoritmo que possibilite distinguir as classes de palavras e anotá-las corretamente.

Apesar de erros como esses, o etiquetador Aelius tem mais probabilidade de “resolver” o problema da ambiguidade por ser híbrido, congregando a abordagem baseada em regras e a probabilística: “as abordagens híbridas exigem um menor esforço na geração de regras e funcionam melhor que a aplicação de uma só abordagem” (DOMINGUES, 2011, p. 56). Os etiquetadores híbridos são centrados em algoritmos de Aprendizado Baseado em Transformação Dirigida por Erro (TBL) desenvolvido por Eric Brill em 1995, mais particularmente esse etiquetador é usado em análise sintática e em etiquetagem morfossintática.

Bick (2005, p. 91) discutindo sobre a ambiguidade dos etiquetadores, diz que: “A maioria das palavras em textos de língua natural são – quando vistas em isoladamente – ambíguas quanto à classe de palavra, flexão, função sintática, conteúdo semântico etc”.

Jurafsky e Martin (2009) exemplificam uma ambiguidade na anotação quando uma palavra é nome em vez de verbo, se ela é precedida de um artigo (a casa/casa). Essas ambiguidades podem ser “resolvidas” de duas maneiras: de forma manual (o usuário edita os grafos correspondentes) e de forma automática por meio de uma gramática de desambiguação definida pelo sistema reprogramado (NUNES *et al*, 2005).

As etiquetas (*tags*) podem ser inseridas nas mais diversas formas: /subs; <subs>; -subs; /CN etc. Uma mesma categoria gramatical pode ser representada por várias *tags,* por exemplo: o verbo pode ser etiquetado como verbo auxiliar, verbo ser, verbo auxiliar, verbo gerúndio (ver Apêndice H). A categoria artigo também comporta várias etiquetas: no caso do etiquetador Aelius, treinado no *Corpus* Tycho Brahe, anota da seguinte forma:

Quadro 4 - Exemplos de etiquetas da categoria gramatical “artigo”

ARTIGO	ETIQUETA (TAG)
O	D
A	D-F
OS	D-P
AS	D-F-P
UM	D-UM
UMA	D-UM-F

Fonte: Elaborado pela autora

As *tags* mudam conforme o modelo do etiquetador, como exemplificado no quadro abaixo:

Quadro 5 - Exemplos de etiquetas conforme o etiquetador

CLASSE	ETIQUETADORES/ETIQUETAS		
	LxTagger	Aelius	Eagles
Substantivo	CN	N	N
Verbo	V	VB	V
Adjetivo	ADJ	ADJ	AJ
Advérbio	ADV	ADV	AV
Conjunção	CJ	CONJ	C
Artigo	DA	D	AT

Fonte: Elaborado pela autora

O processo de etiquetação é dividido em duas fases: treinamento, corresponde à aquisição do modelo linguístico, geralmente a partir de um *corpus* anotado; e teste, quando o algoritmo da etiquetação é aplicado a uma parte reservada do *corpus* para avaliação do aprendizado.

Cada etiquetador morfossintático (*tagger*) tem sua qualidade avaliada a partir, principalmente, de sua precisão geral ou acurácia (*accuracy*) que está relacionada com “o número de palavras classificadas corretamente dividido pelo número de palavras do seu arquivo de teste” (AIRES, 2000, p. 46). Essa ação é formatada como comando que o processador executa devolvendo os resultados.

As anotações são realizadas de forma automática, utilizando ferramentas de PLN, e sempre necessitam de intervenções manuais pelos linguistas no caso de corrigir as distorções de anotação feita pelos processadores, que depois retornam para a correção automática, no intuito de gerar dados concretos para cada vez mais melhorar a acurácia dos etiquetadores.

As aplicações em PLN e LCLC dependem do alto desempenho dos etiquetadores em termos de acurácia. Estudos atuais têm a preocupação de

melhorar cada vez mais o desempenho destas ferramentas, o que representa ainda um desafio para o sucesso de aplicações nessas áreas. Observa-se, portanto, que o problema da etiquetagem é bastante relevante, pois “qualquer pequena melhora em termos de acurácia pode representar uma diferença significativa na aplicação final” (DOMINGUES, 2011, p. 18).

A anotação *POS-tagging* pode ser feita automaticamente, com precisão entre 95 a 98%. Não é o ideal, e para isso requer um laborioso trabalho manual de correção. No caso de uma melhor precisão, que seria de 100%, a correção manual pode atingir talvez entre 99% a 99,5%, considerando-se alguns casos pouco claros ou de ambiguidade que podem surgir em relação à correta anotação (LEECH, 2004). Essa tarefa, claro, aperfeiçoará *a posteriori* o etiquetador quando este se “enriquecer” de análises mais concretas.

O estudo de Domingues (2011) sobre o desenvolvimento de um etiquetador de alta acurácia, diz que as ferramentas de etiquetagem que alcançam entre 96 e 97% de acurácia em textos do gênero jornalístico são consideradas como de alta acurácia. Para textos do gênero literário, o estado da arte relata acurácia de 95%, que com aprimoramentos do etiquetador, obteve melhor performance acima desse percentual: 97,4% (ALENCAR, 2010a, 2011a, 2013).

O *Entropy Guided Transformation Learning* (ETL), ferramenta de aprendizagem em máquina que combina DT (*Decision Trees*) e TBL, utilizando o Tycho para treino, obteve bons resultados com precisão de 96,6% (DOS SANTOS; MILIDIÚ; RENTERIA, 2008).

A utilização, portanto de um *corpus* anotado é imprevisível, pois os “usos futuros são sempre mais variáveis do que o criador do *corpus* poderia ter imaginado” (LEECH, 2004, p. 28).

O processo de anotação também prevê normas úteis, antes de tudo deve ser bem planejado e executado (LEECH, 2004), e dentre algumas citamos:

1º. As anotações devem ser separáveis – é uma ação opcional pra um *corpus*. Deve-se sempre separar as anotações do *corpus* bruto, que pode ser recuperado a qualquer momento sem anotações;

2º. A documentação deve ser detalhada e explícita – seus constituintes léxicos devem ser bem explícitos, detalhando: como foram anotados (ferramentas informáticas); quais os esquemas de codificações (quais *tags* foram utilizadas) e a precisão do etiquetador.

Para testar a precisão do etiquetador e a fidelidade do texto anotado quanto às *tags*, é preciso ter, no mínimo, dois anotadores humanos pós-anotação, de um mesmo texto anotado, para determinar os casos que um concorda com o outro, e com isso melhorar a qualidade da anotação pós-correção. Cada anotador deve escolher a análise que considera correta. Em caso de divergência entre anotadores na correção manual, há que se chamar outro corretor (uma terceira pessoa) para decidir qual a correção correta: “este processo de anotação garante grande confiabilidade nas informações geradas” (GONÇALVES; BRANCO, 2010, p. 466).

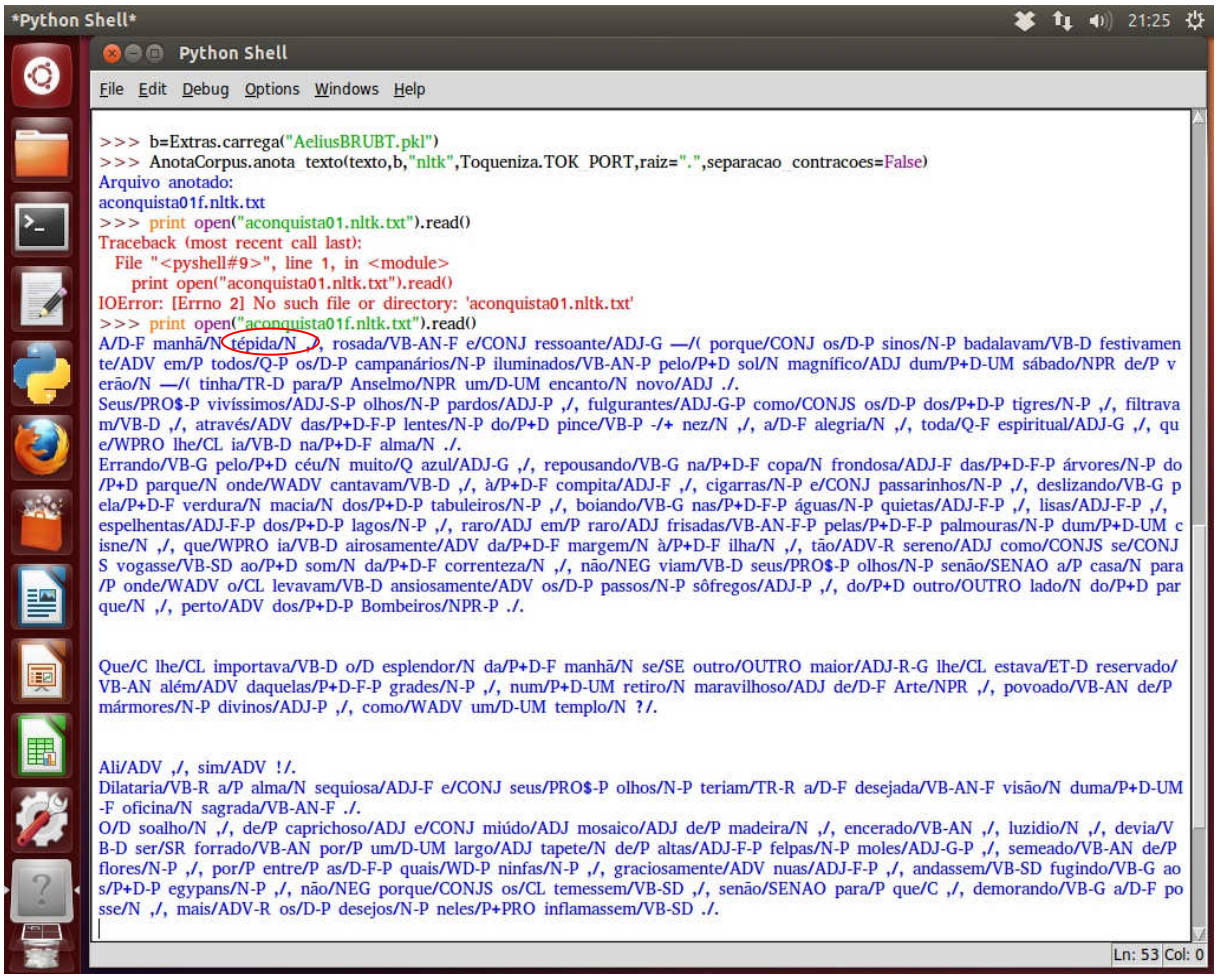
A anotação demanda um tempo muito extenso do pesquisador, mas como é feita de forma automática e considerando-se o tamanho do *corpus* a tarefa pode ser mais rápida. Quanto à correção manual da anotação morfossintática automática, esta requer um tempo maior que demanda, sobretudo, financiamento específico, pois envolve muitos elementos para a consecução desta atividade, que passa por várias etapas, com tecnologia e recursos humanos especializados: isso, claro, vai depender do tamanho do *corpus*.

Para análise de precisão de um etiquetador é necessário dois *corpus* distintos: um *corpus* de teste (utilizado para determinar a precisão do etiquetador) e um *corpus* de treino (utilizado no treino do etiquetador). Esse processo contribuirá para gerar um etiquetador robusto, com desempenho além do já obtido nas testagens de etiquetadores e modelos próprios para textos literários, como o caso do AeliusHunPos.

Esta pesquisa também oportunizará verificar a acurácia do anotador morfossintático Aelius, no modelo AeliusHunPos com textos literários. As análises realizadas por Alencar (2010a) quanto à acurácia do Aelius em textos literários no projeto CORPTEXLIT, com o AeliusBRUBT, detectou uma acurácia de 95,08%. Intenta-se então verificar se a anotação do CCN demonstrará a mesma acurácia ou desempenho superior. Pretende-se, com essa avaliação, colaborar para implementar ainda mais o etiquetador Aelius para que este possa aumentar seu desempenho e se tornar um anotador padrão *gold* na anotação de textos literários dos séc. XIX e XX. Alencar (2011b) constatou que ao se utilizar o HunPos ganhou-se quase 1% de acurácia em relação ao BRUBT em textos anotados sem correção manual.

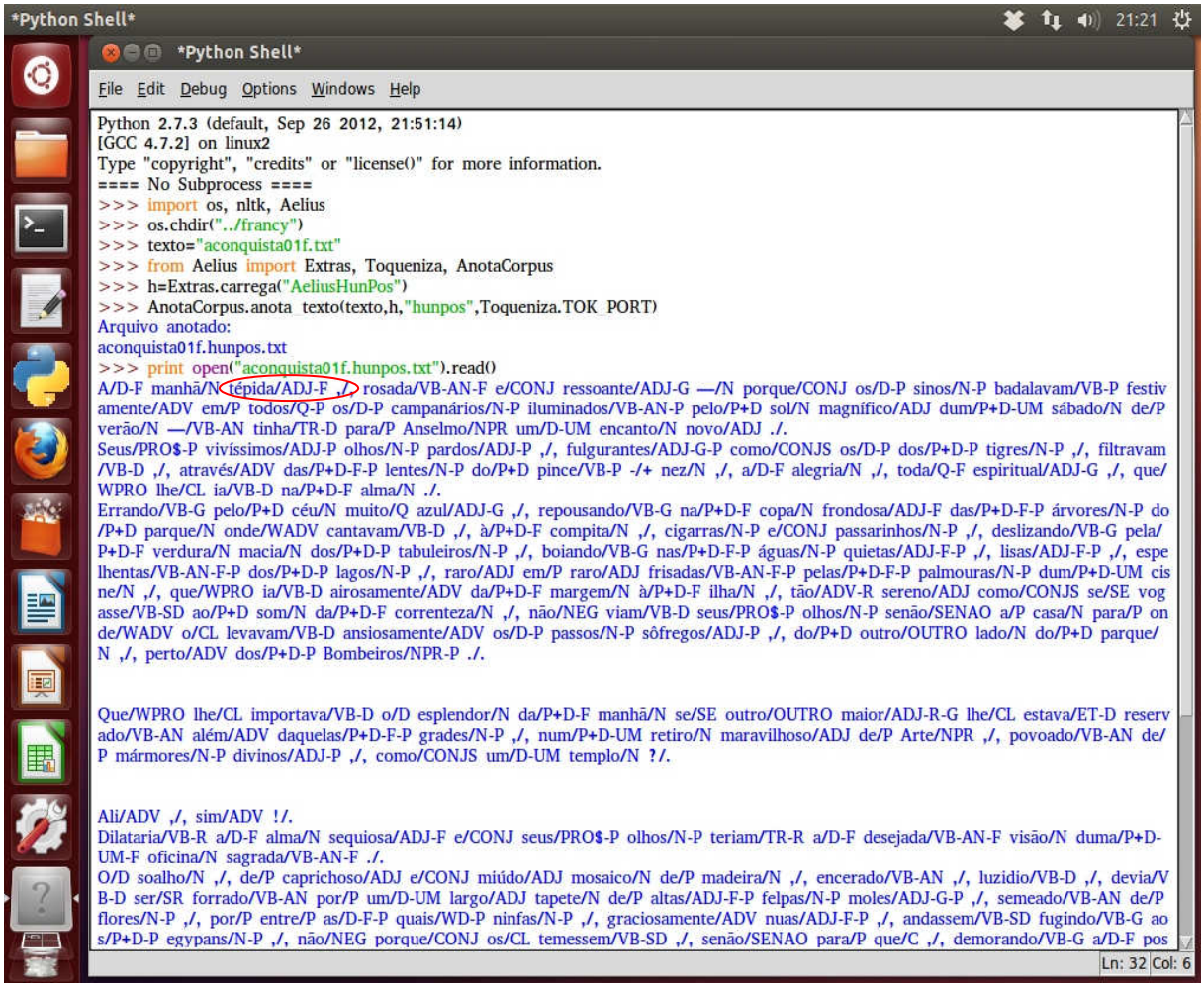
Nas figuras 10 e 11, vemos a demonstração de anotação realizada pelo AeliusBRUBT e AeliusHunPos.

Figura 10 - Anotação do trecho de *A Conquista* utilizando o AeliusBRUBT treinado no *Corpus* Tycho Brahe



Fonte: Elaborada pela autora

Figura 11 - Anotação do trecho de *A Conquista* utilizando o modelo treinado pelo *software* HunPos treinado no *Corpus* Tycho Brahe

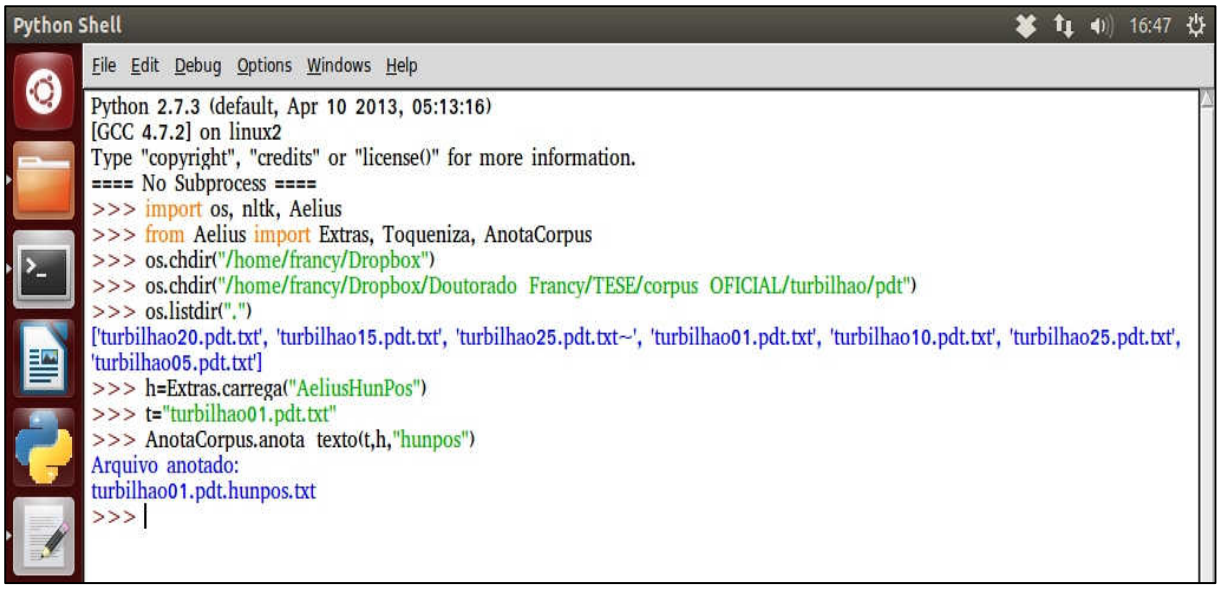


Fonte: Elaborada pela autora

Observa-se nas anotações dos modelos AeliusBRUBT (figura 10) e AeliusHunPos (figura 11) a diferença de etiquetação realizada pelos anotadores. Nas figuras, é possível observar que a palavra “tépida” (circulada de vermelho) no modelo AeliusBRUBT foi anotado como N (nome), no AeliusHunPos foi anotado como ADJ-F (adjetivo feminino), corretamente.

Interessante destacar que Python permite que comandos diversos sejam criados para a realização da anotação morfossintática. Na figura 11 temos um formato de anotação, na figura 12 temos outro, com comandos diferenciados, mas que executam a mesma função, como vemos a seguir.

Figura 12 – Capítulo anotado pelo etiquetador morfossintático Aelius.



Fonte: Elaborada pela autora

Na figura 11, a anotação é exposta na shell de Python, podendo ser visualizada imediatamente. Na figura 12, o comando gera um arquivo que pode ser acessado em outro local. A geração de arquivos é o mais indicado, considerando-se que com a anotação realizada os dados gerados são muitos, e expô-los na shell de Python demandaria muito espaço, e seria difícil sua manipulação.

Isso importa também para processamentos futuros, quando somente se vincula o arquivo aos novos comandos, gerando outros dados, que possibilitarão as análises linguístico-computacionais que serão realizadas neste trabalho.

Conhecer as probabilidades de ocorrência dos eventos linguísticos é importante para o desenvolvimento de sistemas computacionais e para o entendimento da linguagem, por isso a LCLC ser importante nesse processo, pois é possível verificar com maior precisão essas ocorrências em análises futuras (BERBER SARDINHA, 2005a).

Santos (2008, p. 54) discutindo sobre as informações que as análises linguísticas em um *corpus* podem trazer, cita como exemplos “o facto de um adjectivo ser usado como nome, o facto de uma dada construção acontecer mais frequentemente com um dado tempo verbal, ou a constatação de que um verbo co-ocorre maioritariamente com participantes humanos”.

As possibilidades de análise em um *corpus* são diversas. À medida que as análises são feitas outras surgem conforme a necessidade e objetivo pretendido

pelos pesquisadores. As mais ocorrentes dentre as experiências com LCORP nas investigações linguísticas são: concordâncias, lista de frequência de palavras e lista de colocações feitas sobre um *corpus* (SARMENTO, 2008; SILVEIRA, 2008). Essas análises podem contribuir para a melhoria no ensino e no aprendizado de uma língua, bem como sua compreensão. Cada área de especialidade necessita de metodologias de buscas diferenciadas.

A escolha de textos literários para as análises, além do já exposto na Introdução, remete-se, sobretudo, ao fato de que:

> [...] um texto literário pode ser comparado empiricamente a uma norma. Se essa norma é uma coleção de textos do mesmo período ou região ou se se trata da coletânea das obras de um determinado autor, pode-se usar a metodologia de linguística de corpus para interpretações mais demonstráveis e verificáveis (ZYNGIER, 2004, p. 44).

A frequência de palavras (ocorrências) em um *corpus* pode mostrar o contexto de certos itens lexicais utilizados por um determinado autor, permitindo afirmações fundamentais sobre possíveis interpretações, ou então comparar o uso que um autor faz de certas palavras a um conjunto de textos contemporâneos (ZYNGIER, 2004, p. 42): “a previsibilidade está associada à frequência de ocorrências”.

A riqueza vocabular de um escritor pode ser entendida da seguinte forma: “quanto maior o número de vocábulos novos, maior será a riqueza e a variedade do vocabulário” (CÚRCIO, 2013, 84). As *hapaxes* são, portanto, uma referência a essa riqueza.

Sobre estudos com advérbios em *–mente*, em textos literários destacamos:

- Moraes Pinto (2008) percebeu que no séc. XIX a presença desses advérbios diminui em comparação com os textos arcaicos. Em sua tese de doutorado observou as diferentes ordenações que os advérbios qualitativos e modalizadores em *-mente* assumem, e analisou a polissemia e o processo de gramaticalização desses elementos em textos escritos em português dos séculos XV a XX, afirmando que: “No português arcaico, houve maior variedade de itens adverbiais em *–mente*, demonstrando a alta produtividade desses elementos naquela época”, havendo uma gradativa diminuição de uso século a século, e à

primazia dos advérbios, na posição pós-verbal, a partir do século XIX (p. 38);

- Mattos e Silva (1989), também em seu estudo sobre as estruturas gramaticais do português arcaico, chama atenção para o fato de que os advérbios em *-mente* foram mais frequentes no *corpus* arcaico que outros advérbios;
- Hummel (2002, 2033) observou que a frequência dos advérbios em *–mente* em textos literários, brasileiros e portugueses, é bastante baixa, e seu uso resume-se na linguagem culta formal e com tendência mais forte em Portugal;
- Foltran (2010, p. 174) também observa que os advérbios em *–mente* são raros no português brasileiro “a ocorrência fica em torno de 6 a 8% do total de advérbios com essa forma”;
- Machado (2001) confirmou também a moderada frequência desses advérbios em textos literários. Ao analisar dez canções de Camões, no total de 1.793 palavras, o autor encontrou somente dezesseis desses advérbios (menos de 1% do total).

Entendemos que essas pesquisas revelam que a incidência acentuada de advérbios em *–mente* no período histórico do séc. XIX e sec. XX são relevantes de estudo, com destaque para os textos de Coelho Netto que como revela Alencar (2011a), a diversidade lexical apresentada no uso de advérbios em *–mente* é altíssima.

O maior destaque na busca dos advérbios em *–mente* nos textos de CN, não é para a quantidade de advérbios encontrados, mas para as *hapaxes*, ou seja, quantidade de advérbios que aparecem somente uma vez no texto, demonstrando a abundância lexical dos textos de Coelho Netto, e se de alguma forma ele “foge” aos padrões linguísticos do uso dessa classe de palavra nos textos da época.

Considerando-se a crítica de que Coelho Netto tinha “um anseio pela adjetivação numerosa, rara e erudita”, e era submisso ao culto do adjetivo, sendo até alcunhado como um “libidinoso da adjetivação”, neste trabalho buscamos as ocorrências dos adjetivos no CCN e a diversidade apresentada. Realizamos busca semelhante no *Corpus* de Contraste objetivando fazer um comparativo do uso de adjetivos por Aluísio Azevedo e Camilo Castelo Branco. Pretendendo observar, sobretudo, se CN era submisso ao adjetivo ou não.

Di Felippo e Dias-Da-Silva (2005, p. 3) relatam que:

> Os adjetivos ocupam lugar de destaque na exteriorização da visão de mundo do falante, pois parece ser por meio dos adjetivos que se manifestam com bastante clareza as opiniões do locutor (relacionadas à crença, valores, afetividades e registro do que ocorre no mundo exterior).

O processamento automático da língua permite estudos descritivos dos verbos em seus aspectos morfossintáticos-semânticos, pois podem apresentar diferentes comportamentos no uso da língua como abordado nas pesquisas de Borba (1996), Neves (1999), Scher (2004), Pacheco e Laporte (2013).

No estudo realizado por Machado (2001), sobe a obra de Camões, as maiores ocorrências foram de verbos (45%), em seguida substantivos e depois os adjetivos, com aproximadamente 15% de frequência.

Quanto ao uso de vernos nos textos de CN, intentamos constatar se a crítica ao escritor tinha respaldo, e se ele era de fato um “malabarista do verbo”.

No capítulo a seguir, abordamos os aspectos referentes à metodologia utilizada nesta Tese.

.

3 METODOLOGIA

Das etapas metodológicas para a compilação dos *Corpora*, a escolha dos textos é parte inicial deste trabalho. Os *Corpora* compreendem textos do Corpus Coelho Netto (CCN) e do Corpus de Contraste, com textos de Aluísio Azevedo e Camilo Castelo Branco, como detalhados a seguir.

Este item trata basicamente destacar os *Corpora* utilizados na tese, os instrumentos para tratados dos dados e os procedimentos utilizados.

3.1 Do *corpus*-amostra: o *Corpus* Coelho Netto (CCN)

Ao se projetar um *corpus* específico, visando extrair de sua observação generalizações sobre a língua, este serve a objetivos específicos, a análises pontuais, não se excluindo de participar de *corpora* mais extensos com pretensão de análises generalistas (BIDERMAN, 2001).

Dentre os tipos de *corpus* que Sinclair (1995), Tagnin (2010), Berber Sardinha (2000), McEnery e Wilson (2001) e Souza e Gatti (2002) definem, o CCN é caracterizado como resumido no Quadro 6:

Quadro 6 - Tipologia do *Corpus* Coelho Netto

Tamanho (representatividade e amostragem)	Pequeno (menos de 80 mil *tokens*)
Tempo	Histórico
Modo	Escrito
Seleção	De amostragem
Conteúdo	Especializado
Autoria	De língua nativa
Balanceamento	Por gênero
Gênero	Texto Literário
Finalidade	De estudo
Anotação	Anotado morfossintaticamente
Mutabilidade	Aberto
Disponibilidade	Na *WEB*

Fonte: Elaborado pela autora

É, conforme a tipologia definida, um *corpus* de estudo porque serve de base para a pesquisa desenvolvida, é o *corpus* que se pretende descrever e

analisar, e também serve para os propósitos de treinamento, quando no processamento da anotação morfossintática automática e posterior correção manual será utilizado para o desenvolvimento e a melhoria de desempenho de ferramentas de análises.

O CCN não é um *corpus* completo, mas sim um *corpus* que serve de referência para trabalhos nessa área, com amostras significativas tanto de textos anotados quanto de análises realizadas, vislumbrando nos pesquisadores outras possibilidades, quiçá interesse em tornar o CCN mais robusto. Outra característica importante é a de reutilização (ser aberto) de um *corpus,* quando se dispõe a servir a outras pesquisas, além da concebida inicialmente quando da constituição de outro *corpus*. De modo geral, sempre quando se trata do gênero literário, o *corpus* é constituído por amostragens dos títulos selecionados (BACELAR DO NASCIMENTO, 2002).

Definido pela literatura como *corpus* de conteúdo especializado, o CCN contém textos específicos do gênero literário, e foi compilado com um propósito já mencionado na Introdução e que será disponibilizado como um recurso da língua geral para linguistas, lexicógrafos, entre outros.

É denominado de "*corpus*-amostra" porque que se trata da compilação somente de capítulos 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, P., 1964), o *Corpus* que será compilado constitui-se somente como amostra de tudo que foi produzido pelo escritor.

O CCN compreende, parcialmente, dois romances: *A Conquista* (1899 – séc. XIX): capítulos I, VI, XII, XVIII, XXIII, XXVIII; e *Turbilhão* (1906 – séc. XX): capítulos I, V, X, XV, XX, XXV. E três contos do livro *Sertão* (1896 – séc. XIX): *Firmo, o vaqueiro; Mandovi* e *O Enterro.*

Estas obras foram escolhidas considerando-se que estes dois romances e os contos de *Sertão* são considerados como os textos mais lidos do escritor e os que tiveram mais tradução, conforme a crítica já mencionada no item 2.1.2. A opção pela escolha dos capítulos dos dois romances não obedeceu a nenhum critério condicionante. Optamos por dar um intervalo de cinco capítulos do livro para a escolha, mas isso não foi uma regra. No caso dos três contos escolhidos, dos sete que compõe o livro, levamos em consideração o tamanho do conto, pois teríamos

que digitar dois deles e se tivesse uma extensão maior demandaria mais mão-de-obra. A disponibilidade digital dos na *WEB* foi outro fator preponderante.

As principais informações sobre os textos que compõem o CCN foram detalhados no item 2.1.3.

O *Corpus* Coelho Netto contém 53.080 *tokens,* contados pós-edição dos textos digitais: um *corpus,* portanto, relativamente pequeno (BERBER SARDINHA, 2000), como detalhado na Tabela 1. Antes da edição, que incluiu comparação do texto digital com o impresso e as devidas correções, atualização ortográfica, e a uniformização automática de pontuação do texto em *.txt*, o *corpus* continha 52.877 *tokens*.

A diferença na contagem final de *tokens (*mais 203*),* nas etapas necessárias na compilação*,* revelam dados importantes para os devidos ajustes ao final da edição, isso para que seja gerado um *corpus* adequado e compatível com o processamento computacional, tanto para a anotação morfossintática quanto para as análises linguístico-computacionais posteriores.

Tabela 1 - Relação de quantidade de *tokens* pós-edição do CCN

Nº	**TEXTO**	***TOKENS* PÓS-EDIT**
1	aconquista01	5.860
2	aconquista06	2.708
3	aconquista12	3.734
4	aconquista18	2.463
5	aconquista23	3.414
6	aconquista28	3.982
7	turbilhão01	3.637
8	turbilhão05	4.991
9	turbilhão10	7.737
10	turbilhão15	3.737
11	turbilhão20	1.169
12	turbilhao25	2.033
13	oenterro	1.528
14	mandovi	4.219
15	firmo	1.868
	TOTAL	**53.080**

Fonte: Elaborada pela autora

Neste trabalho, o tamanho do *corpus* foi delimitado com a compreensão que tivemos com base em outros estudos de que como *corpus*-amostra para a produção de uma tese este seria suficiente. Sua dimensão poderá ser estendida se demandado em projetos de pesquisas destinados exclusivamente a criação e compilação de *corpora* robustos, como no caso do CORPTEXLIT. Biderman (2001, p. 80) discutindo sobre o tamanho dos *corpora* destaca o quão laborioso é o ato de se criar um grande *corpus*: “requer um esforço enorme – é trabalho longo, fadigoso”.

A compilação e anotação do CCN está também alinhado às preocupações do grupo temático “*Proteccion del patrimonio literario através de formatos digitales*” da Associação Internacional de Literatura Comparada (ICLA), que “pesquisa programas computacionais aplicáveis à digitalização de obras raras em geral” (PAIXÃO DE SOUSA, KEPLER, FARIA, 2012, p. 192), além claro de disponibilizar o CCN na *WEB* para análises linguísticas.

3.2 Do *Corpus* de Contraste

Na tipologia dos *corpora*, o *corpus* utilizado para comparação com o CCN é denominado de *Corpus* de Contraste, por ser “um *corpus* utilizado em comparação ao *corpus* de estudo, mas que não preenche as exigências (como tipo de linguagem e dimensão) para que seja considerado um *corpus* de referência” (SARMENTO, 2008, p. 123).

O *Corpus* de Contraste deve conter a quantidade de *tokens* equivalente à do *corpus* de estudo, para se obter um quantitativo verossímil em relação a algumas ocorrências buscadas, como: diversidade de palavras, diversidade e frequência de verbos, frequência de adjetivos e advérbios em *–mente* entre o *corpus* de estudo e o *corpus* de comparação.

Para a verificação da acurácia do etiquetador Aelius, utilizado neste trabalho para anotar o CCN e o *Corpus* de Contraste e realizar comparação entre eles, será corrigido manualmente 10% dos *Corpora* anotados. Nem sempre é possível igualar a quantidade de *tokens* de um *corpus* de estudo a um *corpus* de comparação, recomenda-se então que se aproxime o máximo possível o *corpus* de comparação, de preferência, para um quantitativo maior de *tokens*, como ocorreu no momento de seleção dos textos do *corpus* de comparação deste trabalho.

Enfim, para a comparação dos *Corpora*, este *corpus* deve obedecer a padrões de extensão que estão de acordo com o que se pretende na pesquisa. Quando se trata de observar diversidade e frequência em um *corpus*, a comparação com um outro só é possível quando o *corpus* de comparação for representativo, garantindo resultados relativamente estáveis quanto aos traços linguísticos observados.

Quanto ao *Corpus* de Contraste, os textos selecionados obedeceram aos critérios de: disponibilização digital na *WEB*, qualidade renomada da obra e serem da mesma época correspondente à época de produção literária de Coelho Netto. Os escritores escolhidos foram Aluísio Azevedo e Camilo Castelo Branco, com as obras *O Cortiço* e *Amor de Perdição*, respectivamente.

Portanto, para análise dos dados referentes à acurácia do anotador em comparação a outros textos literários do séc. XIX e XX e também das análises que serão realizadas, utilizamos: do romance *O Cortiço* (1890 – séc. XX): capítulos 01 a 08; de *Amor de Perdição* (1862 – séc. XX): capítulo de Introdução ao capítulo 06.

Os textos desses escritores foram escolhidos para servir de comparação com os de Coelho Netto, por um ser naturalista (PB) e o outro romântico (PE), sofreram da mesma crítica feita pelos modernistas: de utilizar linguagem muito rebuscada em seus escritos (BROCA, 1958). É importante frisar que o conceito de rebuscamento adotado provinha do entendimento dos modernistas à época, que não convém neste trabalho contestar ou comprovar.

Essa concepção de rebuscamento levava em consideração textos exemplificados na citação a seguir.

> 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, C., 1958, p. 829).

Podemos observar palavras que não são usualmente utilizadas e são pouco conhecidas, que os modernistas já consideravam como rebuscadas e presas às formas, dentre elas: estrincando, enclavinhados, tacões, clichagem, crebas. Isso sem mencionar o preciosismo nas construções frasais.

O *Corpus* de Contraste tem quantidade de *tokens* anotados (53.971) similar ao CCN (Tabelas 1 e 2), e segue a mesma tipologia do *Corpus* Coelho Netto (ver Quadro 1), para a avaliação do desempenho do etiquetador Aelius e para fazer comparação das análises pretendidas como: diversidade lexical, diversidade e riqueza de verbos e adjetivos (*hapaxes*), frequência de verbos em *–mente* dentre outras possíveis análises.

Tabela 2 - Relação de quantidade de *tokens* pós-edição do *Corpus* de Contraste

Nº	**TEXTO**	***TOKENS* PÓS-EDIT**
1	ocortiço01	6.088
2	ocortiço02	4.495
3	ocortiço03	4.950
4	ocortiço04	3.305
5	ocortiço05	2.251
6	ocortiço06	3.153
7	ocortiço07	6.345
8	ocortiço08	5.583
9	amordeperdiçãoINT	536
10	amordeperdição01	2.999
11	amordeperdição02	2.156
12	amordeperdição03	2.704
13	amordeperdição04	2.272
14	amordeperdição05	3.214
15	amordeperdição06	3.920
	TOTAL	**53.971**

Fonte: Elaborada pela autora

Em relação ao CCN, o *Corpus* de Contraste contém 891 *tokens* a mais.

Nas seções a seguir, expomos uma breve panorâmica de quem são os autores e do que tratam as obras do *Corpus* de Contraste.

3.2.1 O Cortiço

Aluísio (Tancredo Gonçalves) de Azevedo é natural de São Luís, nascido em 1857 e morreu em Buenos Aires em 1913. Aos 10 anos compôs uma tragédia em versos, e já era também em tenra idade pintor e cenógrafo, daí surgiu sua paixão

pela pintura. Foi um grande desenhista, somente mais tarde se tornou escritor profissional. Sua influência literária deu-se inicialmente pela alta influência de sua mãe, responsável pela educação literária dos filhos, que desde cedo os obrigava a ler em voz alta para ela ouvir.

Foi viver na Corte (Rio de Janeiro) aos 19 anos, e estreou como caricaturista no Fígaro recebendo um prêmio pela charge *Os trinta botões*. Voltou a São Luís, após a morte de seu pai, em 1878 e começou a escrever sob o pseudônimo de Pitibri. Foi fundador do jornal diário mais importante de São Luís à época: Pacotilha. Seu primeiro romance *Uma lágrima de mulher* foi publicado em 1879 em São Luís e passou despercebido pela crítica, pois diziam que era “fraco e excessivamente romântico” (MEIRELES; FERREIRA; VIEIRA FILHO, 1958, p. 87). O seu mais belo livro ele publicou em 1881, *O Mulato* foi “uma bomba de inopino lançada contra a sociedade maranhense de então, conservadora, dada a beguinagens, negreira e atrasada” (MEIRELES, FERREIRA; VIEIRA FILHO, 1958, p. 87), pois retratava os aspectos mais reprováveis da sociedade. Romance realista, que introduziu na literatura a figura do mulato, considerado uma ousadia na época, diante de tantos senhores poderosos. Como conta Moraes sobre esse momento (1976, p. 121-122), Aluísio Azevedo não deveria permanecer mais na sua cidade natal, “onde o ambiente lhe é muito hostil”. Hoje poder-se-ia dizer que *O Mulato* foi um *best-seller*.

O forte da escrita de Aluísio é o romance de costumes, retratando a vida humilde dos subúrbios ou dos bairros pobres do Rio de Janeiro, cidade que passou a ser sua morada. Neste cenário surgiu o romance *O Cortiço,* cujos capítulos fazem parte do *Corpus* de Contraste deste trabalho. Publicado em 1890, mostra o nascimento, a vida e a morte de um cortiço onde imperam o vício, a miséria e a imoralidade. O personagem central é o comerciante João Romão, dono do cortiço, que para obter riqueza explora os seus empregados e até comete furtos: é um homem inescrupuloso e avarento. A sua amante, Bertoleza, escrava fugida, é uma das mais exploradas pelo comerciante. O autor preocupa-se em descrever o desenvolvimento das ações não de indivíduos isolados, mas de grupos humanos inteiros, que são magnificamente conduzidos por Aluísio no romance: é um exemplo clássico da força descritiva do escritor. Apesar de tudo girar em torno de João Romão e de seu cortiço, no romance não existe um protagonista: o próprio cortiço desempenha o papel de primeira personagem (BRANDÃO, 1979).

Foi membro da Academia Brasileira de Letras fundando a cadeira nº 4; da Academia Amazonense de Letras, cadeiras nº 25 e da Academia Maranhense de Letras, cadeira nº 2 (MORAES, 1976; MEIRELES; FERREIRA; VIEIRA FILHO, 1958). O melhor da sua obra é a tríade naturalista: *O Mulato, Casa de Pensão* e *O Cortiço*.

3.2.2 Amor de Perdição

Camilo (Botelho Ferreira) Castelo Branco é lisboeta, nascido em 16 de março de 1825, filho de portugueses (um pequeno fidalgo e uma mulher do povo), mas criado por uma tia paterna em Vila Real: suicidou-se em 1890, cego e com problemas familiares. Seu sobrenome é originário do marido da tia, Antônio de Azevedo Castelo Branco.

Assim como Coelho Netto e Aluísio Azevedo, Camilo vivia exclusivamente de seus textos. Similar aos de Coelho Netto, os textos de Camilo eram quase sempre improvisados, no dia-a-dia, em ritmo acelerado e raramente passava por revisão. É considerado um dos mais profícuos escritores portugueses, sua obra é vasta e variada (262 títulos): romances e novelas, dramas, comédias, poesia, história, crítica, polêmica, biografia, memórias, crônicas, epistolografia. Seus romances são, em grande parte, uma transposição de sua vida tumultuada (MOISÉS, 1997).

Aos 20 anos fez sua estreia como escritor com a poesia *Os pundonores desgravados*. Apesar de levar uma vida mundana, dispôs-se a se tornar padre aos 25 anos, mas fracassou na tentativa ao se envolver amorosamente com uma freira. Após conhecer Ana Plácido, o grande e complicado amor de sua vida, que lhe rendeu até uma prisão, Camilo publicou *Amor de Perdição* e *Amor de Salvação* que consagraram o gênero novela passional portuguesa (MOISÉS, 1997).

Amor de Perdição, cujos capítulos fazem parte do *Corpus* de Contraste deste trabalho, retrata o amor infeliz entre Simão Botelho e Teresa Albuquerque, que foi prometida em casamento pelos pais ao primo Baltazar Coutinho. Com a recusa de Teresa a se casar com o primo, o pai a encerra em um convento. Os pretendentes se desentendem, culminando na morte de Baltazar e na condenação de Simão à forca, que termina por ser sentenciado ao degredo para a Índia. O fim do romance é trágico: Teresa morre no convento, doente; Simão morre na viagem à

Índia e é atirado ao mar; Mariana personagem apaixonada por Simão (e que estava no mesmo navio) joga-se ao mar e morre abraçada ao cadáver dele.

Sobre o romance, Moisés (1997, p. 9) destaca que “antes de se tornar escritor, Camilo começou a ‘escrever’, com a própria vida, um romance dos mais exaltadamente românticos”. *Amor de Perdição* é um bom exemplo da simbiose arte-vida de Camilo. Toda a sua literatura reflete, de um modo ou de outro, o ambiente geográfico e político de suas andanças a partir da adolescência, destacando seus tipos populares, sua aristocracia decadente e sua burguesia ambiciosa.

3.3 Dos instrumentos e tratamento dos dados

Para proceder à compilação, anotação e analise linguístico-computacional do *Corpus* Coelho Netto utilizamos os instrumentos e ferramentas discriminadas a seguir.

Como Alencar (2009, 2011), decidimos optar por um programa que desenvolvesse as mesmas tarefas de análise automática do *WordSmith Tools* mas que fosse *open source* e possível de rodar também em plataforma livre: Python cumpre portanto essa tarefa. Por ser um *software* livre, com código aberto, “pode ser modificado à vontade pelo usuário, que pode adaptá-lo a suas necessidades ou mesmo aperfeiçoá-lo” (ALENCAR, 2009, p. 137).

A versão utilizada de Python foi a 2.7.3 de 2012 compatível com a versão 2.0.1rc1 do NLTK, um conjunto de ferramentas que auxiliam Python na manipulação dos bancos de dados, que contém um pacote de bibliotecas com inúmeras funções para o processamento da linguagem natural, como detalhado no item 2.4.

Para o processo de etiquetação morfossintática, utilizamos neste trabalho o etiquetador morfossintático Aelius, modelo AeliusHunPos, treinado no Tycho Brahe, porque utilizou textos literários para seus testes e demonstrou uma acurácia superior aos outros modelos (ALENCAR, 2010a, 2013). O HunPos é um “*software* de aprendizado de máquina implementado em Ocaml, distribuído sob licença livre, para o qual o NLTK oferece uma interface amigável” (ALENCAR, 2011b, p. 12) apresentando um bom desempenho. A versão do Aelius utilizada neste trabalho é a versão 0.9.7, atualizada em 25 de fevereiro de 2013[35] (ALENCAR, 2013b).

[35] AELIUS, versão 0.9.7 - http://sourceforge.net/projects/aelius/files/Aelius-February-25-2013.zip/download

Identificados os instrumentos, passamos à demonstração dos procedimentos realizados para a compilação do CCN e do *Corpus* de Contraste.

3.4 Do processo de compilação e anotação morfossintática

O desenvolvimento deste trabalho de pesquisa, de compilação e anotação do CCN e do *Corpus* de Contraste, cumpre etapas distintas como já discriminadas no item 2.5. Para isso, contemplamos todas as fases referenciadas por Berber Sardinha (2004) e Aluísio e Almeida (2006) e acrescidas de outras necessárias à compilação de *corpus* de textos literários, cujo processo detalhamos a seguir.

Compilar um *corpus* na LCORP não é o mesmo que coletar e arquivar textos; envolve muitas outras ações. Coletar é só a parte inicial de todo esse processo. É um processo laborioso e minucioso, pois envolve tarefas que por vezes demanda muito tempo do pesquisador, sobretudo por ser realizado em grande parte de forma manual, requerendo muito cuidado.

O processo de compilação dos *Corpora* consistiu em:

3.4.1 Seleção e coleta de textos:

Considerando-se a tipologia (SINCLAIR, 1995; TAGNIN, 2010; BERBER SARDINHA, 2000; MCENERY; WILSON, 2001; SOUZA; GATTI, 2002) definida na compilação do *Corpus* Coelho Netto e do *Corpus* de Contraste, que tiveram o mesmo rigor na criação, resumimos a caracterização dos *Corpora* deste trabalho da seguinte forma: pequeno, histórico, escrito, de amostragem, especializado, de língua nativa, por gênero, texto literário, de estudo, anotado morfossintaticamente, aberto e disponível na *WEB.* A única exceção quanto à tipologia é que o *Corpus* de Contraste não é de amostragem, é de comparação.

Inicialmente, procedemos com a seleção dos textos que constituem o *Corpus* Coelho Netto. Alguns fatores foram considerados para isso: o primeiro se referia à disponibilidade do texto em formato digital e de livre acesso na *WEB*. Para composição do *corpus*, e por se constituir de textos literários dos séc. XIX e início do séc. XX, já são considerados de domínio público. Outro fator relevante foi o fato de as obras escolhidas serem consideradas renomadas e significativas da antologia do

escritor. Este último fator foi o que nos levou a escolher alguns textos que não estavam disponíveis no formato digital, e que, portanto deveriam ser digitados para serem incorporados ao *Corpus* e analisados automaticamente.

Os romances *A Conquista* e *Turbilhão*, e o conto *Firmo, o Vaqueiro* do livro *Sertão* foram capturados da *WEB:* os romances do site *Wikisource*[36] e o conto do site *Peregrinacultural´s WEBlog*[37].

Os outros textos, dois contos, *O Enterro* e *Mandovi*, foram digitados e digitalizados tendo como referência o texto original publicado em 1926 (5ª edição) pela Livraria Chardron, de Lello & Irmão Ltda, Portugal: a 1ª edição foi publicada em 1897.

Não há muitas limitações no arquivamento de textos no *Wikisource*, há somente uma pequena demora para os administradores do *site* processarem os arquivos e disponibilizar para o público (permissão). Publicamos os três contos de *Sertão,* que não estavam disponíveis no *site.* Os textos que foram anexados ao *Wikisource* são textos já comparados entre impresso e digital, o que gerou textos idôneos quanto à sua fidelidade. Os contos estão disponíveis no link http://pt.wikisource.org/wiki/Sert%C3%A3o.

É comum na edição de textos impressos o uso do programa OCR (*Optical Character Recognition*), recurso automático de reconhecimentos de caracteres, que transformam imagens em textos. Para textos de grafias antigas portuguesas, mais precisamente dos séc. XV e XVI, as tecnologias atuais de reconhecimento de caracteres mostram-se inadequadas: o OCR apresentou um índice de acertos entre 56% a 76% (PAIXÃO DE SOUSA; KLEPER; FARIA, 2012), e por isso todos os textos acabam por serem submetidos à revisão manual para correção e ajustes de caracteres. Levando-se em conta essa possibilidade, nada producente, optamos por digitar manualmente os textos diretamente dos originais, evitando utilizar o OCR ou outro programa com essas características e o processo de “estabilização” dos textos demandar mais tempo.

Realizada essa etapa, passamos então ao processo de manipulação dos textos.

---

[36] Biblioteca livre de acervo digital de livros e textos fontes que são de domínio público. Só em português (europeu e brasileiro) contém 95 733 textos - http://pt.*Wikisource*.org/wiki/P%C3%A1gina_principal

[37] Blog Peregrinacultural´s Weblog - http://peregrinacultural.wordpress.com/2009/12/21/firmo-o-vaqueiro-conto-de-natal-de-coelho-neto-texto-integra

3.4.2 Manipulação dos Corpora:

Esse processo se resume em duas etapas:

1ª etapa - Limpeza, geração e nomeação de arquivos:

Realizamos a limpeza dos Corpora retirando dos arquivos nº de páginas, cabeçalhos, rodapés e outras informações que não são propriamente o texto, para permitir o processamento computacional dos *Corpora*

Todos os arquivos de textos, primários, foram convertidos para o formato "*.txt"*. A partir dessa ação as outras etapas puderam ser realizadas, pois o processamento tanto manual quanto automático de todos os textos necessita da nomeação correta dos arquivos. Essa é a fase menos complexa dessa etapa da compilação dos *Corpora.*

Após a etapa de seleção e limpeza, foram gerados e nomeados os arquivos dos textos chamados "puros", para conseguinte seguirem para a outra etapa da manipulação (edição e atualização), conforme quadro a segur.

Quadro 7 – Nomeação inicial de textos selecionados para os *Corpora*

OBRA	TIPO DE TEXTO	NOMEAÇÃO ARQUIVO
***CORPUS* COELHO NETTO**		
A Conquista	Romance	aconquista01.txt
		aconquista06.txt
		aconquista12.txt
		aconquista18.txt
		aconquista23.txt
		aconquista28.txt
Turbilhão	Romance	turbilhao01.txt
		turbilhao05.txt
		turbilhao10.txt
		turbilhao15.txt
		turbilhao20.txt
O enterro	Conto	oenterro.txt
Firmo, o vaqueiro	Conto	firmo.txt
Mandovi	Conto	mandovi.txt
***CORPUS* DE CONTRASTE**		

O Cortiço	Romance	ocortico01.txt
		ocortico02.txt
		ocortico03.txt
		ocortico04.txt
		ocortico05.txt
		ocortico06.txt
		ocortico07.txt
		ocortico08.txt
Amor de Perdição	Romance	amordeperdicaoINT.txt
		amordeperdicao01.txt
		amordeperdicao02.txt
		amordeperdicao03.txt
		amordeperdicao04.txt
		amordeperdicao05.txt
		amordeperdicao06.txt

Fonte: Elaborado pela autora

Este formato de nomeação dos arquivos dos *Corpora*, feito manualmente, serve como base de geração de outros arquivos, tanto manual quanto automaticamente. Das etapas de geração de arquivos, somente os de formato *".pdt.txt", ".pdt.hunpos.txt"* e *".gold.txt"* foram gerados automaticamente, os outros arquivos todos foram gerados manualmente, ver formatos nos Quadros 8 e 9.

Como os *Corpora* compilados deste trabalho contêm muitos arquivos, eles foram nomeados conforme os modelos no quadro 8, e seguindo orientações do processo de compilação de um *corpus* sugerido por Aluísio e Almeida (2006).

Quadro 8 – Modelo de nomeação dos arquivos para compilação do *Corpus* Coelho Netto.

OBRA	TIPO DE TEXTO LITERÁRIO	ARQUIVO	REFERÊNCIA DE ARQUIVO
		***CORPUS* COELHO NETTO**	
A Conquista	Romance	aconquista01.txt	Texto original capturado da *WEB* ou texto digitado e depois disponibilizado na *WEB*
		aconquista01.edt.txt	Texto limpo, editado e atualizado.
		aconquista01.pdt.txt	Texto pós-editado automaticamente (por *script*)
		aconquista01.pdt.hunpos.txt	Texto anotado automaticamente
		aconquista01.pdt.hunpos.corr_F.txt	Texto corrigido manualmente (1º corretor)
		aconquista01.pdt.hunpos.corr_H.txt	Texto corrigido manualmente (2º corretor)
		aconquista01.pdt.hunpos.corr_FH.txt	Texto final corrigido manualmente (dois corretores)
		aconquista01.gold.txt	Texto padrão *gold*, gerado após correção automática.
***CORPUS* DE CONTRASTE**			
O Cortiço	Romance	ocortico06.txt	Texto original capturado da *WEB* ou texto digitado e depois disponibilizado na *WEB*
		ocortico06.edt.txt	Texto limpo, editado e atualizado.
		ocortico06.pdt.txt	Texto pós-editado automaticamente (por *script*)
		ocortico06.pdt.hunpos.txt	Texto anotado automaticamente
		ocortico06.pdt.hunpos.corr_F.txt	Texto corrigido manualmente (1º corretor)
		ocortico06.pdt.hunpos.corr_H.txt	Texto corrigido manualmente (2º corretor)
		ocortico06.pdt.hunpos.corr_FH.txt	Texto final corrigido manualmente (dois corretores)
		ocortico06.gold.txt	Texto padrão *gold*, gerado após correção automática.

Fonte: Elaborado pela autora

Tanto o C*orpus* CN quanto o de Contraste passam pelo mesmo processo de arquivamento, cumprindo todas as etapas necessárias para obter o texto padrão *gold*.

O quadro 9 descreve como estes arquivos são nomeados:

Quadro 9 – Descrição de nomeação dos arquivos

NOME	DESCRIÇÃO
aconquista	Nome do romance/conto
01	Nº do capítulo do romance/conto
txt	Formato do tipo de arquivo
edt	Texto editado

pdt	Texto pós-editado automaticamente
hunpos	*Software* de aprendizagem de máquina
corr_f	Identificação do 1º corretor manual
corr_h	Identificação do 2º corretor manual
corr_fh	Identificação da correção final unificada dos dois corretores
gold	Texto final pós-correção manual e automática, respectivamente.

Fonte: Elaborado pela autora

Cada texto dos *Corpora* recebeu uma nomeação específica, conforme modelo acima, podendo ser texto único (contos) ou separado por capítulos (romances).

2ª etapa - Edição e Atualização:

Outra ação que consideramos importante e que não é vislumbrada na metodologia de Aluísio e Almeida (2006), mas alinhada às ações realizadas para a construção do *Corpus* Histórico do Português Tycho Brahe, é o processo de edição dos textos que "foi desenvolvido de modo a trazer para o âmbito do trabalho de edição filológica alguns avanços no processamento eletrônico de textos"[38], e tem a finalidade de conferir à edição a responsabilidade pela máxima fidelidade em relação aos textos originais. Também, posteriormente, os resultados dessa edição servirão para a preparação dos textos para outras análises automáticas. Retomamos a fala de Paixão de Sousa, Kleper e Faria (2012) quanto ao uso de textos antigos para composição de *corpus*, quanto à garantia de fidelidade às formas originais dos textos, no sentido de produzir *corpora* idôneos: isso reforça a importância da edição dos textos.

No caso do *Corpus* Coelho Netto e do *Corpus* de Contraste, a edição dos arquivos está relacionada, primeiramente, com a comparação do texto digitado e o texto original impresso. Com uns *Corpora* os mais fieis possíveis aos textos originais, procedemos com a Edição e Atualização destes, como detalhamos aqui.

Esta etapa consistiu em um momento importante na compilação dos *Corpora*. Além de minuciosa, demandou um tempo além do pretendido,

[38] Texto disponível no Sistema de Edições Eletrônicas do Corpus Tycho Brahe: fundamentos, diretrizes e procedimentos, UNICAMP, 2007 - http://www.tycho.iel.unicamp.br/~tycho/corpus/manual/prep/manual_frameset.html

considerando-se as disparidades encontradas na comparação entre os textos, que muito nos surpreendeu.

Os textos comparados do *Corpus* Coelho Netto foram: *A Conquista, Turbilhão* e *Firmo, o vaqueiro*, porque foram capturados da *WEB*, portanto textos digitais. Da mesma forma com os textos do *Corpus* de Contraste: *O Cortiço* e *Amor de Perdição.* Dos 15 (quinze) arquivos (capítulos de romances ou contos) do *Corpus* Coelho Netto, somente dois não foram coletados da *WEB*. Isso significa dizer que 28 (vinte e oito) arquivos, somando-se os do *Corpus* de Contraste, foram submetidos à comparação que foi feita toda manualmente, por mais de uma vez, para evitar que as disparidades entre eles passassem despercebidas.

Via de regra, os textos antigos digitais, disponíveis na *WEB*, passam pelo processo de reconhecimento de caracteres e conversão em formatos editáveis via OCR ou outro *software* similar, e alguns deles não correspondem aos textos impressos de origem, pois são escaneados e depois transformados. Nesse ínterim, muitos caracteres são mal interpretados pelo *software*, gerando diversas incorreções no texto final, mesmo passando por revisão.

Os textos “limpos” foram impressos por nós e, com o apoio dos livros dos autores, foi feita a comparação dos textos. Para o texto *A Conquista* utilizamos o livro impresso do autor, publicado em 1921, em sua terceira edição.

Detectamos no momento da comparação dos arquivos do *Corpus* Coelho Netto e do *Corpus* de Contraste que alguns textos estão digitados com palavras e ou expressões diferentes do texto impresso, original ou de edições posteriores. Editamos o texto digital para que esse ficasse fiel ao que foi publicado na forma impressa. Eis algumas das situações encontradas:

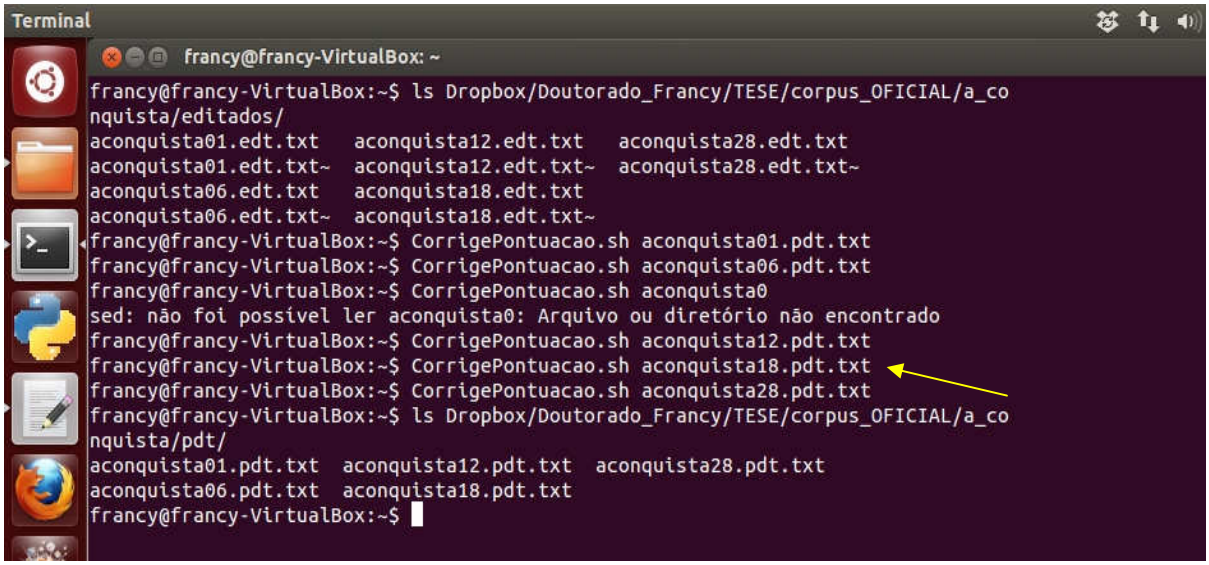
- presença de palavras, frases e até parágrafos que não existem no texto impresso e continham no texto digital;
- palavras diferentes no texto digital em comparação com o impresso, respectivamente: fúria por desespero; despertar por acordar; balordamente por bestamente; descambar por decair; intangida por entanguida; ressumbrava-lhe por molhava-se; encaminhavam por chegavam; venho por chego; de mais por demais; manchil por machadadas;
- diferença na ordem das frases;
- supressão de plural no texto digital;

- supressão e ou acréscimo de palavras no texto digital;
- diferença nas pontuações, especialmente pontos, dois pontos, vírgulas. O texto digital continha muitas diferenças em relação ao impresso;
- o texto digital não continha uma grafia correta em todas as reticências, por vezes aparecendo dois pontos seguidos e até quatro.

Pudemos observar que muitos dos “erros” encontrados no texto digital são decorrentes das dificuldades do reconhecedor de caracteres quando do momento de transformação de texto escaneado para o digital, especialmente pontuações, maiúsculas, e palavras com poucas diferenças entre si.

Todas essas correções foram realizadas em um arquivo de texto criado para esse fim, com terminação *“.edt.txt*”, feitas manualmente, palavra por palavra. A exceção são as correções das reticências, dos travessões em forma de hífen, aspas diferentes e outros, que foram uniformizadas e adequadas por meio do *script* “*CorrigePontuacao.sh*” que é executado no terminal do Linux, criando um texto adequado à leitura do etiquetador morfossintático, como mostrado na figura 13. O *Script* está detalhado no Apêndice L.

Figura 13 - Execução do *script* “*CorrigePontuacao.sh*”



Fonte: Elaborada pela autora

Observando a figura, é possível perceber que essa ação gera automaticamente, a partir do texto editado, um arquivo com terminação *“pdt.txt”*

(seta amarela). Esse arquivo final gerado é o propício para iniciar o processo de anotação morfossintática, etapa seguinte de compilação do *Corpus*.

Para o texto *Turbilhão,* como comparação, utilizamos o livro impresso do autor, publicado em 1958. Considerando-se que a primeira edição foi publicada em 1906, já no século XX, é uma republicação organizada com a vigência das normas ortográficas do Acordo Ortográfico de 1945. Por isso, quando do momento de comparação com o texto digital, poucas diferenças ortográficas foram encontradas. As diferenças encontradas na comparação são:

- presença de palavras, frases e até parágrafos que não existem no texto impresso e continham no texto digital;
- palavras diferentes no texto digital em comparação com o impresso, respectivamente: exercita por exercia; antigo por artigo; válcula por válvula; respungou por resmungou; folhas por calhas; emblocados por embiocados; estas por essas; monte por morte; escadada por encadaria; apuado por amuado; e por de; ora por hora; indolante por indolente; em por com;
- diferença na ordem das frases;
- supressão de plural no texto digital;
- supressão e ou acréscimo de palavras no texto digital;
- diferença nas pontuações, especialmente pontos, dois pontos, vírgulas. O texto digital continha muitas diferenças em relação ao impresso;
- o texto digital não continha uma grafia correta em todas as reticências, por vezes aparecendo dois pontos seguidos e até quatro.

Pudemos observar também que a maioria dos “erros” encontrados no texto digital, sobretudo na diferença entre as palavras do texto digital x texto impresso, são as mesmas já relatadas na comparação do texto *A Conquista*. Estas correções também foram realizadas em um arquivo de texto criado para esse fim, com terminação *“edt.txt”*. Todas feitas igualmente como feito com *A Conquista*.

O arquivo do conto *Firmo, o vaqueiro*, capturado do *site Peregrinacultural´s WEBlog* (à época da busca, foi o único site que encontramos na *WEB* que disponibilizava o texto) foi comparado com o texto impresso de 1926 publicado no Porto – Portugal, pela Livraria Chardron. No *site* não há informação de

que edição e ano o livro foi digitalizado. Atualmente, o site *Wikisource* já contém o conto, por inserção nossa, fazendo parte desse trabalho.

Quando da comparação impresso x digital, as diferenças encontradas foram:

- palavras diferentes no texto digital em comparação com o impresso, respectivamente: vara por Iara; aprecia por aparecia; cou por lou; às vezes por as velhas; mesa por sala; em por no;
- supressão e ou acréscimo de palavras no texto digital;
- diferença nas pontuações, especialmente pontos, dois pontos, vírgulas. O texto digital continha muitas diferenças em relação ao impresso;
- o texto digital não continha uma grafia correta em todas as reticências, por vezes aparecendo dois pontos seguidos e até quatro.

Também as correções quanto às pontuações foram feitas por meio de *script* já mencionado na correção dos outros textos.

Os textos dos contos *O enterro* e *Mandovi* não foram capturados de nenhum *site*, pois após busca na *WEB* não encontramos os textos disponíveis digitalmente. Para isso, foi necessária a digitação e depois a digitalização destes. Para a digitação, utilizamos o livro impresso em 1926, em sua quinta edição, já que foi publicado a primeira vez em 1896. Em função disso, o texto foi digitado na forma original, somente sofrendo atualização ortográfica compatível com o processamento computacional de textos antigos e observando a norma vigente, etapa seguinte do processo de compilação. Portanto, não houve a necessidade de se fazer comparações. Estes contos também já disponibilizamos no site *Wikisource.*

O *Corpus* de Contraste passou pelo mesmo processo do *Corpus* Coelho Netto. Os arquivos da obra *O Cortiço* foram capturados também da *Wikisource*, que não informa qual a data de publicação do livro utilizado na digitalização. O livro impresso para comparação com o digital é uma publicação de 1985: não encontramos a edição deste livro, pois a primeira publicação da obra é de 1890.

Observamos que na comparação poucas divergências foram encontradas, as mais comuns e poucas, foram: troca de singular por plural; palavras trocadas; troca de minúsculas por maiúsculas; e supressão de letras. A maioria, entendemos, em função das dificuldades que o *software* de reconhecimento de

caracteres encontra ou então quem supostamente tenha digitado o texto. Não foram encontradas diferenças ortográficas entre os textos.

Os capítulos de *Amor de Perdição* também foram retirados do *Wikisource*, e não informa também qual a data de publicação do livro utilizado na digitalização. O livro impresso para comparação com o digital é uma publicação de 1997, e informa que o texto está atualizado conforme o Acordo de 1971; a edição é a 24ª. Observamos que na comparação poucas divergências foram encontradas, praticamente as mesmas encontradas nos arquivos de *O Cortiço*.

Outra ação dessa etapa de compilação está relacionada à uniformização dos textos de acordo com a norma ortográfica vigente.

A atualização requerida é em relação à ortografia vigente à época da publicação dos livros, no período entre 1862 a 1906, a maioria pertencente ao séc. XIX. A ortografia deste século faz parte do período chamado de pseudoepistemológico "que buscava a grafia correta na origem das palavras, e das paroxítonas e proparoxítonas sem diacríticos" (CAMPOS; ANDRADE, 2011, p. 1502). A simplificação ortográfica só ocorreu em 1907, antes disso não havia uma regra ortográfica formalmente estabelecida. As principais características da ortografia no século XIX, observadas nos manuscritos e jornais, são as seguintes: consoantes duplicadas; consoantes mudas ou nulas; paroxítonas sem acento, terminadas em ditongo ou em l; uso do h para indicar hiato; ditongo com semivogal i, y, e, o e u (CAMPOS; ANDRADE, 2011).

Ressaltamos que todos os textos foram atualizados conforme o Novo Acordo Ortográfico da Língua Portuguesa, assinado em 2009, mas só com previsão de implantação, de fato, após o período de transição para as novas regras pelos países da Comunidade de países de Língua Portuguesa (CLP) que finaliza em 31 de dezembro de 2015. No Exemplo 5, demonstramos algumas atualizações realizadas nos textos.

Exemplo 5:

- texto original impresso de 1921: *"Era um sombrio* ***predio*** *entre velhíssimas arvores copadas, cujos ramos altos faziam uma* ***abobada impenetravel*** *ao sol. As paredes, pintadas dum vermelho* ***amarellado****, pareciam cobertas de limo. Os canteiros esquecidos estavam invadidos pelo* ***matto****, as aléas eram* ***humidas*** *e tinham placas lutulentas, dum* ***avelludado*** *fino".*
- texto digital de 1913: "*Era um sombrio prédio entre velhíssimas árvores copadas, cujos ramos altos faziam uma abóbada impenetrável ao sol. As paredes, pintadas de um verde amarelado, pareciam*

cobertas de limo. Os canteiros esquecidos estavam invadidos pelo mato, as ***aléias*** *eram úmidas e tinham placas lutulentas, de um aveludado fino".*

Observamos que, apesar de o texto digital ser mais antigo que o texto impresso, a ortografia daquele estava bastante atualizada, mas contendo algumas palavras que mantinham as características da ortografia do séc. XIX. O destaque em negrito é para demonstrar que a grafia do texto impresso correspondia ao que disseram Campos e Andrade (2011), quanto à ortografia vigente. Considerando-se que a simplificação ortográfica ocorreu em 1907, e o texto digital (do exemplo) foi publicado em 1913, cremos que o texto já cumpria com a simplificação ortográfica. Entendemos que o texto de 1921 ainda mantinha a ortografia do séc. XIX porque foi editado e publicado em Lisboa, pela Livraria Chardron, em Portugal, que mantinha os padrões tradicionais da língua. Atualizando esse exemplo conforme o Acordo Ortográfico atual, a mudança seria somente na palavra "aléias", que perderia o acento agudo.

As comparações e atualizações são etapas muito importantes no processo de compilação do *corpus*. Muitas diferenças encontradas, fazendo a comparação do impresso com o digital, influenciam sobremaneira no processo de anotação morfossintática dos *Corpora,* pois se se anotasse o texto sem cumprir estas etapas os resultados seriam bastante diferentes, o que implicaria em resultados não concretos, só para citar como exemplo: diferença na quantidade de *tokens*, na diferenciação entre classes de palavras anotadas, nas *tags* utilizadas na anotação, nas ocorrências e em outras informações que resultassem dessas diferenças.

Quanto às análises linguístico-computacionais as diferenças poderiam contribuir para resultados errôneos sobre os textos analisados, destacando também: diferença na quantidade de *tokens*, nas *hapaxes*, na busca pela diversidade lexical e nas ocorrências das classes de palavras.

No Exemplo 6 relacionamos alguns "erros" encontrados na comparação que certamente influenciariam nos resultados obtidos como já citados. Para melhor entendimento, o texto correto, adotado para as anotações e análises é o do texto impresso. As palavras em destaque são as que sofreram modificações.

Exemplo 6:

- quantidade de tokens:

* texto digital: *"iluminada **por seis bicos de gás, respirando** por duas janelas largas", "caminhou vagarosamente **a** porta", "emoção **malcontida**", "algum plano de violência", "não viam **seus olhos** senão a casa", "e telas de artistas célebres sóbrias".*
* texto impresso: "*iluminada por duas janelas largas", "caminhou vagarosamente **para** a porta", "emoção mal contida", "algum **novo** plano de violência", "não viam senão a casa", "e telas de artistas célebres **em molduras** sóbrias".*

- diferenciação entre classes de palavras anotadas

* texto digital: "***ora**", "do profeta **a** reboavam", "emoção **malcontida**", "**tem** notável por sua hierarquia", "sob a **dava** do Hércules", "mais um discurso e **demais** abraços", "o chapéu **derreado à** nuca".*
* texto impresso: "*hora", "do profeta e reboavam", "emoção mal contida", "tão notável por sua hierarquia", "sob a clava do Hércules", "mais um discurso e de mais abraços", "o chapéu sobre a nuca".*

A grande maioria destas diferenças nos textos é de supressão ou acréscimo de palavras presentes no texto digital, o que de alguma forma influencia nas análises morfológicas no momento da anotação. Destacamos, no exemplo 6, trechos pequenos, ou palavras isoladas, mas detectamos no livro *A Conquista* trechos inteiros, parágrafos até, que foram suprimidos no texto digital, como por exemplo:

Exemplo 7:

* texto digital: *"[...] e em climas ásperos, tiritando, tarantulamente, à neve. Ele ao menos, tinha a benignidade do clima paradisíaco, sem invernias que o encarangassem ou congelassem, como acontecera aos soldados de Napoleão na Rússia, e tinha a esperança de vencer grandes prélios literários, impondo-se à Pátria e ao mundo com os períodos da sua pena".*
* texto impresso: "*Ele ao menos, tinha um leito macio e a clemência do céu, vivia num país que a neve não flagela, dormia repousadamente e não tiritava à nevasca. Tinha a mercê do clima amável e a eterna primavera que tinge o céu de azul e reverdece os campos, e tinha a esperança, consolação suprema".*

Além de distorcer o conteúdo do romance, o texto digital mostra um texto que em quase nada se compara com o impresso. E nesse caso, as distorções são gravíssimas, sobretudo para o trato computacional, por tudo que já explicamos sobre as incoerências quanto aos erros encontrados nos textos digitais.

Esse processo nos alertou para muitas questões importantes que devem servir de objeto de muito cuidado quando da busca e do manuseio de textos da *WEB.* Dentre algumas, que é preciso sempre verificar a qualidade/idoneidade, em relação aos originais, dos textos disponibilizados. Nossa pesquisa mostrou que muitos deles não são fieis aos originais. Simplesmente coletar um texto da *WEB* e criar um arquivo não consiste necessariamente em compilação de um *corpus*. O detalhamento das etapas concluídas demonstra a laboriosa tarefa de se compilar um *corpus* e se obter dele dados concretos que, significativamente, venham a contribuir com as investigações linguísticas.

Após a conclusão dessas ações, os arquivos finais foram gerados em formato *"pdt"*, e então procedemos com o processo de anotação morfossintática, etapa seguinte de nosso trabalho.

3.5 Anotação Morfossintática

A anotação realizada no *Corpus* de Estudo e no *Corpus* de Contraste é uma anotação morfossintática, conforme explorado no item 2.6. A anotação é realizada de forma automática, utilizando o AeliusHunPos, adequado a anotações de textos literários, porque teve como *corpus* de treinamento o *Corpus* do Português Histórico Tycho Brahe.

As etiquetas utilizadas para a anotação destes *Corpora* têm como referência as de anotação do *Corpus* Tycho Brahe. Alencar (2010a) explica que o conjunto de etiquetas (*tags*) do Tycho é complexo, causando muitas ambiguidades durante as anotações. Ao implementar o AeliusHunPos, Alencar (2013a) procurou diversificar os conjuntos de etiquetas utilizadas por este Etiquetador, solucionando algumas ambiguidades e buscando sempre melhorar seu desempenho. De acordo com o Manual do Sistema de Anotação Morfológica, o Tycho contém atualmente 377 etiquetas[39], com última atualização em maio de 2008. O Apêndice H apresenta uma relação simplificada destas etiquetas.

Após a anotação morfossintática de todos os *Corpora* deste trabalho, com o processamento computacional por meio de Python e utilizando a ferramenta AeliusHunPos e utilizando a biblioteca NLTK, foram gerados novos arquivos dos

[39] Relação de *tags* do Manual do Sistema de Anotação Morfológica disponível em: http://www.tycho.iel.unicamp.br/~tycho/corpus/manual/alltags.txt

textos anotados. A criação desses arquivos é toda feita automaticamente, diferente do processo inicial de seleção e compilação dos textos, que é feito manualmente. O processador Python gera os arquivos conforme ação demandada, conforme Quadro 8, de modelo de nomeação de arquivos

Após a criação dos arquivos com os textos anotados (“*.pdt.hunpos.txt”*), fizemos a seleção dos arquivos que deveriam ser corrigidos manualmente. Selecionados três arquivos do *Corpus* Coelho Netto, totalizando em torno de 10% do *corpus* pós-editado, com 5.405 *tokens*. Essa é uma amostra representativa de correção para apurar a acurácia do Etiquetador. Tomamos como base o percentual de correção manual realizado pelo ComPlin com o CORPTEXLIT (ALENCAR, 2010b).

Selecionamos do *Corpus* de Contraste a quantidade de *tokens* o mais aproximado possível do número de *tokens* dos textos anotados do *Corpus* Coelho Netto: dois capítulos de *O Cortiço*, contendo 5.404 *tokens* pós-edição dos arquivos. No total, corrigimos 10.809 *tokens.* Vale lembrar que *tokens* compreendem palavras e pontuações.

Feita a correção manual, realizada no próprio arquivo anotado, após execução de comandos específicos em Python, geramos um novo arquivo com o texto corrigido. Este arquivo deverá gerar um outro, que servirá de padrão para a verificação e comparação com o *Corpus* de Contraste, denominado de *gold,* para a avaliação da acurácia do Etiquetador. A correção manual das anotações foi feita por dois corretores (humanos), criando um arquivo que convergissem as duas correções, e daí então submeter o arquivo à correção automática das etiquetas, gerando o arquivo padrão *gold.*

Reafirmamos que a quantidade de *tokens* para a correção manual de um *corpus* de comparação deve ser igual, senão similar ao quantitativo de *tokens* do *corpus* de Estudo. Porquanto, após a seleção dos arquivos adequados à quantidade necessária, relacionamos estes na Tabela 3:

Tabela 3 - Relação de arquivos selecionados para correção manual dos *Corpora*

OBRA	TIPO DE TEXTO	NOMEAÇÃO ARQUIVO	*TOKENS*
		***CORPUS* COELHO NETTO**	
A Conquista	Romance	aconquista06.pdt.hunpos.txt	2.708
Turbilhão	Romance	turbilhao20. pdt.hunpos.txt	1.169
O enterro	Conto	oenterro.pdt.hunpos.txt	1.528
		TOTAL	**5.405**
		***CORPUS* DE CONTRASTE**	
O Cortiço	Romance	ocortico05.pdt.hunpos.txt	2.251
		ocortico06.pdt.hunpos.txt	3.153
		TOTAL	**5.404**

Fonte: Elaborada pela autora.

Observa-se nesta tabela, que o quantitativo de *tokens* dos arquivos do *Corpus* de Contraste é praticamente igual ao do *Corpus* Coelho Netto. Para conseguirmos isso, a escolha dos textos do *Corpus* de Contraste não foi aleatória, foi em função de tentativas em equiparar os totais, condicionado por quantitativo de *tokens*.

Após a correção de todos os arquivos selecionados, pelo 1º corretor, geramos manualmente os arquivos com o modelo "*.pdt.hunpos.corr_F.txt".* Os mesmos arquivos foram enviados para o segundo corretor, que após conclusão de sua correção, geramos arquivos com o formato "*.pdt.hunpos.corr_H.txt"*[40]*.*

A etapa semifinal consistiu em contrastar as duas correções gerando arquivos uniformes, com a nomeação de "*aconquista06.pdt.hunpos.corr_FH.txt*": o arquivo de consenso. Não houve necessidade de um terceiro anotador para dirimir as possíveis discordâncias dos outros dois anotadores.

Para uma melhor percepção da correção manual, o quadro abaixo mostra os *tokens* corrigidos manualmente seguindo o protocolo observado no Exemplo 8:

[40] O primeiro corretor foi a autora da tese, e o segundo corretor foi o aluno de graduação de Letras-Francês e Bolsista de Iniciação Científica CNPq do projeto "Técnicas em softwares livres para a linguística de corpus", da UFC, Hélio Leonam Barroso, portanto com a inicial do nome H, correspondente à nomeação do arquivo.

Quadro 10 - Exemplo de *tokens* com anotação automática e correção manual

Os/D-P canteiros/N-P esquecidos/VB-AN-P estavam/ET-D invadidos/VB-AN-P pelo/P+D mato/N ,/, as/D-F-P aleias/N-P eram/SR-D ***úmidas/VB-AN-F-P@ADJ-F-P*** e/CONJ tinham/TR-D ***placas/ADJ-F-P@N-P lutulentas/N-P@ADJ-F-P*** ,/, de/P um/D-UM aveludado/N fino/ADJ ./.

Fonte: Elaborado pela autora

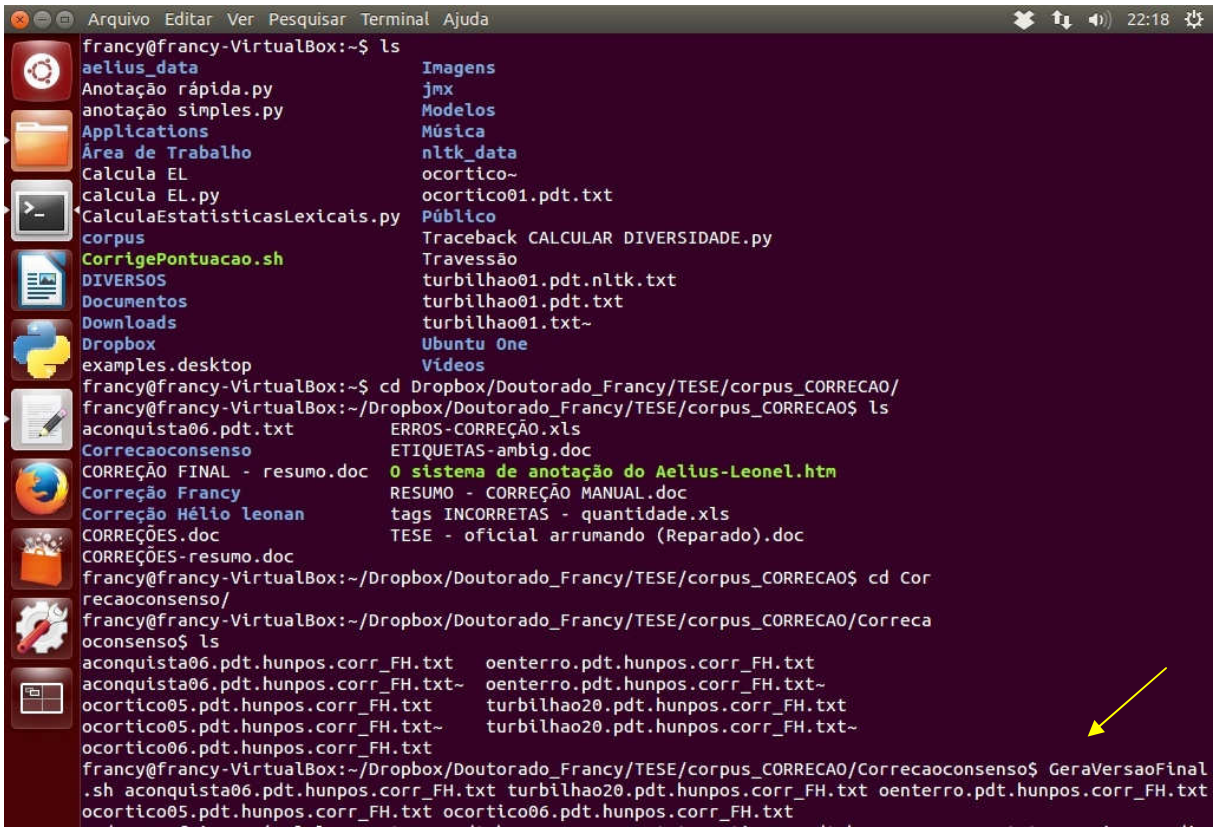
A fase final consistiu em corrigir automaticamente os arquivos uniformizados. Estas correções consistem em retirar dos textos corrigidos o símbolo "@" que vem após cada *token*. Essa ação é feita submetendo o arquivo de consenso ao processamento no terminal com o script *"GeraVersaoFinal.sh",* que automaticamente gera arquivos com extensão ".gold.txt.". Vejamos o exemplo abaixo:

Exemplo 8:

- *Toledo/**VB-AN@NPR*** - Toledo é o *token*; VB-AN é a etiqueta atribuída; @ é a arroba, símbolo que intermedeia a *tag* errada da correta; e NPR é a etiqueta corrigida manualmente.
- *Toledo/**NPR*** - versão final do *token* anotado, após processamento do *script GeraVersaoFinal.sh*

Para configurar melhor, a figura 14 mostra no terminal como o processo é realizado.

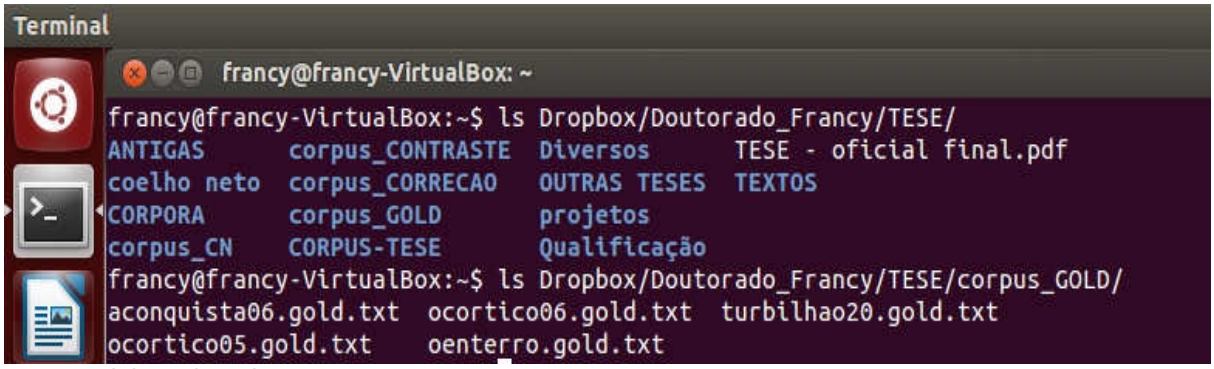
Figura 14 - Execução do *script* que Gera Versão Final dos textos corrigidos manualmente – arquivos *Gold.*



Fonte: Elaborada pela autora

Por meio de *script* no terminal, automaticamente o programa gerou os arquivos padrão *gold ("aconquista06.gold.txt"),* como podem ser vistos na relação de arquivos na Figura 15.

Figura 15 – Relação dos arquivos com extensão "*.gold.txt"* gerados pelo scipt "*GeraVersaoFinal.sh*".



Fonte: Elaborada pela autora

Como isso, os arquivos estavam preparados para passar pelo processo de avaliação de desempenho do etiquetador Aelius, modelo AeliusHunPos, tanto

quanto os outros arquivos gerados em todo esse processo estavam preparados para as análises e discussão do proposto nesta Tese.

O capítulo 4 é destinado à apresentação dos resultados e discussão dos dados encontrados com os procedimentos metodológicos adequados para a LCLC. Exporemos todas as etapas do processo de criação do *Corpus* Coelho Netto, relatando as ações e as dificuldades e ou limitações encontradas nesse percurso, bem como os resultados pretendidos no objetivo da pesquisa, envolvendo desde a compilação até as análises.

4 RESULTADOS E DISCUSSÃO

Considerando-se todas as etapas envolvidas no processo de compilação, anotação e análises linguístico-computacionais de um *corpus*, entendemos que os resultados detalhados em cada etapa seria a forma mais clara de se mostrar os dados, como veremos a seguir.

4.1 Anotação e Resultados

Após o cumprimento do processo de anotação morfossintática detalhado no item 3.5 e da correção manual, destacamos os resultados da correção manual e dos erros cometidos pelo etiquetador:

Tabela 4 – Cinco primeiras *tags* que apresentaram mais erros.

TAG	OCORRÊNCIA
N	**29**
NPR	**23**
ADJ-F	**20**
ADJ	**18**
N-P	**18**

Fonte: Elaborada pela autora com base em dados automatizados e manuais

De forma geral, as *tags* que apresentaram mais erros, por ordem de frequência, são: N, *tag* atribuída a nomes; NPR, atribuída a nomes próprios; ADJ-F, atribuída a adjetivos femininos; ADJ, atribuída a adjetivos sem gênero; e N-P, atribuída a nomes no plural.

Quanto às classes de palavras especificadas abaixo (objeto de análises nos *Corpora)* que apresentaram erros no momento da anotação, por ordem de ocorrência foram:

1. ADJ (adjetivos), apresentando 62 erros, 8 *tags* diferentes (ADJ-F, ADJ, ADJ-F-P, ADJ-G, ADJ-P, ADJ-G-P, ADJ-R-F-P, ADJ-R-G) distribuídas por ordem de ocorrência;

2. VB (verbos), apresentando 59 erros, com 12 *tags* diferentes (VB-P, VB-AN, VB-SP, VB, VB-AN-F, VB-AN-F-P, VB-D, VB-G, VB-I, VB-PP, VB-R, VB-SD) por ordem de ocorrência;
3. ADV (advérbios), com 1 erro, e 1 tag (ADV).

Esses dados foram obtidos organizando-se as tabelas de erros cometidos pela anotação automática e as correções realizadas manualmente. Após a correção manual dos textos, estruturamos estas tabelas e fizemos as comparações, que resultaram nos dados expostos acima.

Para cada *tag* erroneamente etiquetada, os corretores manuais atribuíram a *tag* correta. Na Tabela 5, listamos as *tags* mais demandadas para correção da etiquetação incorreta.

Tabela 5 - Cinco *tags* mais demandadas para substituírem corretamente as *tags* atribuídas incorretamente aos *tokens*.

TAG	OCORRÊNCIA
N	**71**
NPR	**25**
ADJ	**16**
N-P	**11**
ADJ-F-P	**9**

Fonte: Elaborada pela autora com base em dados automatizados e manuais

Observando esta tabela e fazendo um comparativo com a tabela 4, destacamos que a ocorrência de erros e mudanças são em relação às mesmas *tags*, tendo uma variação somente nas etiquetas flexionais. As *tags* N e NPR pontuam como as de maior dificuldade quanto à ambiguidade, incorrendo, portanto, em muitos erros.

Alencar (2010a, p. 4) já se preocupava com esse problema, quando verificou a grande quantidade de erros relacionadas à *tag* NPR, e para isso desenvolveu e implementou em Python um “algoritmo para distinguir nomes próprios de outras palavras de inicial maiúscula”. Apesar dessa implementação, o etiquetador ainda tem essa dificuldade, pois cometeu 23 erros de atribuição da etiqueta NPR. Mesmo com esses erros, temos um percentual de ambiguidade pequeno (3%), considerando-se que já foi considerado grande. Já com a *tag* N, e suas flexões (N – P), o etiquetador cometeu 47 erros de etiquetagem.

Por outro lado, a etiquetagem cometeu muitos erros com atribuição de etiquetas que demandaram a *tag* NPR para corrigir esses erros. Com levantamento feito com os dados da correção manual, esta etiqueta substituiu outras 25 erradas.

O etiquetador também incorreu em um erro de anotação quanto ao reconhecimento de algumas reticências. Dos cinco textos anotados e corrigidos manualmente, o erro só ocorreu em um texto, e em duas situações. Os erros aconteceram nos trechos do arquivo *"aconquista06.pdt.hunpos.txt"* – texto pós-editado e anotado morfossintaticamente. Nas anotações, a *tag* vem sempre posicionada após a barra.

"Ah! tempos..." *- texto sem anotação*
"Ah/INTJ !/. Tempos./. ../.". *– texto anotado incorretamente*

Pode-se observar que a anotação para o *token* "Tempos" foi considerada com uma pontuação, utilizando a etiqueta *"./."*: o correto seria atribuir a etiqueta "N-P". Ocorreu que o toquenizador falhou na segmentação da palavra e das reticências. Testamos diversas vezes o etiquetador, refazendo o processo, modificando as posições da palavra e das reticências para que o toquenizador lesse corretamente os dados. Chegamos à conclusão que ao se afastar as reticências da palavra, a etiquetagem ocorria normalmente, que foi o que aconteceu quando assim procedemos, ficando a anotação da seguinte forma:

"Ah/INTJ !/. Tempos/N-P .../." *– texto anotado corretamente*

Ainda no mesmo arquivo, o etiquetador fez a mesma leitura com a sentença:

"– Besta! Homero...?" *- texto sem anotação*
"–/(Besta/N !/. Homero./.../. ?/." *– texto anotado incorretamente*

O correto, após o mesmo procedimento realizado com a sentença anterior, ficou assim:

"–/(Besta/N !/. Homero/N .../. ?/." - *texto anotado corretamente*

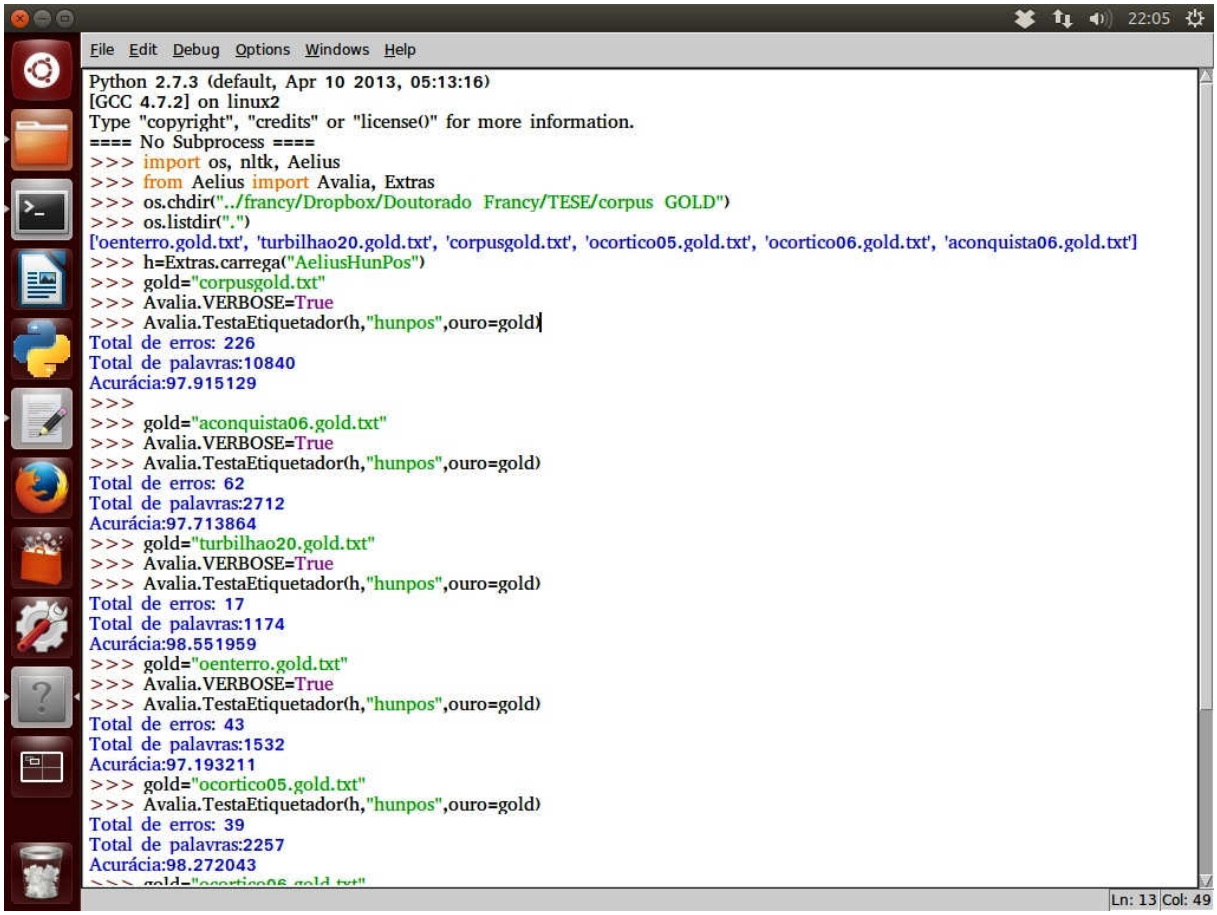
Certamente que o AeliusHunPos deve passar por processo de revisão nesse sentido (processo comum), sendo alimentado com *scripts* específicos que devem sanar essa dificuldade de toquenização. Situações como essas, percebidas na correção manual é que contribuem para a melhoria do desempenho do etiquetador aumentado sua qualidade. No universo de 10.809 *tokens* anotados e corrigidos, somente surgiram essas duas situações.

Os dados compilados na correção manual revelaram que os textos do *Corpus* Coelho Netto, escolhidos para a correção, foram os que mais precisaram de correção manual; as *tags* atribuídas aos *tokens* incorreram em 123 erros. Essa substituição ocorreu mais em relação às *tags* ADJ e N, que foram atribuídas incorretamente aos *tokens*.

Após a avaliação de todo o processo de correção manual, buscando os dados relatados acima, procedemos com a etapa subsequente a essa correção: a criação do arquivo padrão *gold*, como já demonstrado no item 3.4.3.

Gerados os arquivos, automaticamente avaliamos a acurácia do modelo AeliusHunPos, com os textos corrigidos manualmente, como mostra a execução dos comandos do NLTK na figura 16.

Figura 16 - Modelo de processamento de avaliação do etiquetador AeliusHunPos, anotando os 10% do CCN e do *Corpus* de Contraste – arquivos padrão *gold*.



Fonte: Elaborada pela autora

Os resultados compilados do processamento estão distribuídos na Tabela 6:

Tabela 6 - Demonstrativo da acurácia do etiquetador Aelius, modelo AeliusHunPos anotando os 10% do CCN e do *Corpus* de Contraste – arquivo padrão *gold.*

CORPUS	ARQUIVO	PALAVRAS ETIQUETADAS	TOTAL DE ERROS	ACURÁCIA DO ETIQUETADOR – média %
CORPUS COELHO NETTO	aconquista06.gold.txt	2.712	63	97,6
	turbilhao20.gold.txt	1.174	17	98,5
	oenterro.gold.txt	1.532	43	97,1
CORPUS DE CONTRASTE	ocortico05.gold.txt	2.257	39	98,2
	ocortico06.gold.txt	3.165	63	98,0
TOTAL		**10.840**	**225**	**97,9**

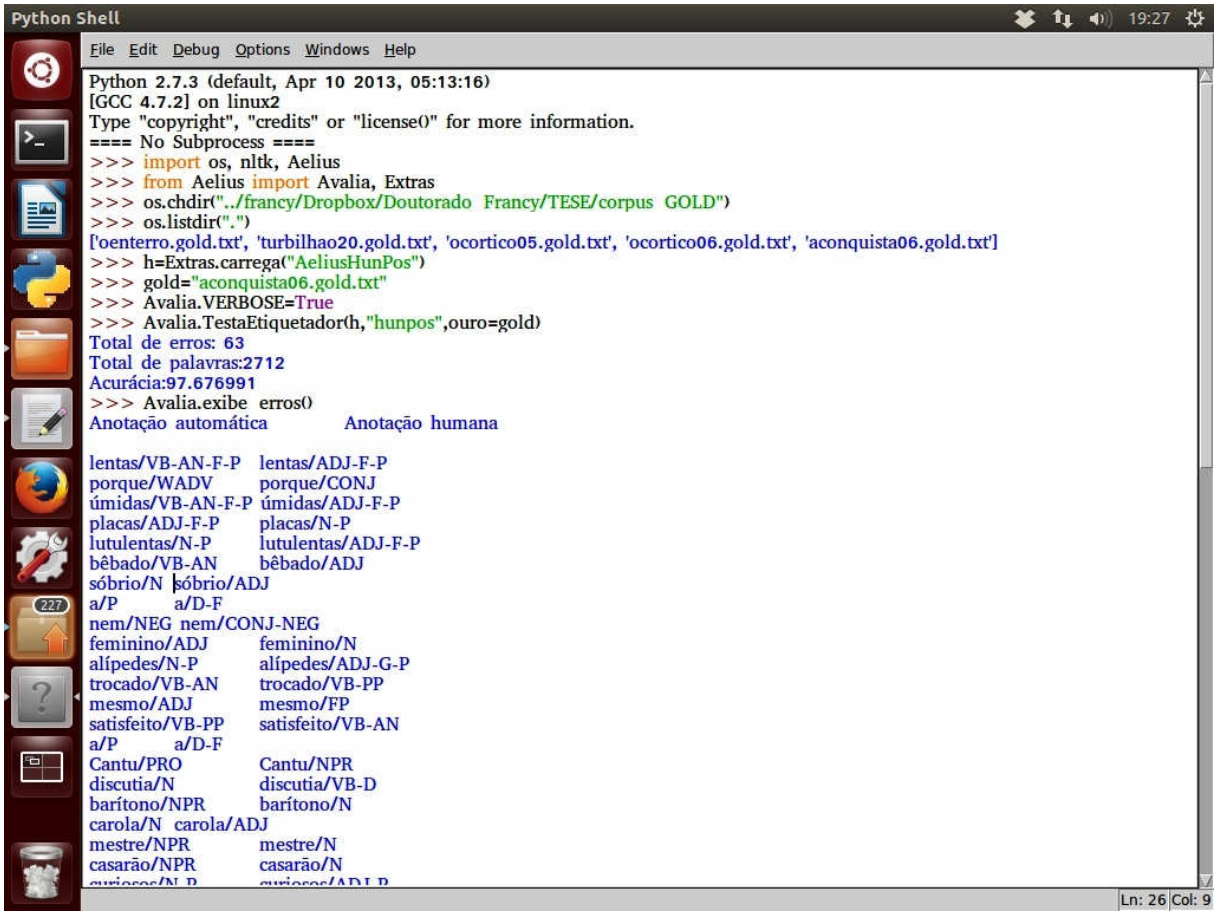
Fonte: Elaborada pela autora

Uma síntese dos erros de anotação (*tags* incorretas) detectados após correção manual é encontrada no Apêndice K.

Com estes dados podemos afirmar que o etiquetador Aelius, modelo AeliusHunpos, ao ser aplicado no arquivo padrão *gold*, corrigidos manualmente, obteve um desempenho (97,9%) superior ao que se refere a literatura. O melhor desempenho alcançado pelo etiquetador até então se obteve com a anotação dos dois primeiros capítulos do romance Luzia-Homem, com 97,47% de acurácia (ALENCAR, 2013a). Temos aí um avanço na melhoria de desempenho do etiquetador de 0,5%.

A função *Avalia.exibe_erros()* do NLTK avaliou equivocadamente alguns dados, que de forma manual observamos ao compararmos os erros cometidos. Damos destaque para alguns erros cometidos pela Anotação Humana ou pela Anotação Automática (figura 17).

Figura 17 - Comando de execução da função *Avalia.exibe_erros*, relacionando os erros cometidos pelo etiquetador



Fonte: Elaborada pela autora

Dentre alguns equívocos, listamos os cometidos no arquivo *"aconquista06.gold.txt"* que foi submetido à correção manual. Gerado o arquivo, o

sistema processou automaticamente as seguintes informações, que foram analisadas fazendo-se a comparação entre a relação dos erros gerada automaticamente e a relação de erros gerada manualmente:

Quadro 11 - Lista de equívocos cometidos pela função *Avalia.exibe_erros*, no arquivo *"aconquista06.gold.txt"* .

Anotação automática	**Anotação humana**	**OBSERVAÇÕES**
porque/WADV	porque/CONJ	O sistema exibiu como erro, mas os corretores manuais não precisaram fazer nenhum tipo de correção.
analisar/VB	analisar/VB-SR	O sistema exibiu como erro, mas os corretores manuais não precisaram fazer nenhum tipo de correção.
Nem/CONJ	Nem/CONJ-NEG	O sistema exibiu como erro por duas vezes, mas os corretores manuais não precisaram fazer nenhum tipo de correção.
Homero/N	Homero/NPR	O sistema exibiu como erro, por duas vezes, mas os corretores manuais não precisaram fazer nenhum tipo de correção.
Toledo/VB-AN	Toledo/NPR	A anotação humana era para ser considerada pelo processador. E a anotação automática para ser considerada como erro. Isso não ocorreu. Não aparece na lista de erros cometidos.
Velhos/ADJ-P	Velhos/N-P	A anotação humana era para ser considerada pelo processador. E a anotação automática para ser considerada como erro. Isso não ocorreu. Não aparece na lista de erros cometidos.

Fonte: Elaborado pela autora

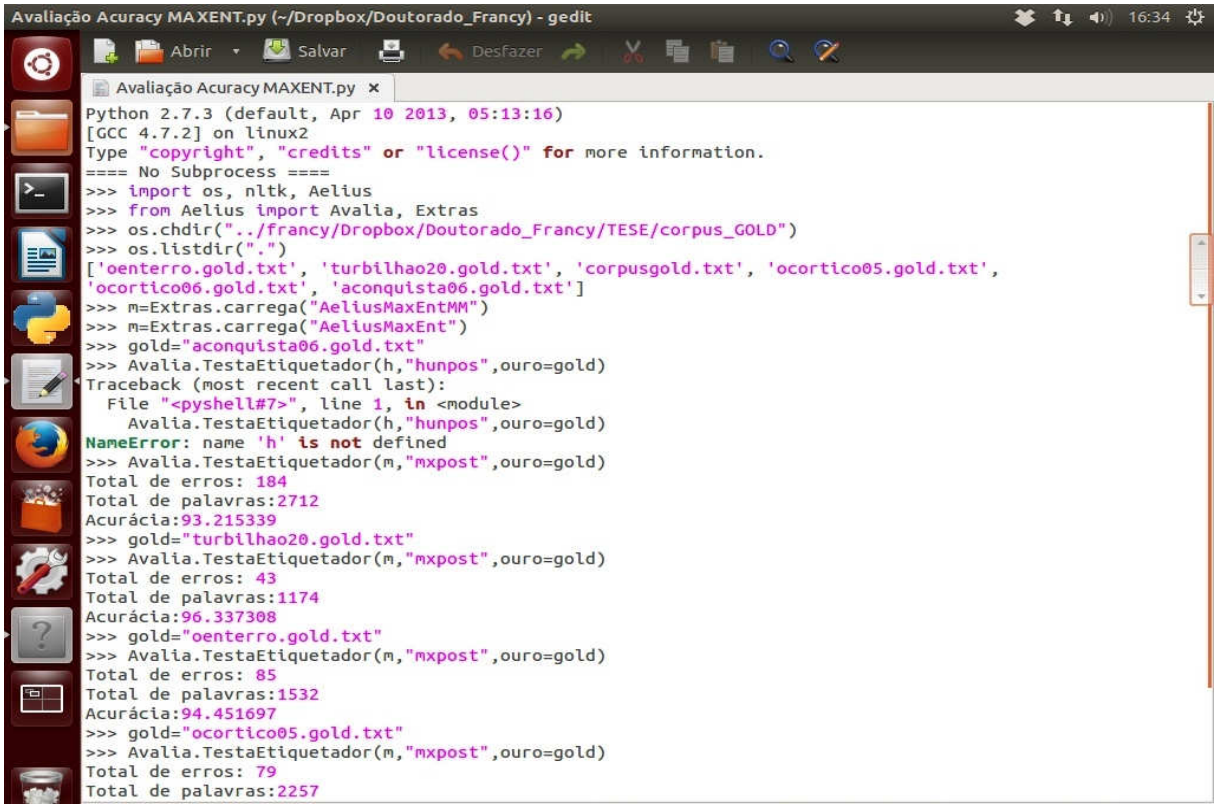
Dos 63 erros que o sistema confirmou que o etiquetador cometeu para este arquivo, manualmente pudemos constatar uma diferença de quantitativo (ver figura acima): na anotação humana ocorreram 60 erros. Isso nos chamou atenção no sentido de observamos pontualmente o ocorrido. Dos quatro primeiros erros do quadro, considerados como erro de anotação automática e corrigido pela anotação humana, são falsos. Os *tokens* não precisavam de correção da anotação, pois estavam etiquetados corretamente, da forma que aparece na coluna da Anotação Humana. Os outros dois *tokens* da lista do quadro (Toledo e Velhos) deveriam ser listados como erros pelo processador, mas não encontramos estes *tokens* na lista de erros gerados pelo processador. Considerando-se que o arquivo *gold* não continha erros no processamento final de correção das etiquetas, tampouco o

arquivo de consenso continha incorreções, entendemos que o Aelius incorreu nesses equívocos de listagem dos erros. É importante ressaltar que dos cinco arquivos submetidos à correção manual, somente no arquivo *"aconquista06.gold.txt"* apareceu essa diferença. Apesar da diferença de *tokens* a mais na lista de erros, entendemos que é importante que o quantitativo seja igual, mas que, sobretudo, a etiquetagem esteja correta.

Entendemos também que os resultados obtidos subsidiarão ações futuras tanto de melhoria do desempenho do Etiquetador AeliusHunPos, quanto de reorganizações de *scripts* ou comandos que corrijam algumas distorções, em especial destaque para a função *Avalia.exibe_erros().*

Avaliamos os mesmos arquivos com o etiquetador Aelius, modelo AeliusMaxEnt, que também foi treinado no Tycho, e obtivemos uma acurácia bem inferior ao HunPos, média de 95,4% (ver Tabela 8), reforçando a qualidade e boa acurácia do AeliusHunPos para anotação morfossintática de textos literários, especialmente os produzidos no séc. XIX e XX.

Figura 18 - Modelo de processamento de avaliação dos arquivos padrão *gold* com o AeliusMaxEnt



Fonte: Elaborada pela autora

Os resultados do processamento demonstrados na figura 18 estão dispostos na Tabela 7:

Tabela 7 - Demonstrativo da acurácia do etiquetador Aelius, modelo AeliusMaxEnt.

CORPUS	ARQUIVO	PALAVRAS ETIQUETADAS	TOTAL DE ERROS	ACURÁCIA DO ETIQUETADOR – média %
CORPUS COELHO NETTO	aconquista06.gold.txt	2712	184	93,21
	turbilhao20.gold.txt	1174	43	96,33
	oenterro.gold.txt	1532	85	94,45
CORPUS DE CONTRASTE	ocortico05.gold.txt	2257	79	96,49
	ocortico06.gold.txt	3165	109	96,55
TOTAL		**10.840**	**500**	**95,40**

Fonte: Elaborada pela autora

O modelo AeliusMaxEnt treinado no Tycho Brahe e anotando textos literários (dois parágrafos iniciais do romance *Luzia-Homem*) obteve uma acurácia de 96,20% (ALENCAR, 2013a), superior aos resultados encontrados neste trabalho. Ressaltamos que este romance é datado de 1903, mesmo período de publicação dos romances dos *Corpora* deste trabalho. Podemos observar então que o MaxEnt se comportou de forma aquém do HunPos quando anotou os textos do padrão *gold* do CCN e do *Corpus* de Contraste, como mostra a Tabela 8.

Tabela 8 – Comparativo de acurácia entre os modelos AeliusHunPos e AeliusMaxEnt, com os arquivos padrão *gold*.

CORPUS	ARQUIVO	PALAVRAS ETIQUETADAS	ACURÁCIA HUNPOS – média %	ACURÁCIA MAXENT – média %
CORPUS COELHO NETTO	aconquista06.gold.txt	2.712	97,7	93,21
	turbilhao20.gold.txt	1.174	98,5	96,33
	oenterro.gold.txt	1.532	97,1	94,45
CORPUS DE CONTRASTE	ocortico05.gold.txt	2.257	98,2	96,49
	ocortico06.gold.txt	3.165	98,0	96,55
TOTAL		**10.840**	**97,9**	**95,40**

Fonte: Elaborada pela autora

Esperamos que os relatórios dos erros relatados, pelo modelo AeliusHunPos, com as devidas intervenções no Etiquetador, criando algoritmos específicos para resolver as ambiguidades encontradas, aumentem mais ainda o

seu desempenho, quiçá chegando ao patamar considerado por Leech (2004) como de melhor precisão atingindo entre 99% a 99,5%.

Utilizando o programador Python, a biblioteca NLTK, o programa Aelius e especificamente o modelo AeliusHunpos, e todos os outros dispositivos necessários para o tratamento computacional dos traços linguísticos do CCN e do *Corpus* de Contraste, como já referenciado, realizamos a etapa das análises linguístico-computacionais.

Seguir corretamente todas essas etapas possibilitou-nos realizar o trabalho de acordo com os objetivos estabelecidos para sua consecução, que consiste primordialmente na compilação e anotação do *Corpus* Coelho Netto, para daí então realizar qualquer outra etapa subsequente.

4.2 Análises e Resultados

Especificamente, nas análises deste trabalho, serão relevantes os resultados da observação quanto à 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 riqueza vocabular afirmada pelo seu filho Coelho Netto e pela crítica literária que se debruçou em analisar sua obra, ou então negar tudo isso. Como forma de testar essas hipóteses, criou-se o *Corpus* de Contraste, detalhado no item 3.2, referência de escritores do mesmo momento literário de CN e que sofreram das mesmas críticas a que foi submetido Coelho Netto: no caso, Aluísio Azevedo e Camilo Castelo Branco.

As análises linguístico-computacionais realizadas tanto no *Corpus* Coelho Neto quanto no *Corpus* de Contraste são basicamente:

- Verificação da diversidade lexical: busca-se a quantidade de *tokens* (unidades lexicais inclui palavras e sinais de pontuação); quantidade de *tokens* alfabéticos (conta somente palavras); quantidade de *hapaxes* (mostra a quantidade de palavras, sem repetição: palavras diferentes no texto que só ocorrem uma vez); e cálculo da diversidade lexical (mostra a estatística da diversidade lexical do *corpus* em análise);
- Frequência e diversidade de verbos;
- Frequência e diversidade de adjetivos;

- Frequência e diversidade de advérbios em *–mente*.

Além destas, considerando-se a automatização de busca de informações por meio dos processadores automáticos utilizados neste trabalho, outras análises foram realizadas, como: a frequência das classes de palavras nos Corpora e o detalhamento de quais advérbios em *–mente*, verbos e adjetivos são mais utilizados em cada *corpus*.

Após a análise detalhada dos *Corpora*, procedemos com a comparação entre eles para validação ou não do que a crítica modernista apregoava a Coelho Netto e de outros que se debruçaram a analisar seus textos, com discursos como os já destacados no item 2.1.2 deste trabalho, que retomamos aqui:

- "rebuscador de palavras difíceis";
- "não costumava usar a mesma palavra em trechos pequenos";
- "um anseio pela adjetivação numerosa, rara e erudita";
- "malabarista do verbo";
- "submissão ao culto do adjetivo";
- "libidinoso da adjetivação";
- "possuidor de um vernáculo portentoso, ressuscitando um verbalismo inatual".

4.2.1 Da Diversidade Lexical:

Retomando o que já detalhamos no item 2.6, a diversidade lexical dos *Corpora* está vinculada aos seguintes dados encontrados: quantidade de *tokens;* quantidade de *tokens* alfabéticos; quantidade de *hapaxes;* e cálculo da diversidade lexical (mostra a estatística da diversidade lexical em porcentagem).

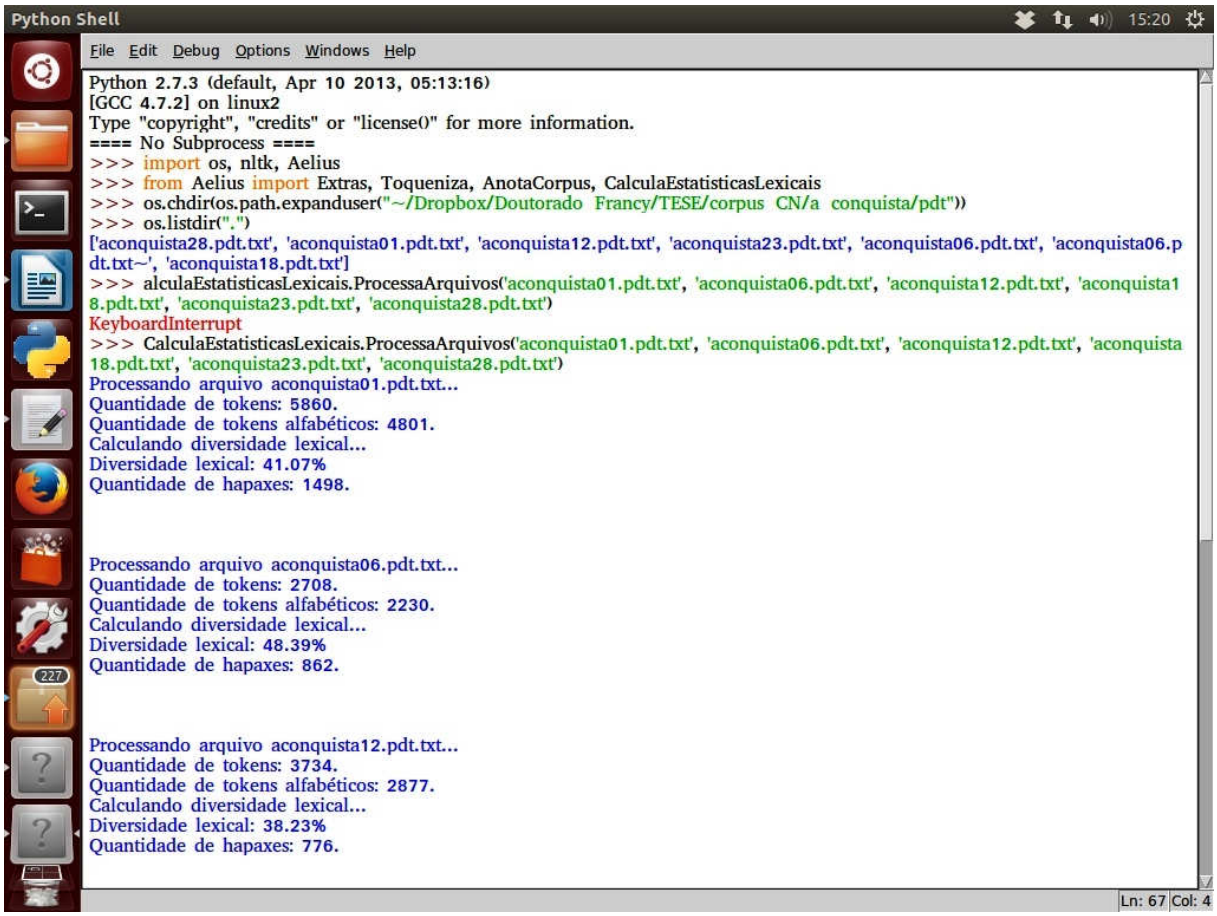
As análises que realizamos nos *Corpora* desta Tese, precipuamente, que mostram a Diversidade Lexical, podem ser observadas no exemplo de comando realizado na figura 19.

Esses dados, obtivemos processando os arquivos já editados (comparados impresso com digital) mas não anotados. Nesse caso, utilizamos os arquivos no formato *".pdt.txt"*. Preferimos esse formato, pois a quantidade de *tokens*, alfabéticas, diversidade lexical e *hapaxe*s modificam-se em relação ao texto digital arquivado no formato *".txt"*. Os resultados servem como base de comparação para

as etapas seguintes. A cada formato de arquivo gerado, novos elementos vão surgindo, modificando os dados e os resultados encontrados.

Outros indicadores vão influenciar nesses dados, como as informações obtidas após a fase de comparação impresso com digital e após a fase de correção manual.

Figura 19 – Exemplo de extração de dados dos arquivos dos *Corpora.*



Fonte: Elaborada pela autora

Como resultado dessa análise, o quando abaixo mostra os seguintes dados:

Tabela 9 - Demonstrativo de diversidade lexical do CCN por capítulos/contos

CORPUS COELHO NETTO	TOKENS	TOKENS ALFABÉTICOS	DIVERSIDADE LEXICAL %	QTADE DE HAPAXES
aconquista01	5.860	4.801	41,07	1.498
aconquista06	2.708	2.230	48,43	862
aconquista12	3.734	2.877	38,23	776
aconquista18	2.463	1.946	44,19	628

aconquista28	3.414	2.732	44,18	911
aconquista28	3.982	3.143	40,15	935
SUBTOTAL	**22.161**	**11.729**	**42,7**	**5.610**
turbilhao01	3.637	2.936	47,55	1.130
turbilhao05	4.991	3.971	38,28	1.108
turbilhao10	7.737	6.023	33,55	1.407
turbilhao15	3.737	2.825	44,96	961
turbilhao20	1.169	897	51,62	355
turbilhao25	2.033	1.573	42,15	469
SUBTOTAL	**23.304**	**18.225**	**43,0**	**5.430**
oenterro	1.528	1.257	52,51	526
mandovi	4.219	3.303	34,12	742
firmo	1.868	1.495	45,22	494
SUBTOTAL	**7.615**	**6.055**	**43,9**	**1.762**
TOTAL	**53.080**	**42.009**	**-**	

Fonte: Elaborada pela autora

Com esses dados primários, e por capítulos, constatamos que o escritor tem praticamente a mesma média de diversidade lexical em todos os seus textos, confirmando uma regularidade na sua riqueza vocabular.

Observamos nos dados acima que se totalizássemos a diversidade lexical e a quantidade de *hapaxes* teríamos dados incoerentes com a realidade, pois uma mesma palavra pode ocorrer simultaneamente em outro capítulo, por isso geramos uma tabela organizando os arquivos por obra ao invés de capítulos, assim temos uma panorâmica por obra, e um detalhamento melhor da diversidade lexical de forma generalista.

Tabela 10 - Demonstrativo de diversidade lexical do CCN, por obra.

CORPUS COELHO NETTO	TOKENS	TOKENS ALFABÉTICOS	DIVERSIDADE LEXICAL %	QTADE DE HAPAXES
Aconquista	22.161	11.729	28,55	3.461
turbilhao01	23.304	18.225	26,44	3.177
Sertaocontos	7.615	6.055	33,15	1.342
TOTAL	**53.080**	**42.009**	**-**	

Fonte: Elaborada pela autora

Com os dados coadunados, os resultados demonstram uma totalização e percentual bastante diferente do exposto na tabela 9, isso em função da repetição de alguns termos em capítulos diferentes. De qualquer forma, os resultados nos permitem confirmar a regularidade de diversidade lexical no CCN.

Submetemos o *Corpus* de Contraste aos mesmos processos e obtivemos os dados dispostos nas tabelas 11 e 12.

Tabela 11 - Demonstrativo de diversidade lexical do *Corpus* de Contraste por capítulos.

***CORPUS* DE CONTRASTE**	**TOKENS**	**TOKENS ALFABÉTICOS**	**DIVERSIDADE LEXICAL %**	**QTADE DE HAPAXES**
ocortico01	6.088	5.120	33,22	1.192
ocortico02	4.495	3.747	36,54	994
ocortico03	4.950	4.114	37,77	1.121
ocortico04	3.305	2.707	39,45	755
ocortico05	2.251	1.920	43,33	631
ocortico06	3.153	2.624	41,50	784
ocortico07	6.345	5.282	34,89	1.310
ocortico08	5.583	4.483	34,02	1.050
SUBTOTAL	**36.170**	**29.997**	**37,6**	**7.837**
amordeperdicaoINT	536	444	56,53	192
amordeperdicao01	2.999	2.566	39,01	745
amordeperdicao02	2.156	1.831	46,42	674
amordeperdicao03	2.704	2.251	37,49	607
amordeperdicao04	2.272	1.923	41,91	602
amordeperdicao05	3.214	2.645	34,14	596
amordeperdicao06	3.920	3.164	32,81	686
SUBTOTAL	**17.801**	**14.824**	**41,2**	**4.102**
TOTAL	**53.791**	**44.821**	**-**	

Fonte: Elaborada pela autora

Tabela 12 - Demonstrativo de diversidade lexical do *Corpus* de Contraste, por obra

CORPUS DE CONTRASTE	TOKENS	TOKENS ALFABÉTICOS	DIVERSIDADE LEXICAL %	QTADE DE HAPAXES
Ocortiço	36.170	29.997	19,90	3.605
amordeperdicao	17.801	14.824	24,91	2.499
TOTAL			**-**	

Fonte: Elaborada pela autora

Depurando mais ainda os dados, geramos outra tabela agora com os dados concentrados por *Corpus*, e obtivemos os seguintes resultados.

Tabela 13 – Resultado da diversidade lexical do CCN e do *Corpus* de Contraste

CORPORA	DIVERSIDADE LEXICAL %	QTADE DE HAPAXES
CCN	21,37	5.622
Corpus de Contraste	18,19	4.004
TOTAL	**-**	

Fonte: Elaborada pela autora

Em comparação ao *Corpus* de Contraste, que apresenta uma diversidade lexical em média de 18,19%, o CNN tem uma diversidade lexical maior, de 21,37%.

Os textos de Aluísio Azevedo apresentam menor diversidade que os de Camilo Castelo Branco, respectivamente 19,9% e 24,9%, confirmando o que disse Broca (1958) e Vilela (2008) sobre a riqueza léxica de Camilo, mostrando-se similar à de Coelho Netto.

Nenhum destes dois escritores, portanto, se sobrepõem à riqueza vocabular de Coelho Netto, confirmando assim o que dizia a crítica, levando-nos a depreender que a afirmação de seu filho de fato se aproxima da verdade, quando disse que CN possuía “o mais rico vocabulário da língua portuguesa, calculado em 20.000 palavras” (NETTO, P., 1958, p. XCIX-C), e também Mendes e Ignácio (2010) que ressaltam que o escritor buscava sempre expor novos vocábulos. Para que calculássemos de fato todo o léxico do escritor, teríamos de compilar toda a sua obra: aproximadamente 73 livros entre romances, contos, narrativas históricas, crônicas, teatro; e mais de 8.000 artigos (NETTO, P., 1964, 1958; NETTO, C., 1958; BIBLIOTECA NACIONAL DO RJ, 1964).

A riqueza lexical de um texto está diretamente relacionada com a *hapax legomena.* Essas palavras que têm apenas uma ocorrência em um *corpus* influenciam no resultado da riqueza lexical, e essa proporção depende de características estilísticas ou linguísticas, e do comprimento dos textos (CÚRCIO, 2013). A riqueza lexical de um escritor pode ser entendida da seguinte forma: “quanto maior o número de vocábulos novos, maior será a riqueza e a variedade do vocabulário” (CÚRCIO, 2013, 84). Em sua tese de doutorado, Cúrcio (2013) fez uma análise da obra de João Guimarães Rosa para detectar dados sobre o léxico e o estilo rosiano. Dos textos analisados, foi possível perceber que mais da metade do vocabulário de Rosa não se repete (aproximadamente 52,8%).

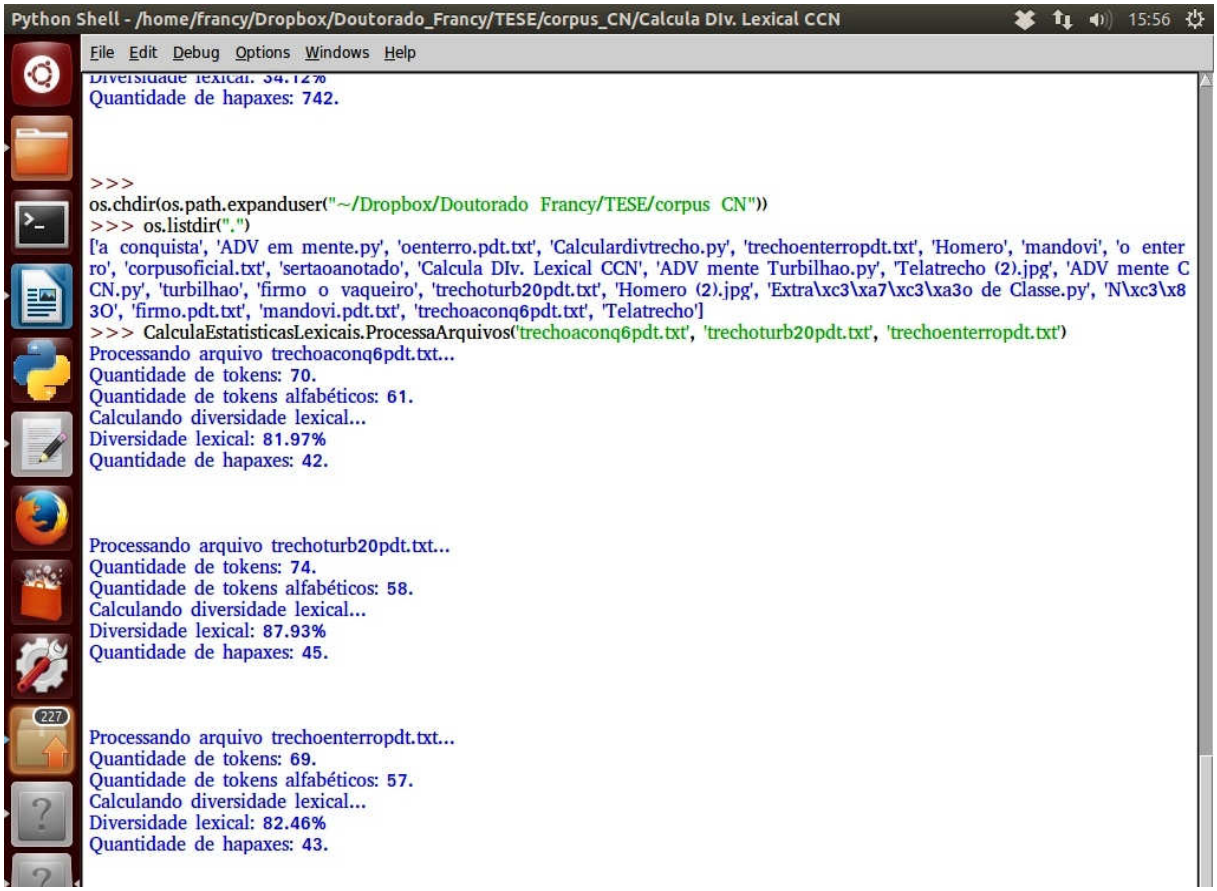
A ocorrência corresponde à forma que aparece uma vez em um *corpus*, e de acordo com a quantidade de vezes em que aparece consideramos isso como uma frequência (CÚRCIO, 2013). Os dados expostos na Tabela 9 mostram que dos mais de 53 mil *tokens* (soma de palavras e pontuações) do CCN, 42 mil são somente palavras. Destas, 6.622 palavras aparecem somente uma vez no CCN (*hapaxes*), ou seja, 13% do total de palavras. Quando se compara a riqueza lexical de Coelho Netto com a de Guimarães Rosa, concluímos que o vocabulário deste é muito mais criativo.

Diante da referência de que CN “não costumava usar a mesma palavra em trechos pequenos” (MENDES; IGNÁCIO, 2010, p. 2), buscamos no CCN verificar se essa informação procedia, para isso utilizamos comandos em *Python* para essa verificação. Destacamos três arquivos com maior diversidade lexical, de cada livro, para verificar a diversidade em pequenos trechos de cada arquivo.

Os arquivos escolhidos foram: *aconquista06.pdt.txt, turbilhao20.pdt.txt*, e *oenterro.pdt.txt*, coincidentemente os mesmos textos escolhidos para a correção manual da anotação morfossintática. Coincidência, porque não tínhamos ainda, no momento da escolha para correção, analisada a diversidade lexical dos textos. O que levamos em consideração era que os textos para a correção manual deveriam ter aproximadamente 10% do *corpus,* e a escolha foi feita fazendo somatórios mais aproximados deste percentual, e que de alguma forma pudesse contemplar um arquivo de cada livro.

A figura 20 mostra a interface de Python com os comandos realizados para a obtenção dos dados descritos acima:

Figura 20 - Exemplo de comandos de processamento para cálculo da diversidade lexical de trechos do CCN.



Fonte: Elaborada pela autora

Em uma primeira tentativa de encontrar trechos similares no tamanho e quantitativo de *tokens*, a escolha do trecho do livro *A Conquista* não foi satisfatória. Para a escolha dos trechos, levamos em consideração principalmente a quantidade de linhas de cada um (quatro linhas). Nesse caso o trecho tinha quatro linhas, mas com um quantitativo de *tokens* inferior aos outros escolhidos. Procedemos então com nova escolha, e encontramos trechos equitativos.

Resumindo os dados encontrados, observamos a seguinte diversidade nos trechos destacados na Tabela 14:

Tabela 14 - Diversidade lexical de trechos do CCN

TRECHOS *CORPUS* COELHO NETTO	TOKENS	TOKENS ALFABÉTICOS	DIVERSIDADE LEXICAL %	QTADE DE HAPAXES
aconquista06	70	61	81,97	42
turbilhao20	74	58	87,93	45
Oenterro	69	57	82,46	43

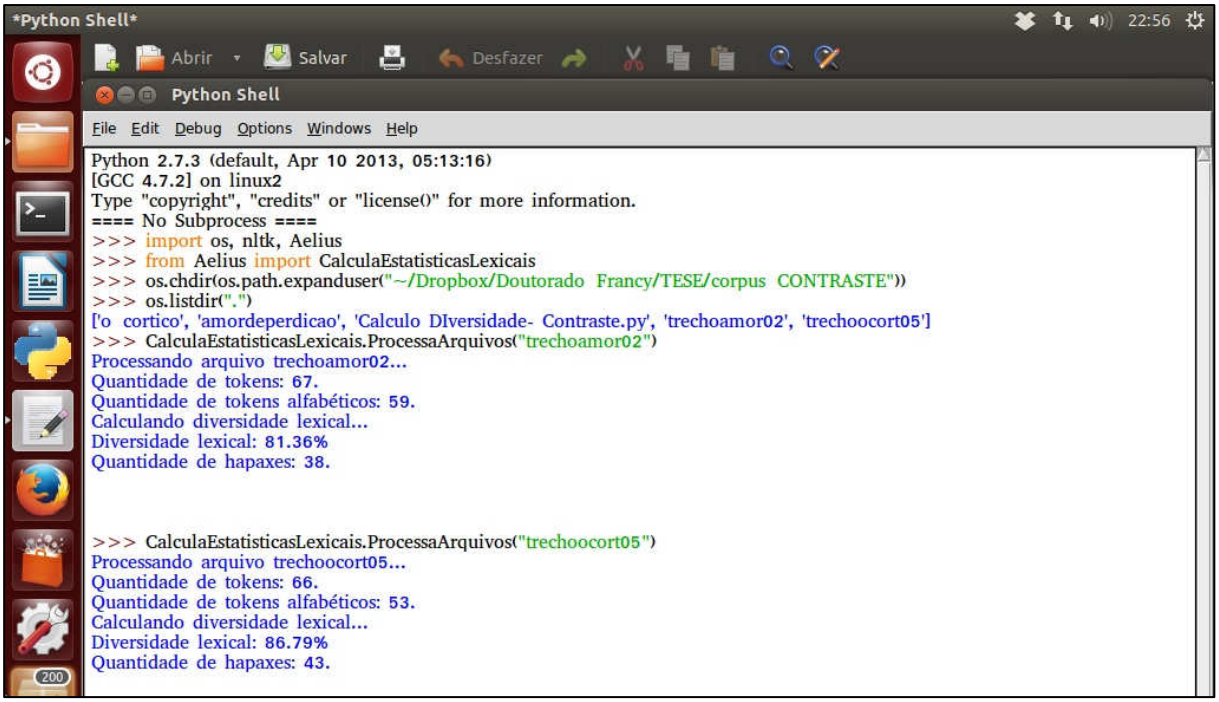
Fonte: Elaborada pela autora

Esses dados nos revelam que, de fato, o grau de repetição é muito pequeno em cada trecho. Com um percentual de, em média, 84% de palavras que aparecem somente uma vez nos trechos (*hapaxes*), percebemos então a diversidade vocabular de Coelho Netto ao produzir seus textos. CN foi considerado também como um dos "cultores da forma" (CANDIDO, 1999, p. 61) e foi enquadrado na escola parnasiana, sobretudo pelo cuidado da escrita e pelo preciosismo vocabular. Importante destacar que, como bem mencionou Maria Eugenia Boaventura em uma entrevista falando sobre Macunaíma, no Modernismo os padrões narrativos não primavam pela riqueza vocabular (BOAVENTURA, 2008).

Realizando o mesmo processo no *Corpus* de Contraste, destacamos um trecho de cada livro deste *Corpus*, para fazer a comparação quanto à repetição ou não de muitas palavras em pequenos textos. O critério de escolha foi o mesmo do CCN: quantidade de linhas e maior diversidade lexical em cada capítulo. Os trechos escolhidos são dos arquivos: "*amordeperdicao02.txt"* e "*ocortico05.txt"*.

A Figura 21 mostra a interface de Python com os comandos realizados para obtenção dos dados pretendidos:

Figura 21 - Exemplo de comandos de processamento para cálculo da diversidade lexical de trechos do *Corpus* de Contraste



Fonte: Elaborada pela autora

Resumindo os dados encontrados, observamos a seguinte diversidade, nos trechos destacados na Tabela 15:

Tabela 15 - Diversidade lexical de trechos do *Corpus* de Contraste

TRECHOS CORPUS DE CONTRASTE	TOKENS	TOKENS ALFABÉTICOS	DIVERSIDADE LEXICAL %	QTADE DE HAPAXES
amordeperdicao02	67	59	81,36	38
ocortico05	66	53	86,79	43

Fonte: Elaborada pela autora

Fazendo uma comparação entre os *Corpora*, percebemos que a riqueza vocabular é similar, com a mesma média de 84%, por isso talvez Aluísio Azevedo e Camilo Castelo Branco tenham sido submetidos à mesma crítica destinada a Coelho Netto: os modernistas os consideravam como rebuscados e prolixos.

Biderman (1998, p. 176) justifica o uso do vocábulo rico em textos literários, porque a

> linguagem literária também registra elevado número de palavras raras. Às vezes são criações idiossincráticas resultantes de uma característica típica da arte: a busca da inovação. O artista viola a norma por razões estéticas, aproveitando as virtualidades de criação que o sistema lexical lhe permite e propicia. O criador literário deseja exatamente não escrever como o vulgo e evita o vocábulo banal, usual.

Zyngier, Viana e Silveira (2011, p. 100) também se referem a essa inovação quando afirmam: “os textos literários são marcados por inovação e imprevisibilidade no uso da língua”.

Brandão (1979, p. 115), reforçando o modo de produção dos escritores maranhenses, destaca que:

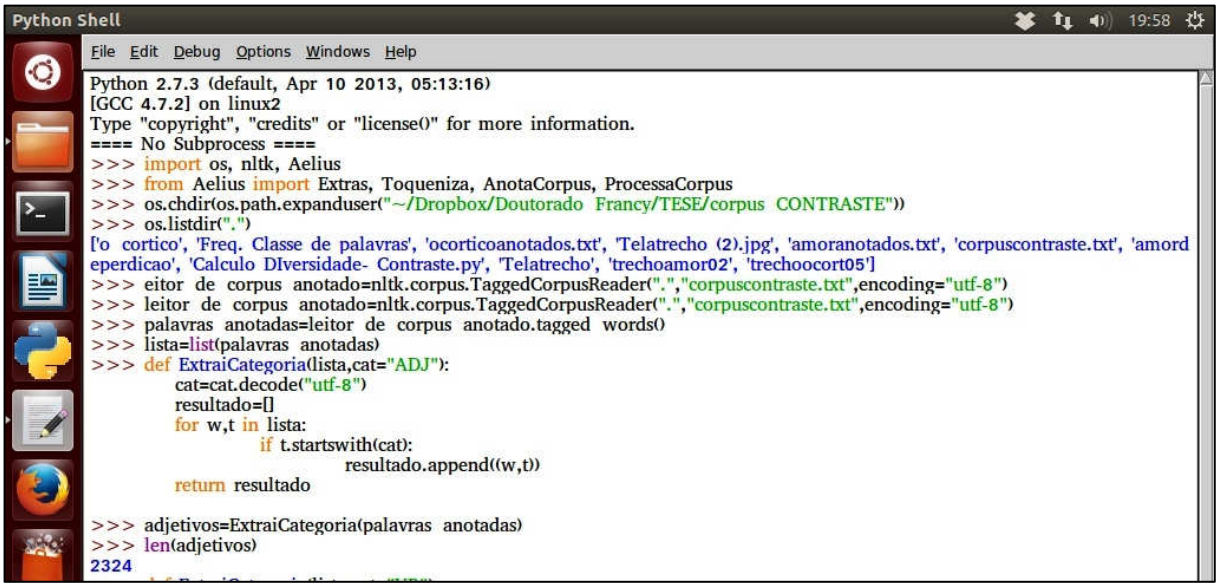
> Sob o ponto de vista da língua, os maranhenses colocam-se como cultores do bom português, de tendências conservadoras, mas nem por isso anacrônico. A língua que utilizam traz a vitalidade necessária para fazer dela um instrumento eficaz e atual. Esse é o ponto em que, apesar das divergências ideológicas e estéticas, todos eles se encontram de maneira mais estreita.

E nesse cenário, encontramos tanto Coelho Netto quanto Aluísio Azevedo.

Realizamos análises para identificar a frequência das “palavras instrumentais” (BIDERMAN, 1998, p. 161), que são os artigos, conjunções, preposições, pronomes, interjeições, numerais, pois devido ao seu número reduzido de formas e de ter função sintática, aparecem com frequência bem maior que as “palavras lexicais e plenas” (p. 167), que são os substantivos, adjetivos, verbos, advérbios. Para isso, utilizamos os *Corpora* anotados, com arquivos no formato “*.pdt.hunpos.txt*”.

Inicialmente, buscamos os dados de frequência dessas classes de palavras nos *Corpora*. A figura 22 mostra a execução dos comandos em Python para a obtenção dos dados dispostos na Tabela 16.

Figura 22 - Exemplo de comandos para obtenção das frequências das classes de palavras nos *Corpora.*



Fonte: Elaborada pela autora

Nesta tabela mostramos a lista de ocorrências das classes de palavras nos *Corpora* em análise, destacando o quantitativo de cada *Corpus*, observando qual classe de palavra é mais ocorrente em cada um.

Tabela 16 - Frequência de classes de palavras dos *Corpora*.

	CLASSE DE PALAVRAS	***CORPUS* COELHO NETTO**	***CORPUS* DE CONTRASTE**
PALAVRAS INSTRUMENTAIS	ARTIGOS	4.423	4.384
	CONJUNÇÕES	2.285	2.472
	PREPOSIÇÕES	6.930	9.020
	PRONOMES	890	1.449
	INTERJEIÇÕES	75	46
	NUMERAIS	174	371
PALAVRAS PLENAS	SUBSTANTIVOS	12.323	13.366
	ADJETIVOS	2.552	2.324
	VERBOS	8.487	8.401
	ADVÉRBIOS	1.674	2.002
	TOTAL	**38.487**	**42.835**

Fonte: Elaborada pela autora

O etiquetador considera além das palavras outros elementos na anotação, como contração, palavras estrangeiras, pontuação etc (ver apêndice F). Destacamos somente as classes de palavras, para dar uma dimensão generalizada do quantitativo lexical presente no *Corpus* Coelho Netto e no *Corpus* de Contraste.

Em uma visão generalista, é possível notar que o léxico do *Corpus* de Contraste (com obras de Aluísio Azevedo e Camilo Castelo Branco) possui mais itens que o *Corpus* Coelho Netto, e isso depõe contra a crítica modernista sobre a riqueza vocabular exacerbada imposta a esse escritor. Na verdade, os textos de CN só se destacam com frequência maior em artigos, interjeições, verbos e adjetivos.

Quanto à frequência das palavras em um texto, Biderman (1998, p. 169) afirmava que “as palavras gramaticais (ou instrumentais)[41] e certo grupo de verbos são as classes mais estáveis da língua, são consideradas palavras multiuso, pois aparecem em qualquer texto”. E elas são igualmente frequentes em todos os tipos de textos literários, seja qual for o gênero ou o tema tratado. Cúrcio (2013, p. 47) corrobora essa informação sobre frequência quando diz que “na maioria dos casos, as formas mais frequentes que surgem no topo da listagem são as palavras gramaticais, como os artigos, os dêiticos, as preposições, os pronomes, as conjunções etc”.

O quadro abaixo relaciona as 60 palavras mais frequentes no CCN, por ordem decorrência, confirmando assim as afirmações desses pesquisadores:

Quadro 12 - Lista das 60 palavras mais frequentes no CCN por ordem de ocorrência.

ORDEM	PALAVRA	ORDEM	PALAVRA	ORDEM	PALAVRA
1	a	21	ao	41	há
2	o	22	no	42	minha
3	e	23	lhe	43	sem
4	de	24	por	44	porque
5	que	25	mas	45	Anselmo
6	se	26	dos	46	casa
7	um	27	eu	47	até
8	com	28	quando	48	então
9	não	29	mais	49	ou

[41] As palavras instrumentais, no português brasileiro contemporâneo, são os “artigos, pronomes, preposições, conjunções, advérbios, numerais, e algumas palavras lexicais ou plenas das classes substantivo, adjetivo e verbo” (BIDERMAN, 1998, p.167).

10	os	30	me	50	grande
11	da	31	olhos	51	lá
12	do	32	era	52	muito
13	em	33	homem	53	rua
14	uma	34	ele	54	às
15	as	35	foi	55	sobre
16	como	36	das	56	dona
17	é	37	aqui	57	está
18	para	38	já	58	onde
19	à	39	nos	59	cabeça
20	na	40	aos	60	sua

Fonte: Elaborado pela autora

Berber Sardinha (2004) relata que no *Corpus* de Referência do Português Contemporâneo (CRPC) as dez palavras mais frequentes são, por ordem de ocorrência: *de, a, o, e, que, do, da, em, para, no*.

Ao compararmos o CRPC com o CCN, observamos que as cinco palavras mais frequentes nos *corpora* coincidem, não por ordem de ocorrência, mas estão entre as primeiras mais frequentes.

Quadro 13 – Comparação das palavras mais frequentes entre CRPC e CCN

ORDEM	**CCN**	**CRPC**
1	a	de
2	o	a
3	e	o
4	de	e
5	que	que

Fonte: Elaborado pela autora

No que se refere à frequência de adjetivos e verbos, o CCN é muito mais rico que o *Corpus de Contraste*. Esses dados reforçam o que a crítica dizia sobre a riqueza vocabular de sua produção, sobretudo quanto ao uso de adjetivos e verbos. Se considerarmos que a diferença entre os *Corpora* quanto à riqueza vocabular é de quase 2% (vide Tabela 13), é possível afirmar que a excelência vocabular de Coelho Netto não era assim tão acentuada, principalmente comparado com escritores do mesmo período literário, no caso Aluísio e Camilo. Portanto, não parecia “fugir à regra”.

Avaliando os dados dos Corpora deste trabalho (Tabela 16) constatamos que as categorias mais frequentes são as das palavras plenas ou lexicais, com a frequência maior de substantivos e, em sequência, dos verbos. Esses dados confrontam-se com os resultados das pesquisas relatadas acima, que destacam as palavras gramaticais como as mais frequentes nos textos literários.

4.2.2 Da frequência e diversidade de Adjetivos:

Quanto ao uso “exagerado” de adjetivos, que para os críticos CN tinha verdadeiro culto, ou então por ser considerado um “libidinoso da adjetivação”, buscamos no CCN a frequência dos adjetivos e sua diversidade.

Na tabela 17, mostramos uma lista de ocorrências nos *Corpora* em análise, destacando o quantitativo e as *hapaxes* dos adjetivos por livro, e por *corpus*. Essa ação permite que visualizemos mais detalhadamente o léxico de cada texto, e não o *corpus* como um todo.

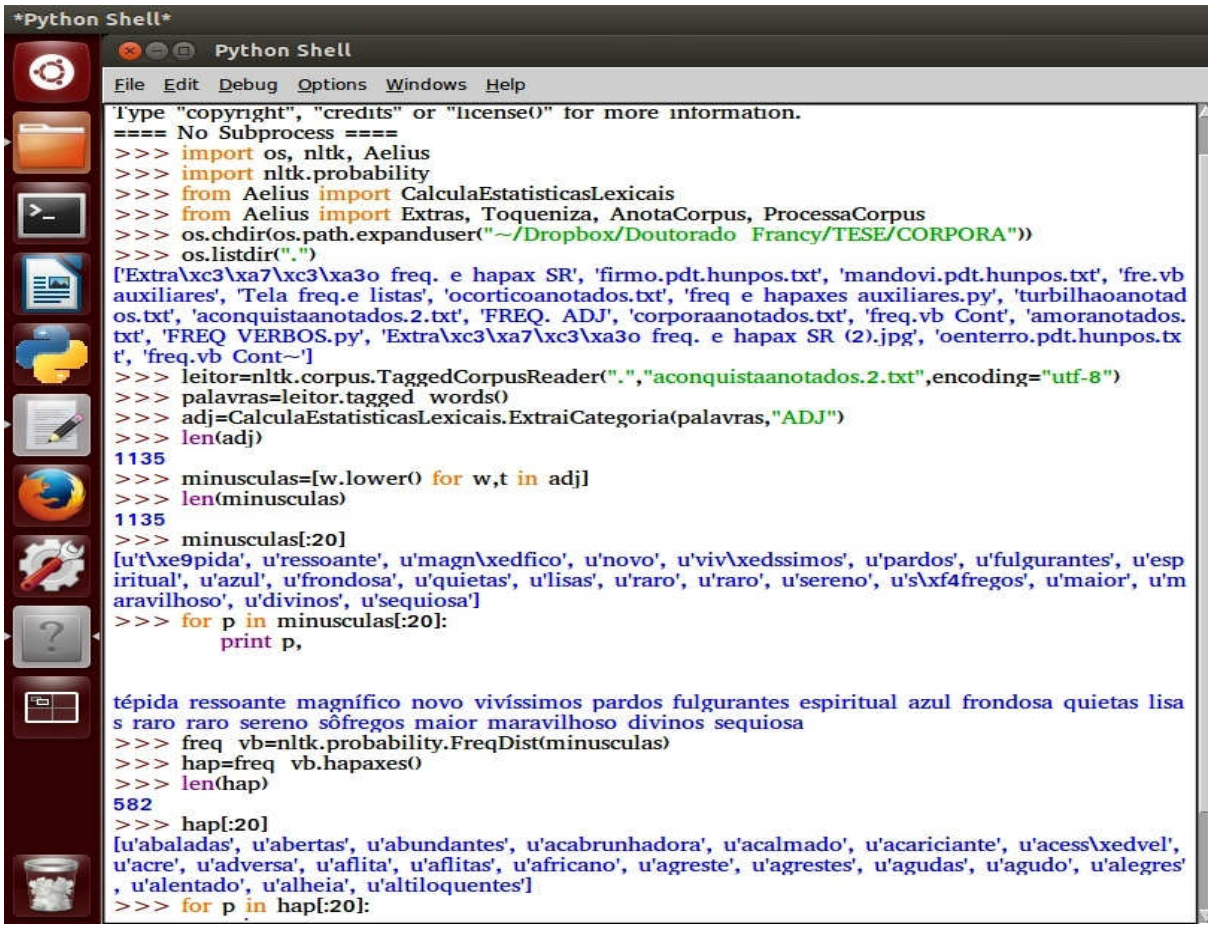
Tabela 17 - Lista de ocorrências de **adjetivos** e *hapaxes* nos *Corpora*

TEXTO	ADJETIVOS	HAPAX	% HAPAX
A Conquista	1.135	594	52,3
Turbilhão	1.010	459	45,4
O enterro	129	108	83,7
Mandovi	180	107	59,9
Firmo, o vaqueiro	98	64	65,3
TOTAL CCN	**2.552**	**1.332**	**61,3**
O cortiço	1.692	530	31,3
Amor de Salvação	632	355	56,1
TOTAL CORPUS CONTRASTE	**2.324**	**885**	**43,7**

Fonte: Elaborada pela autora

Esses resultados foram extraídos do processamento dos comandos mostrados na figura 23. Esses comandos possibilitam extrair outros dados, dentre eles: lista de palavras *hapax* e lista dos verbos mais frequentes em cada arquivo.

Figura 23 - Comandos de extração de frequências, *hapaxes*, lista de adjetivos.



Fonte: Elaborada pela autora

Os dados encontrados revelam uma frequência similar de adjetivos do CCN em comparação ao *Corpus* de Contraste, respectivamente 2.552 e 2.324 ocorrências.

Destacando as *hapaxes*, observamos uma incidência bem elevada de palavras raras no CCN em comparação ao *Corpus* de Contraste. Organizamos as tabelas por obras e não por *Corpus,* no intuito de visualizarmos melhor a frequência de cada um. Se organizássemos por *Corpus* teríamos um percentual diferente, mas a média percentual não seria tão díspare.

Apesar de uma frequência similar de adjetivos nos *Corpora*, o percentual de *hapaxes* não está diretamente relacionado com a quantidade. Isso é percebido analisando-se as diferenças nos valores das *hapaxes*: 61,3% e 43,7% respectivamente. Os dados confirmam então a habilidade de CN em utilizar palavras raras nos seus textos. Não repetir as palavras era uma faceta na produção do

escritor, não propositalmente, mas em função de seu rico vocabulário e sintaxe ampla (LIMA, 1958).

Os comandos executados nos possibilitam ter a lista completa de todas as *hapaxes* de cada arquivo. Para exemplificar alguns adjetivos usados no CCN somente uma vez, citamos 20 deles.

Exemplo 9:

abaladas, abertas, absurdas, abundantes, acabrunhadora, acalmado, acariciante, acessível, acre, adversa, africano, afável, agonizante, agradável, agreste, agrestes, agudo, alentado, alheio, altiloquentes.

Considerando-se os dados encontrados quanto à quantidade de adjetivos presentes nos seis capítulos da obra *Turbilhão*, podemos rebater a crítica feita à Coelho Netto quanto a esse livro feita por Mendes e Ignácio, ao dizer que ele mostrava “um anseio pela adjetivação numerosa, rara e erudita” (MENDES; IGNÁCIO, 2010, p. 2). Comparando *Turbilhão* com *A Conquista* observamos que este contém mais adjetivos e mais *hapaxes* que aquele (Tabela 17).

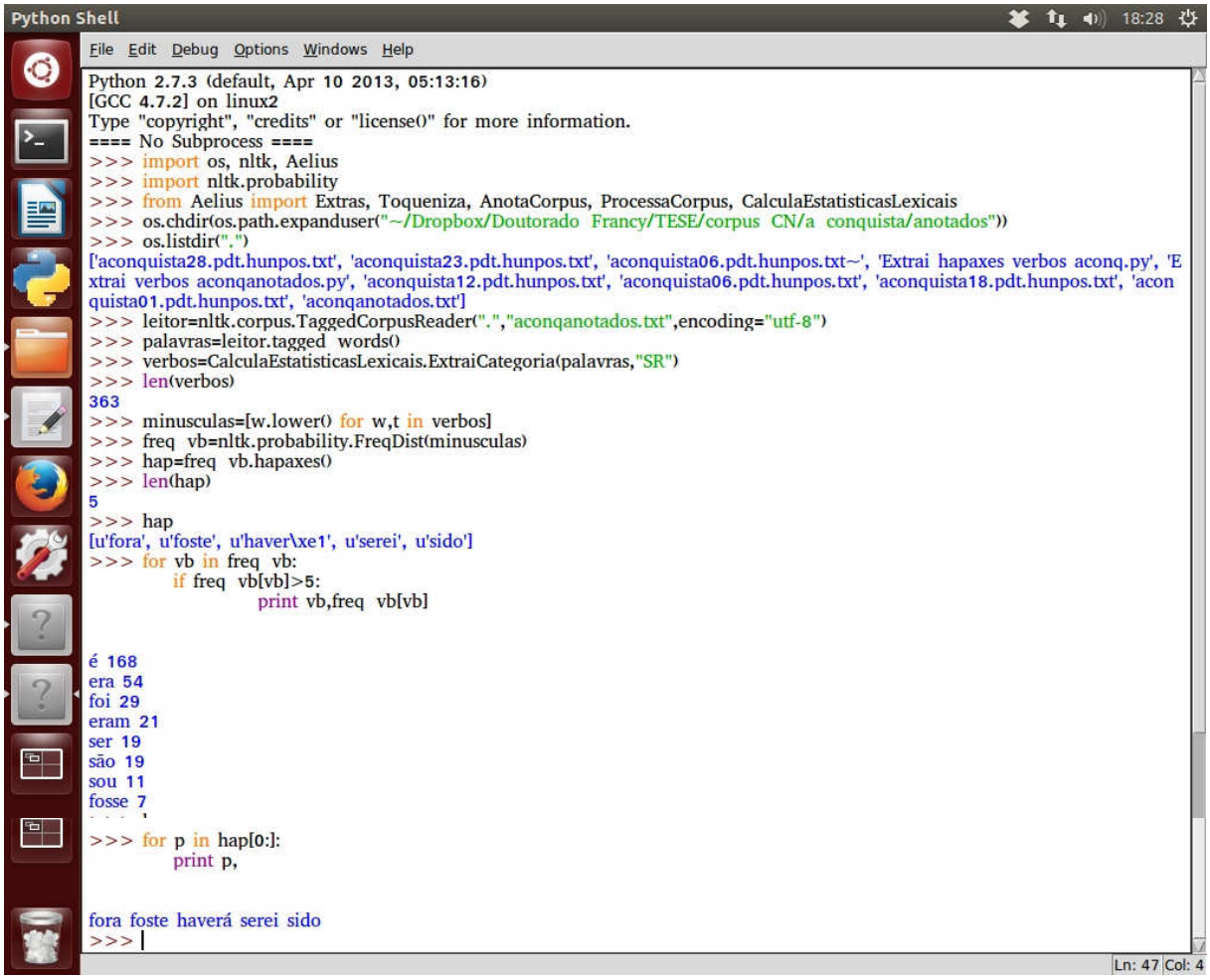
4.2.3 Da frequência e diversidade de Verbos:

Diante do “malabarismo verbal” de CN, e por estar sempre em busca de novas palavras, verificamos no *Corpus* mais que a quantidade de verbos que aparecem em seus textos, mas principalmente a diversidade dos verbos apresentados, podendo confirmar diante dos dados generalistas expostos na tabela 18, essa profícua habilidade do escritor. Com os dados obtidos, encontramos também as *hapaxes*.

Mediante os dados extraídos dos *Corpora* verificamos se além da frequência maior de verbos no CCN, as *hapaxes* são significativas para afirmar se CN cultivava uma acentuada verborragia em relação a esta classe de palavra.

O processamento computacional dos *Corpora* para verificação dessas informações pode ser visualizado na figura a seguir:

Figura 24 - Comandos de extração de frequências, *hapaxes*, lista de **verbos**.



Fonte: Elaborada pela autora

Essa figura mostra que com a execução de poucos comandos podemos extrair de um *corpus* quantidade de *tokens*, quantidade e lista de *hapaxes*, e a relação das palavras mais frequentes do arquivo.

O programa de processamento por vezes não interpreta corretamente a codificação das palavras, geralmente palavras com acentuação, nesse caso executamos um comando com a função (*print*) que exibe os termos corretamente, como visto na figura 24.

Com a extração dos dados dos capítulos de *A Conquista*, já anotados, no caso somente do verbo SER (*tag* SR), fazemos o seguinte resumo: 363 verbos, 4 palavras raras (*hapaxes*), lista das *hapaxes* (*fora, foste, serei, sido*), lista das palavras mais frequentes (oito primeiras) com a quantidade de vezes que aparece no texto (*é, era, foi, eram, ser, são, sou, fosse*).

É importante frisar que estes arquivos não passaram por correção manual tampouco automática, isso significa dizer que as anotações podem conter erros.

Nesse caso, o processador enumera o verbo “haverei” com a *tag* SR-R, significando verbo ser no futuro. Por isso aparecem cinco *hapaxes* nos resultados.

Os capítulos de cada livro foram organizados em um só arquivo para o processamento por obra e para que os dados ficassem concentrados, minimizando assim a tarefa do processamento.

Os resultados obtidos a partir do processamento dos *Corpora* anotados revelaram-nos os seguintes dados:

Tabela 18 - Lista de ocorrências de **verbos** e *hapaxes* nos *Corpora*

TEXTO	VERBOS	HAPAX	% HAPAX
A Conquista	3.584	1.482	41,7
Turbilhão	3.637	1.541	42,3
O enterro	214	186	86,9
Mandovi	730	391	53,5
Firmo, o vaqueiro	322	225	68,8
TOTAL CCN	**8.487**		**58,6**
O cortiço	5.514	1.946	35,29
Amor de Salvação	2.887	1.273	44,0
TOTAL CORPUS CONTRASTE	**8.401**		**39,65**

Fonte: Elaborada pela autora

Para esse levantamento, além da categoria VB (verbo), acrescentamos nessa lista de frequências outras categorias de verbo (os auxiliares), que recebem etiquetas diferentes: SR (verbo ser), ET (verbo estar), TR (verbo ter), HV (verbo haver).

As possibilidades de ocorrências que podemos encontrar nesses *Corpora* são inúmeras. À medida que executamos os comandos vamos vislumbrando outras possibilidades de análises com uso tanto dos verbos, quanto de qualquer outra categoria gramatical.

Sobre as *hapaxes*, Biderman (1998, p. 176) afirma que: “A esmagadora maioria das palavras raras, *hapax legomena*, são substantivos. Eventualmente ocorrem alguns adjetivos e muito raramente um verbo”. Pudemos constatar que no

CCN as *hapaxes* são em maior quantidade os substantivos (nomes) que os verbos, mas com uma diferença pequena, uma relação de 2.725 substantivos para 2.441 verbos.

Como importa nesse trabalho a verificação de adjetivos, verbos e advérbios em –*mente*, vamos nos prender à analise dessas categorias.

Salientamos o fato de que a busca de ocorrência e a frequência de verbos nos *Corpora* levam em consideração as formas verbais flexionadas, e não somente os verbos no Infinitivo. Quando falamos em verbos, estamos nos referindo às formas também flexionadas, que contém etiquetas flexionais. Não nos detivemos a relacionar todos os verbos em sua forma infinitiva nos *Corpora*, especialmente no *Corpus* Coelho Netto, objeto maior de nossas análises. Para isso teríamos que submeter o CCN ao processo de lematização, que consiste em aplicar regras que identifiquem as mesmas formas gráficas correspondentes às diferentes flexões de um mesmo lema (CÚRCIO, 2013). O CHPTB não está lematizado (ALENCAR, 2009), e o Aelius carece de incremento no desenvolvimento de lematização (ALENCAR, 2013d).

Uma lista das dez *hapaxes* verbais presentes no CCN, iniciadas com a letra A, exemplificam melhor essa questão:

Exemplo 10:

Abafadas, abafado, abafando, abaixa, abala, abalada, abalar, abalou, abanando, abanava.

Se lematizássemos essas formas, teríamos a frequência dos verbos no Infinitivo, relacionando a ocorrência destes. Nesse caso, teríamos como resultado os verbos: *abafar, abaixar, abalar, abanar.*

Apresentando a mesma similaridade quanto à frequência de adjetivos nos *Corpora*, os verbos também apresentam poucas diferenças de frequência nas análises comparativas. O que observamos é que também as palavras raras são mais frequentes no CCN que no de Contraste: 58,9% e 39,6% respectivamente. Confirmamos assim a habilidade de CN em utilizar palavras raras nos seus textos, tanto adjetivos quanto verbos. Com esses dados, constatamos que nos *Corpora* as *hapaxes* de verbos são bastante elevadas em relação aos adjetivos, e não raramente como defende Biderman (1998).

Prova disso, é que do total de 53.080 *tokens* do CCN, 42.009 são *tokens* alfabéticos e 12.802 são *hapaxes*. Destas, 3.219 (3.825) são somente verbos. Temos então um percentual de 29,8% de formas verbais que aparecem somente uma vez no texto; os verbos, portanto, não podem ser considerados como formas raras nesse *Corpus*.

O que diferencia a frequência dos verbos em relação aos adjetivos é que aqueles sofrem muitas flexões, aumentando o quantitativo de ocorrências em relação a estes.

Fazendo uma análise mais minuciosa dos dados encontrados na exploração dos *Corpora*, quanto à relação dos verbos mais utilizados, fizemos um breve comparativo com os resultados da pesquisa de Freitas (2007), que estudando o estilo contista de Machado de Assis e com base na pesquisa de Maciel (1986) sobre o vocabulário de Érico Veríssimo, relacionou os dez verbos mais frequentes nos volumes de contos publicados por Machado de Assis, a partir das listas de altas frequências de cada grupo: *ser, ter, dizer, estar, perguntar, saber, haver, poder, ir, ver.*

Em nossa pesquisa, destacamos os sete verbos mais frequentes na obra de Coelho Netto, e em comparação com os dados de Freitas (2007), temos:

Quadro 14 - Comparação de verbos mais frequentes entre Machado de Assis e Coelho Netto

VERBO	**MACHADO DE ASSIS**	**COELHO NETTO**
Ser	1º	1º
Ter	2º	5º
Dizer	3º	6º
Estar	4º	2º
Haver	7º	3º
Ir	8º	4º
Ver	10º	7º

Fonte: Elaborado pela autora

Assim como na pesquisa de Freitas (2007) o verbo SER que aparece com mais frequência nos textos de Machado de Assis, são os mais utilizados também por Coelho Netto. A ordem de frequência de uso dos verbos por CN é a seguinte: *ser,*

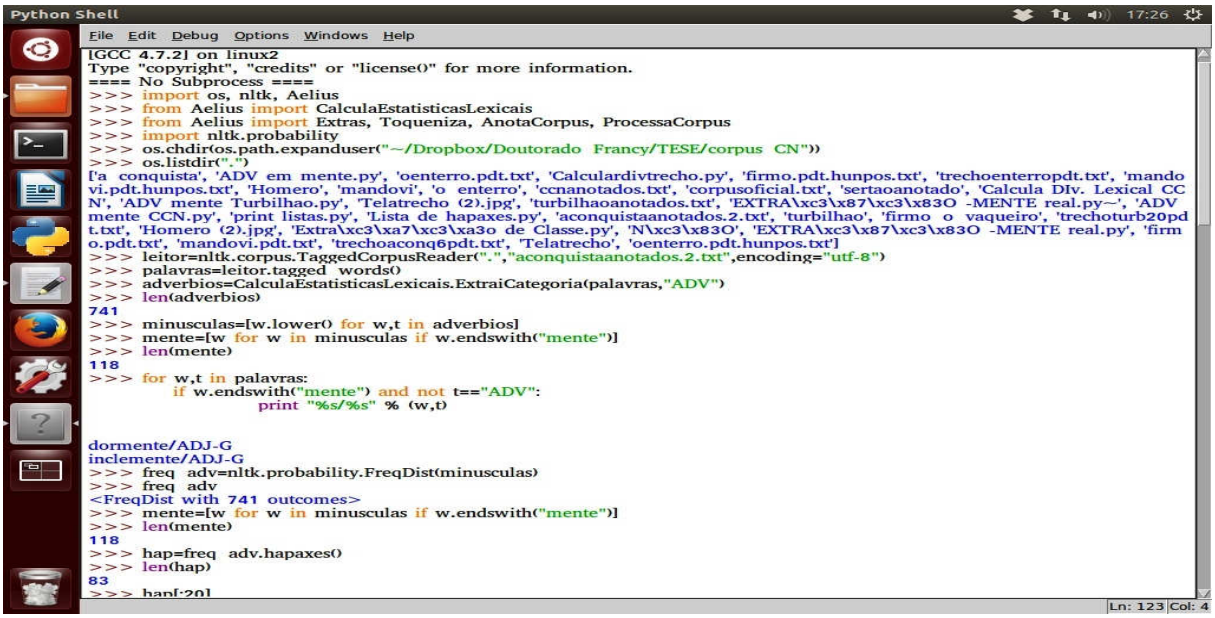
estar, haver, ir, ter, dizer, ver. O verbo ser é o que mais aparece, com 677 ocorrências. Causou-nos surpresa quando da realização dessas análises, ao perceber que os verbos ESTAR e HAVER não apareciam em nenhum momento no conto *O Enterro.* O verbo mais frequente nesse conto é o verbo IR.

4.2.4 Da frequência e diversidade de Advérbios em -mente:

Em uma busca rápida, grosso modo, no capítulo 01 do livro *A Conquista* de CN, Alencar (2011a), dos 186 advérbios do texto, observou uma incidência de 26% de advérbios em *–mente*. Quanto à diversidade lexical encontrada nessa busca, no que se refere à *hapaxes*, verificou um percentual de 93,8%, ou seja, dos 49 advérbios em *–mente* encontrados neste capítulo, 46 deles não foram repetidos, repetiram-se apenas três advérbios em *–mente*, o que é considerado por Alencar (2011a, p. 15) como uma "diversidade lexical altíssima". Esses dados nos instigaram a investigar a frequência desses advérbios nos textos de CN e compará-los com os textos de escritores do mesmo período, no caso Aluísio Azevedo e Camilo Castelo Branco, integrantes do *Corpus* de Contraste.

O processamento computacional dos *Corpora* para a extração de ocorrências desses advérbios, visualizamos na figura a seguir:

Figura 25 – Exemplo de comandos de extração de frequências, *hapaxes*, lista de advérbios em *–mente.*



Fonte: Elaborada pela autora.

Os dados obtidos estão compilados na tabela 19. Entendemos que ao especificar os dados por livro ou conto, teríamos uma dimensão mais detalhada de cada produção. Os capítulos dos livros (*A Conquista* e *Turbilhão*) foram organizados em arquivo único cada, para um melhor processamento dos dados.

Tabela 19 - Lista de ocorrências de **advérbios em *–mente*** e *hapaxes* no CCN

TEXTO	ADVÉRBIOS	PALAVRAS EM –*MENTE*	ADVÉRBIOS EM *-MENTE*	HAPAX	ADV. NÃO HAPAX
A Conquista	741	118	116	83	21
Turbilhão	673	132	131	77	22
O enterro	52	11	11	9	1
Mandovi	153	20	20	17	1
Firmo, o vaqueiro	55	2	2	2	0
TOTAL	**1.674**	**283**	**280**	**188**	**45**

Fonte: Elaborada pela autora

Do total de 1.674 advérbios encontrados no CCN, foram encontradas 283 palavras terminadas em *–mente*. O processador ao selecionar os advérbios terminados em *–mente*, acaba por resgatar todo o vocabulário do arquivo com terminação em –mente, para sanar esse problema executamos um comando que mostra somente os advérbios (ver figura 27). A tabela mostra em cada arquivo o quantitativo de palavras que não são advérbios em *-mente*.

Destacando os dados de *A Conquista*, temos 741 advérbios, destes 116 são advérbios terminados em *–mente*, mas no universo de palavras com essa terminação no arquivo o processador encontrou duas palavras que não são advérbios: *dormente e inclemente,* que são adjetivos no texto (ver dados na figura 27). Dos 116 advérbios, 83 são *hapaxes* (palavras que aparecem somente uma vez no texto) e 21 não são *hapaxes*, aparecem mais de uma vez, como mostrado na lista abaixo:

Tabela 20 - Lista de advérbios em *–mente* que não são *hapaxes* em *A Conquista.*

ADVÉRBIO	FREQUÊNCIA NO TEXTO
justamente	7
felizmente	4
constantemente	3
molemente	3
violentamente	3
decididamente	2
perfeitamente	2
desafogadamente	2
diariamente	2
disfarçadamente	2
finalmente	2
gravemente	2
imediatamente	2
inesperadamente	2
lentamente	2
maquinalmente	2
perfeitamente	2
preguiçosamente	2
principalmente	2
prodigamente	2
rapidamente	2

Fonte: Elaborada pela autora

Resumindo todas as informações processadas sobre o *CCN* e o uso destes advérbios, destacamos: das 283 palavras terminadas em *–mente*, somente 280 são advérbios, as outras categorias encontradas foram*: dormente e inclemente (ADJ) e mente (N)*. Destes 280 advérbios, 188 eram *hapaxes*, que significam 67% dos advérbios no *Corpus* Aquém do encontrado por Alencar (2011a), mas significativo quando somente 33% desses advérbios aparecem mais de uma vez. A pesquisa de VLCEK (2011, p. 91) sobre a presença desses advérbios no português e no francês, revela que em português “num universo de 90.000 [palavras], 36 eram advérbios em *-mente*, ou seja, aproximadamente 0,04%”. Num universo de 53.080 palavras no CCN, 280 são advérbios em *–mente*, aproximadamente 0,52%, indicando um percentual bem maior que o encontrado por Vlcek (2011). Apesar de

o tamanho destes *corpora* não serem equivalentes, o percentual encontrado no CCN é um indicador de riqueza vocabular em suas obras.

Compilamos os dados do *Corpus* de Contraste da mesma forma que o CCN, para que fosse possível a comparação entre estes. Essa comparação, no caso com os advérbios em *–mente,* está relacionada com o disposto na teoria relatada no item 2.6 sobre esses advérbios em textos literários, também em função da sugestiva riqueza vocabular de CN. A tabela 21 mostra os dados encontrados.

Tabela 21 - Lista de ocorrências de **advérbios em *–mente*** e *hapaxes* no *Corpus* de Contraste

TEXTO	ADVÉRBIOS	PALAVRAS EM *–MENTE*	ADVÉRBIOS EM *-MENTE*	HAPAX	ADV. NÃO HAPAX
O Cortiço	1.466	102	102	58	15
Amor de Salvação	536	48	48	36	5
TOTAL	**2.002**	**150**	**150**	**94**	**20**

Fonte: Elaborada pela autora

Diferente dos dados obtidos no CCN, os do *Corpus* de Contraste não continham nenhuma palavra terminada em *–mente* que não fosse advérbio. A quantidade de advérbios presente no *Corpus* de Contraste é superior ao do CCN, que tem 1.674. Apesar do *Corpus* de Contraste conter mais advérbios, a frequência de advérbios em *–mente* é bem inferior, mas esse dado não reflete na quantidade de *hapaxes*, que proporcionalmente apresenta um percentual muito aproximado ao do CCN: 62,6%. Em resumo, o *Corpus* de Contraste apresenta 2.002 advérbios, sendo que 150 são advérbios em *–mente*, destes 94 são advérbios raros, só aparecem uma vez no texto, e 20 aparecem repetidas vezes, relacionados abaixo:

Tabela 22 - Lista de advérbios em *–mente* que não são *hapaxes* no *Corpus* de Contraste.

ADVÉRBIO	FREQUÊNCIA NO TEXTO
justamente	7
perfeitamente	5
naturalmente	5
tristemente	4

escassamente	4
ruidosamente	3
principalmente	3
eternamente	3
totalmente	2
mormente	2
levemente	2
lentamente	2
igualmente	2
freneticamente	2
finalmente	2
febrilmente	2
exclusivamente	2
completamente	2
alegremente	2

Fonte: Elaborada pela autora

Esta lista contém 19 advérbios que não são *hapaxes*, diferentes do total da tabela 18, isso em função da palavra "*justamente*" ocorrer nos dois arquivos, e preferimos juntar as ocorrências.

Fazendo uma analogia ao que disse Broca (1958, p. 24) sobre a crítica em Portugal sobre as obras de Coelho Netto e Camilo Castelo Branco, poderíamos dizer que diante dos resultados destas análises, no Brasil, ninguém deveria se espantar com a riqueza lexical de um Coelho Netto, de um Aluísio Azevedo, nem de um Camilo, sobretudo porque produziram na mesma época. E Coelho Netto e Aluísio Azevedo ainda conviveram com as mesmas realidades políticas e sociais.

Nesse capítulo, com base nos dados obtidos pudemos confirmar de fato a riqueza vocabular de Coelho Netto, como também a de Aluísio e Camilo por meio do *Corpus* de Contraste. Essas informações podem de alguma forma lançar por terra a crítica voltada a CN, sobretudo quando diziam que ele vivia "eternamente debruçado sobre os dicionários" (NETTO, P., 1958, p. XCIX), ou pelo fato de ter uma produção vultosa, que seus "oponentes" modernistas não se preocuparam em ler (GURGEL, 2012), e como disse Broca (1958, p. 3), CN foi vítima de "juízos levianos e falhos".

5 CONCLUSÃO DA TESE

O enriquecimento cada vez mais do *Corpus* Coelho Netto ou de outros *corpora* de textos especializados contribuirá significativamente para beneficiar os pesquisadores da LCLC ou os que porventura se interessarem por essas áreas de pesquisa.

Inquietações à parte configuraram a intenção em realizar essa pesquisa, como os já mencionados na Introdução da tese. Enveredamos por uma área ainda com pouca divulgação no meio acadêmico. Apesar de eventos específicos como os que destacamos no texto produzirem ricas e diversificadas pesquisas, estas ainda se concentram significativamente na região Sudeste.

Quando iniciamos as pesquisas, observamos um crescimento progressivo de produções vinculadas à área, e, sobretudo, a disponibilização destes textos. Tem crescido a curiosidade e o interesse de se trabalhar com a Linguística de *Corpus*. Chegamos a essa conclusão diante da diversidade de trabalhos disponíveis na *WEB*, que as buscas constantes demonstraram em tempos diferentes de acesso. O aumento de grupos de pesquisas e o surgimento de novos *corpora* também revelam esse crescimento.

Ao mesmo tempo em que a *WEB* é uma porta de acesso a esses trabalhos, por outro lado muita coisa está ou desatualizada ou não dá mais acesso pelos *links* indicados: de etiquetadores, de *corpora,* de trabalhos acadêmicos, de artigos, de livros etc.

Quando buscamos dados sobre os *corpora* de textos literários do Português Brasileiro (*corpora* especializados) constatamos que os que existem são somente fragmentos de textos incorporados a um *Corpus* Geral, em que se concentram muitos e diversos tipos de textos. E os textos literários que fazem parte desses grandes *corpora* não são possíveis identificá-los, em sua maioria. Uma busca detalhada no Linguateca[42], que reúne a maior parte de *corpus* da LP, dos 46 corpus disponíveis não encontramos nenhum com a mesma especificidade do CCN, tampouco no seu Repositório[43], que agrega contribuições de pesquisadores do mundo inteiro.

[42] Corpora da LP - http://www.linguateca.pt/
[43] Repositório de recursos, publicações e programas- http://dinis.linguateca.pt/Repositorio/

Dos textos de Coelho Netto, nos *corpora* pesquisados, destacamos: o *Corpus* Araraquara que acresceu ao seu banco de dados o livro de contos *Cidade Maravilhosa* e o *Corpus* de Referência do Português Contemporâneo, que contém os principais textos desta tese: de Coelho Netto, *A Conquista;* de Aluísio Azevedo, *O Cortiço;* e Camilo Castelo Branco, *Amor de Perdição*. Ainda assim, a presença de textos de CN são minoria. Seria interessante que tivéssemos acesso livre ao *corpus Araraquara*, porque numa busca rápida na *WEB* não o encontramos. Quanto ao *Corpus* de Referência este está completamente disponível na *WEB* para consulta e uso.

As tarefas principais deste trabalho, praticamente todas feitas de forma manual, demandaram um tempo e cuidado maiores. As etapas subsequentes são feitas de forma automática, a partir de um dos objetivos concluídos, os outros automaticamente estarão concluídos também. Os processos que demandam mais precisão são os de compilação e de 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, mas não menos laboriosa tampouco simples, e sempre precisando de revisão humana.

É primordial um trabalho acurado nas etapas primeiras, para que os resultados do processamento computacional sejam efetivos. Pequenas amostras do que já foi revelado no processo de compilação, na anotação morfossintática e na correção manual, podem ser vistos nos Apêndices.

Quanto à avaliação da acurácia do etiquetador Aelius, os resultados da pesquisa mostraram que, ao anotar o CCN, o AeliusHunPos demonstrou um desempenho melhor que em outros *corpus* já testados: 97,9% em comparação com 96,3% detectado por Alencar (2013a).

Mediante a correção manual dos 10% dos *Corpora* deste trabalho, acredita-se em um aumento considerável no desempenho deste etiquetador, pois os aproximados 3% de erros de etiquetagem encontrados pós-correção manual não é tão expressivo ao ponto de modificar os resultados buscados. Principalmente porque as *tags* com mais ambiguidade foram: N (nome), NPR (nome próprio) e ADJ (adjetivos). Nesse caso, somente os resultados quanto aos adjetivos utilizados nos *Corpora* sofreriam modificações, pois boa parte dos erros cometidos na anotação de

ADJ deveria ser corretamente anotada como N (ver Apêndice K). Em suma, a cada 48 palavras anotadas, o Etiquetador cometeu um erro.

Os *Corpora* compilados e anotados nos possibilitam uma busca generosa de dados sobre seu léxico, mediante comandos específicos de processamento, como os mostrados nas diversas figuras expostas no trabalho, como forma de se visualizar todo o processamento realizado.

Importou-nos aqui cumprir com os objetivos previamente estabelecidos, validando ou não os questionamentos levantados na Introdução da tese.

Os resultados finais encontrados nas análises linguístico-computacionais nos permitiram validar a afirmação que Coelho Netto possuía um rico léxico, mas não exacerbado como os críticos, sobretudo os modernistas, atribuíram a ele. Os resultados mostraram também que os textos de Aluísio de Azevedo e Camilo Castelo Branco tinham riqueza lexical similar a Coelho Netto. Frisamos que a crítica da época não foi “por demais agressiva” (ADONIAS FILHO, 1964, p. 7) com estes escritores quanto foi com Coelho Netto. Tampouco, foram vítimas de uma campanha que “extrapolando os limites da negação de seus méritos literários, resvalaria para ofensas pessoais” (MORAES, 1976, p. 148).

Quanto à presença acentuada de verbos, adjetivos e advérbios em *–mente* nos textos de Coelho Netto, estes apresentam de fato riqueza e diversidade lexical. Quando se comparam essas ocorrências nos textos de Aluísio e Camilo as diferenças não são tão acentuadas. Se se comparássemos com escritores modernistas certamente as diferenças seriam marcantes.

Buscando na literatura dados sobre uso de advérbios em *-mente* e de que eles são mais ocorrentes no português arcaico, constatamos nos *Corpora* a proeminência destes advérbios nos textos dos escritores analisados, considerando-se os movimentos literários que vivenciaram, o que de alguma forma contesta o que pesquisas sobre a incidência destes advérbios relatam.

As extensas pesquisas na *WEB,* no período de realização deste trabalho, revelaram que se lê um Aluísio Azevedo, mas não se lê um Coelho Netto, dois maranhenses de produções da mesma época. Sobretudo porque as obras deste não estão disponíveis na *WEB*; são poucas e sempre as mesmas, e mais ainda em função da campanha crítica depreciativa feita à CN.

Entendemos que ao concluirmos este trabalho contribuiremos para que algumas “impressões” sobre Coelho Netto mudem, pois o fato de disponibilizarmos

seus textos para acesso livre, e ainda preparado/estruturado para análises linguísticas se configura em uma ação producente no âmbito da Linguística e também da Literatura.

Destacamos o âmbito da Literatura, principalmente por concordar com o que diz Santos (2008) ao falar sobre a convergência das pesquisas entre Linguística e Literatura, ao declarar estar convencida do erro fulcral que é a separação entre literatura e linguística, pois ela entende que devia ser (também) ao se estudar a literatura que compreendíamos a língua.

Esta tese, quando ainda em andamento, foi objeto de nossa participação em dois eventos internacionais: IV SIMELP - Simpósio Mundial de estudos de Língua Portuguesa (apresentação da tese em andamento) e o STIL 2013 - *9th Brazilian Symposium in Information and Human Language Technology* (*paper* publicado e apresentação de pôster da tese em andamento). Este último, o *paper,* foi submetido à avaliação de três revisores que fizeram algumas observações a respeito do trabalho: que o *short-paper* apresentava uma boa motivação e bom potencial de atingir resultados interessantes; que é um *corpus* necessário para a pesquisa literária; que o tema do trabalho é relevante e original, por se tratar de textos históricos que não são anotados morfossintaticamente, e que esta anotação irá permitir uma análise mais aprofundada desses texto. Isso foi motivador para o desenvolvimento do trabalho.

Além destes incentivos, um revisor destacou a possibilidade de aumentar o *corpus* (incluindo outros tipos de textos), isso para cobrir um intervalo de tempo, gêneros etc, no *corpus,* resolvendo a restrição do tamanho do *corpus.* Na correção do *paper* justificamos o tamanho do *corpus* da mesma forma relatada no item 3.1.

Ao realizarmos trabalhos nessa área, percebemos que a LCLC engloba vários campos e áreas, evidenciando sua riqueza e alcance, gerando uma diversidade que é fonte de enriquecimento dos estudos linguísticos da língua portuguesa. Enriquecimento este que nos permite ampliar nossos horizontes, proporcionando mudanças em uma situação que Phillips (1989) apontava, ao dizer que a linguística tradicionalmente está restrita à investigação de trechos de linguagem que podem ser acomodados confortavelmente em um quadro negro comum.

Apoiando o discurso de Santos (2008, p. 60), enfatizamos que para queremos compreender a língua não podemos fixar-nos ou fiar-nos em áreas

estanques, é preciso que estejamos abertos ao diálogo e à compreensão de métodos e conceitos aparentemente de ‘outras’ áreas. Área entendida como algo dinâmico que está sempre em evolução.

Porquanto, das contribuições linguísticas resultantes deste trabalho estão incluídas as análises provenientes tanto das anotações morfossintáticas realizadas pelo etiquetador, quanto do uso das bibliotecas do programa NLTK que *a posteriori* serão otimizadas com os resultados obtidos das análises, que ao processar os comandos busca informações linguísticas já estruturadas em seus pacotes. Além do que a criação e compilação dos *Corpora* deste trabalho servem como produtos científico-tecnológicos para fins linguísticos, corroborando as pesquisas dos linguistas de *corpus*. Santos (2008) reforça esse intento, ao dizer que a ação mais simples que possamos imaginar na LCORP, que é a de contar palavras ou identificar a pontuação, pressupõe uma teoria linguística.

Fazemos alusão a Berber Sardinha (2005a) e Zyngier (2004) ao dizer que os “desafios específicos” impostos pelas tecnologias na área de língua portuguesa constituem-se na proposição de novas formas de se analisar um fenômeno linguístico, obtendo-se resultados que certamente levarão à formulação de novas perguntas às questões estudadas, possivelmente contradizendo leituras tradicionais de um determinado texto e colocar novamente em julgamento sua qualidade literária.

Nas décadas por vir, os dados aumentarão exponencialmente, em todos os domínios e línguas. O futuro estará seguramente na informatização de todos esses dados, em sua sistematização e na diversificação dos acessos: a lexicografia informatizada já é resultante desse processo.

Biderman, em 1998, mencionava o quão é difícil dar conta de todo o acervo de palavras da língua portuguesa para um tratamento lexicoestatístico e computacional nos moldes que desejamos. Hoje, não damos conta do todo, mas avançamos sobremaneira no sentido de amostralmente compilarmos *corpus* que convergirão com outros já criados e se constituirão em grandes *Corpora* do Português Brasileiro.

Entendemos também como Martin (2003) que as perspectivas da Linguística Computacional são múltiplas, beneficiando tanto os métodos quanto os fundamentos teóricos, levando a crer que a linguística do futuro se encontrará profundamente renovada. A Linguística Descritiva se beneficiará notadamente com

as perspectivas de análise que propicia um *corpus* compilado, anotado e aberto. Os recentes encontros, seminários, simpósios e congressos como o LiPral, ELC, STILL, PROPOR, em suas sessões técnicas, com pôsteres e exposições orais, demonstram isso.

Com este trabalho não tivemos a pretensão de realizar análises mais profundas com os dados obtidos, pois acreditamos se tratar de objeto de futuras investigações mais minuciosas dos *Corpora* disponibilizados no formato atual. Estas tarefas não couberam aqui, porque isso demanda tempo e recurso humano. Para exemplificar outras ações de anotação morfossintática e análises linguístico-computacionais, a exploração dos *Corpora* pode se voltar para busca da diversidade lexical de todas as classes de palavras; da relação de concordância das palavras buscando dados mais concretos do ponto de vista tanto sintático quanto semântico; do esmiuçamento de ocorrências de tempos verbais, e outras tantas análises ocorrentes na Linguística em busca da compreensão da língua. Os dados obtidos para a avaliação da acurácia do Etiquetador AeliusHunPos neste trabalho possibilitará a melhoria de seu desempenho, como já mencionado.

Também não pretendíamos neste trabalho defender ou condenar Coelho Netto. Tampouco condenar ou “absolver” os críticos modernistas. Mas, sobretudo, revelar dados que estudos literários mais especializados podem realizar tomando como base os *corpora* compilados, anotados e analisados.

Enfim, ao concluirmos esta tese, esperamos ter contribuído com a Linguística de *Corpus*, no sentido de investigar os gêneros por meio de um grande número de textos, os quais formam um *corpus* eletrônico. E também, como enfatizado por Alencar (2013c) “preencher importante lacuna na ‘paisagem de *corpora*’ do português, onde os *corpora* anotados disponíveis, os quais permitem um processamento computacional mais sofisticado, ainda são insuficientes”.

REFERÊNCIAS

ABREU, D. Iuri. **Normalização lexical em traduções de Dom Casmurro, de Machado de Assis**: um estudo baseado em *corpus*. Dissertação (Mestrado em Estudos da Tradução). Universidade Federal de Santa Catarina, Florianópolis, 2007. Disponível em: <https://repositorio.ufsc.br/handle/123456789/89771?show=full>. Acesso em: 29 set. 2013.

ADONIAS FILHO. Apresentação da Exposição Coelho Neto. *In:* BIBLIOTECA NACIONAL DO RJ. **Exposição comemorativa do centenário de nascimento de Coelho Neto**. Rio de Janeiro: Gráfica Editora Olímpica, 1964. Disponível em: <http://objdigital.bn.br/acervo_digital/div_iconografia/icon1285844.pdf>. Acesso em: 21 abr. 2014.

AFONSO, Susana *et. al*. Floresta Sintá(c)tica: um “treebank” para o Português. do *In:* XVII Encontro da Associação Portuguesa de Linguística, Lisboa: APL, 2002, pp.533-45. **Actas...** Disponível em: <http://visl.sdu.dk/~eckhard/pdf/AfonsoetalAPL2001.ps.pdf>. Acesso em: 02 out. 2013.

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

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

_____. de. **Aelius Brazilian Portuguese POS-Tagger and Corpus Annotation Tool**, versão 0.9.7. Fortaleza: [s.n.], 2013b. Disponível em: <http://aelius.sourceforge.net/>. Acesso em: 25 fev. 2013.

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

_____. **Aelius User's Manual**. UFC, 2013d. Disponível em: <http://aelius.sourceforge.net/manual.html>. Acesso em: 30 abr. 2014.

_____. **Fundamentos de Python para a Linguística de Corpus**. Manuscrito. Fortaleza: Universidade Federal do Ceará, 2011a.

_____. Novos recursos do Aelius para o processamento computacional raso do português. *In:* LAPORTE, Éric (org). **Dialogar é preciso**: linguística para o processamento de línguas. Vitória-ES: PPGEL/UFES, 2013a.

_____. **Parecer sobre tese de doutorado em andamento**: compilação, anotação e análises linguístico-computacionais de um corpus-amostra de textos literários dos

séc. XIX e XX: o Corpus Coelho Netto. Manuscrito. Fortaleza: Universidade Federal do Ceará, 2013c.

_____. Técnicas em software livre para exploração de *corpora* do português livremente disponíveis na WWW. UFC. **Veredas on-line** – linguística de *corpus* e computacional – 2/2009, p. 134-150. Disponível em: <http://www.ufjf.br/revistaveredas/files/2009/11/ARTIGO-Leonel-Figueiredo-de-Alencar.pdf>. Acesso em: 20 jan. 2013.

_____. Utilização de informações lexicais extraídas automaticamente de *corpora* na análise sintática computacional do português. **Rev. Est. Ling.,** Belo Horizonte, v. 19, n. 1, p. 7-85, jan./jun. 2011b. Disponível em: < http://www.periodicos.letras.ufmg.br/index.php/relin/article/view/2553/2505>. Acesso em: 20 jan. 2013.

ALUÍSIO, Sandra M. *Corpus* históricos, recursos léxicos e ferramentas para a tarefa de criação de dicionários. *In:* I Escola Brasileira de Linguística Computacional. **Apresentações...** Faculdade de Filosofia, Letras e Ciências Humanas: USP, 2007. Disponível em: <http://www.nilc.icmc.usp.br/nilc/projects/hpc/ebralc_2007/ebralc_2007_alusio.pdf>. Acesso em 03 out. 2013.

ALUÍSIO, Sandra M.; ALMEIDA, Gladis M. B. O que é e como se constrói um *corpus*? Lições aprendidas na compilação de vários *corpora* para pesquisa linguística. **Calidoscópio (UNISINOS**). Vol. 4, n. 3, p. 155-177, set/dez, 2006. Disponível em: <http://www.unisinos.br/publicacoes_cientificas/images/stories/pdfs_calidoscopio/vol4n3/art04_aluisio.pdf>. Acesso em: 20 jan. 2013.

ALVES, Daniel, TAGNIN, Stella E. O. Pela visibilidade da Linguística em trabalhos acadêmicos. *In:* X ENCONTRO DE LINGUÍSTICA DE *CORPUS*: aspectos metodológicos dos estudos de *corpora*. **Anais...** Belo Horizonte: Faculdade de Letras da UFMG, 2012. P. 13-33.

ARRUFAT, M. J. Gallego. **Cuestiones y polémicas en la investigación sobre medios de enseñanza.** Granada: Facultad de Ciencias de la Educación**,** 1997. Disponível em: <http://www.lmi.ub.es/te/any97/gallego_force/>. Acesso em: 19 set. 2013.

ASSIS, Machado. Crônica de 11 de agosto de 1895. **Gazeta de Notícias**, 1897. Disponível em: < http://www.cronicas.uerj.br/home/cronicas/machado/rio_de_janeiro/ano1895/11ago1895.html>. Acesso em 23 abr. 2014.

_____. Crônica de 14 de fevereiro de 1897. **Gazeta de Notícias**, 1897. Disponível em: <http://www.cronicas.uerj.br/home/cronicas/machado/rio_de_janeiro/ano1897/14fev1897.html>. Acesso em 23 abr. 2014.

ATKINS, Sue; CLEAR, Jeremy; OSTLER, Nicholas. Corpus design criteria. **Literary and Linguistic Computing**, 7: 1-16, 1991. Disponível em: <www.natcorp.ox.ac.uk/archive/vault/tgaw02.pdf>. Acesso em 23 abr. 2014.

BACELAR DO NASCIMENTO, Mª Fernanda. **O lugar do *corpus* na investigação linguística**. Centro de Linguística da Universidade de Lisboa, 2002. Disponível em:

<http://www.clul.ul.pt/equipa/fbacelar/apl_2002_nascimento.pdf>. Acesso em: 10 abr. 2014.

BDTD: Biblioteca Digital Brasileira de Teses e Dissertações. Disponível em: http://bdtd.ibict.br/. Acesso em: 23 abr. 2014.

BENNETT, Gena R. **Using corpora in the language learning classroom**: corpus linguistics for teachers. Michigan: ELT, 2010. Disponível em: <http://www.nhlrc.ucla.edu/media/files/Using-corpora-in-the-language-learning-classroom--Corpus-linguistics-for-teachers-my-atc.pdf>. Acesso em: 02 out. 2013.

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

_____. **Linguística de *Corpus***. São Paulo, Manole, 2004. Disponível em: <http://books.google.com.br/books?id=i8uJXgeok48C&printsec=frontcover&dq=lingu%C3%ADstica+de+corpus&hl=pt-BR&sa=X&ei=AML-UKOUKJK29gTY5ICACg&ved=0CDIQ6AEwAA#v=onepage&q=lingu%C3%ADstica%20de%20corpus&f=false>. Acesso em: 20 jan. 2013.

_____. Linguística de *Corpus*: histórico e problemática. **DELTA [online]**. 2000, 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.

_____. **Pesquisa em Linguística de *Corpus* com WordSmith Tools**. Campinas: Mercado de Letras, 2009. Disponível em: <sis.posugf.com.br/AreaProfessor/Materiais/Arquivos_1/13879.pdf>. Acesso em 02 out. 2013.

_____. Trazendo a língua portuguesa para o computador. In: BERBER SARDINHA, T. (org.). **A língua portuguesa no computador**. Campinas, SP: Mercado de Letras, 2005b. p.269-295.

_____. Ver a língua portuguesa no computador. In: BERBER SARDINHA, T. (org.). **A língua portuguesa no computador**. Campinas, SP: Mercado de Letras, 2005a. p.7-32.

BERBER SARDINHA, Tony; ALMEIDA, Gladis M. de B. A linguística de *corpus* no Brasil. *In*: TAGNIN, Stella E. O; VALE, Oto A. **Avanços da linguística de *corpus* no Brasil**. São Paulo: Humanitas, 2008. p. 17-40.

BEZERRA, Eliezer. **Coelho Netto e a onda modernista**. São Paulo: Ítalo-Latino-Americana PALMA, 1982.

BIBLIOTECA NACIONAL DO RJ. **Exposição comemorativa do centenário de nascimento de Coelho Neto**. Rio de Janeiro: Gráfica Editora Olímpica, 1964. Disponível em: <http://objdigital.bn.br/acervo_digital/div_iconografia/icon1285844.pdf>. Acesso em: 21 abr. 2014.

BICK, Eckhard. Gramática Constritiva na Análise Automática de Sintaxe Portuguesa. *In*: BERBER SARDINHA, Tony (org.). **A Língua Portuguesa no computador**. Campinas, SP: Mercado de Letras, São Paulo: 2005. Fapesp. p. 91-112.

BIDERMAN, M. Teresa C. A face quantitativa da linguagem: um dicionário de frequências do português. **Alfa** (São Paulo), v.42 (n. esp.), p.161-181, 1998. Disponível em: <seer.fclar.unesp.br/alfa/article/download/4049/3713>. Acesso em: 23 abr. 2014.

_____. **Teoria Linguística**: teoria lexical e Linguística Computacional. São Paulo: Martins Fontes, 2001.

BIRD, Steven *et al*. Multidisciplinary Instruction with the Natural Language Toolkit. **Proceedings of the Third Workshop on Issues in Teaching Computational Linguistics (TeachCL-08)**, pages 62–70, Columbus, Ohio, USA, June 2008. Association for Computational Linguistics. Disponível em: <www.aclweb.org/anthology/W/W08/W08-0208.pdf>. Acesso em: 30 abr. 2014.

BIRD, Steven; KLEIN, Ewan; LOPER, Edward. **Natural language processing with Python**: analyzing text with the Natural Language Toolkit. Sebastopol, CA: O'Reilly, 2009.

BOAVENTURA, M. Eugênia. Macunaíma: ficção divertida e riqueza vocabular. **Revista do Instituto Humanitas Unisinos**. São Leopoldo, RS, ago. 2008. Entrevista concedida a André Dick e Márcia Junges. Disponível em: <http://www.ihuonline.unisinos.br/index.php?option=com_content&view=article&id=2022&secao=268>. Acesso em: 23 abr. 2014.

BORBA, Francisco da S. A informação gramatical nos dicionários. **Revista Alfa**, São Paulo, 51 (1): 137-149, 2007. Disponível em: <http://pt.scribd.com/doc/61494454/A-INFORMACAO-GRAMATICAL-NOS-DICIONARIOS-BORBA> Acesso em: 20 jan. 2013.

_____. **Uma gramática de valências para o português**. São Paulo: Ática, 1996.

BORGES, Luiz E. **Python para desenvolvedores**. Rio de Janeiro: Edição do Autor, 2010. Disponível em: <http://ark4n.files.wordpress.com/2010/01/python_para_desenvolvedores_2ed.pdf>. Acesso em: 20 jan. 2013.

BRANDÃO, Jacyntho J. L. **Presença maranhense na literatura nacional**. São Luis: UFMA/SIOGE, 1979.

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

CAMPOS, Angelita H.; ANDRADE, Elias A. de. Características ortográficas da língua portuguesa: séculos XVIII ao XX. **Cadernos do CNLF**, Vol. XV, Nº 5, t. 2. Rio de Janeiro: CiFEFiL, 2011. Disponível em: <http://www.filologia.org.br/xv_cnlf/tomo_2/121.pdf>. Acesso em: 31 jan. 2013.

CAMPOS, Humberto de. **O Sr. Coelho Neto e o seu estilo**. Crítica (Primeira Série). 2.ed. Rio de Janeiro, São Paulo, Porto Alegre: W.M. Jackson Inc. Editores, 1945, pp. 285-300.

CANDIDO JR, A.; ALUÍSIO, S. M. Um ambiente computacional para o processamento de *corpus* do Português Histórico. *In*: **VI Best MSc Dissertation/PhD Thesis Contest** (CTDIA 2008), 2008, Salvador. Proceedings of CTDIA (2008). Hidelberg: Springer, 2008. v. 1. p. 1-10. Disponível em:

<http://www.nilc.icmc.usp.br/nilc/projects/hpc/ctdia_2008/ctdia_2008.pdf>. Acesso em 13 set. 2013.

CANDIDO, Antonio. **Iniciação à literatura brasileira**: resumo para principiantes. 3.ed. São Paulo: Humanitas/ FFLCH/USP, 1999. Disponível em: <http://www.afoiceeomartelo.com.br/posfsa/Autores/Candido,%20Antonio/Inicia%C3%A7%C3%A3o%20-%20Literatura%20Brasileira%20-%20Antonio%20Candido.pdf>. Acesso em: 23 abr. 2014.

CARVALHO, Cid I. da C.; VASCONCELOS, Davis, M.; ALENCAR, Leonel F. Superando o estado da arte na etiquetagem morfossintática por meio de regras de pós-etiquetagem. *In:* X Encontro de Linguística de Corpus: aspectos metodológicos dos estudos de *corpora*. **Anais...** Belo Horizonte: Faculdade de Letras da UFMG, 2012. p. 122-134.

CÚRCIO, Verônica R. **Palavras de Rosa**: análise estilométrica da obra de João Guimarães Rosa. 2013. 158 f. Tese (Doutorado do Programa de Pós-Graduação em Literatura da Universidade Federal de Santa Catarina) – Universidade Federal de Santa Catarina: Florianópolis: 2013.

DI FELIPPO, Ariani; DIAS-DA-SILVA, Bento C. O processamento automático de línguas naturais enquanto engenharia do conhecimento linguístico. Unisinos. **Calidoscópio**, Vol. 7, n. 3, p. 183-191, set/dez 2009. Disponível em: <http://revistas.unisinos.br/index.php/calidoscopio/article/view/4871>. Acesso em 28 jan. 2014.

_____. Os adjetivos valenciais do português e sua representação linguístico-computacional. **Revista do GEL 02 (2005)**, pp. 55-81. Disponível em: <www.letras.ufscar.br/pdf/Revista_do_GEL.pdf>. Acesso em: 15 abr. 2014.

DIAS-DA-SILVA, Bento C. O estudo linguístico-computacional da linguagem. **Letras de Hoje**, Porto Alegre. v.41(2): 103-138, junho 2006. Disponível em: <http://revistaseletronicas.pucrs.br/ojs/index.php/fale/article/viewFile/597/428>. Acesso em 28 jan. 2014.

DOMINGUES, Miriam L. **Abordagem para o desenvolvimento de um etiquetador de alta acurácia para o Português do Brasil**. 2011. 140 f. Tese (Doutorado do Programa de Pós-Graduação em Engenharia Elétrica.) – Universidade Federal do Pará, Instituto de Tecnologia: Belém, 2011.

DOMINGUES, Miriam L.; FAVERO, Eloi L.; MEDEIROS, Ivo P. de. O desenvolvimento de um etiquetador morfossintático com alta acurácia para o português. *In:* TAGNIN, S. E. O.; VALE, O. A. (Eds**.). Avanços da Linguística de *Corpus* no Brasil**. São Paulo: Humanitas, 2008. p. 267-286.

DOMÍNIO PÚBLICO: Biblioteca Digital do Governo do Brasil. Disponível em: http://www.dominiopublico.gov.br/pesquisa/PesquisaPeriodicoForm.jsp. Acesso em 28 jan. 2014.

DOS SANTOS, C. N.; MILIDIÚ, R.; RENTERIA, R. Portuguese part-of-speech tagging using entropy guided transformation learning. *In*: Teixeira, A. et. Al (Eds). **Computational processing of the portuguese language**, volume 5190. SpringerBerlin / Heidelberg, PROPOR, 2008, p.143–152.

DUARTE, Júlio C. **O Algoritmo Boosting at Start e suas aplicações**. Tese (Doutorado do Programa de Pós–graduação em Informática) – PUC-Rio. Rio de

Janeiro: PUC-Rio, 2009. Disponível em: <http://www2.dbd.puc-rio.br/pergamum/tesesabertas/0510956_09_cap_07.pdf>. Acesso em 03 out. 2013.

DUBOIS, J. et al. **Dicionário de linguística**. 16ed. São Paulo: Cultrix, 2011.

FRANKENBERG-GARCIA, Ana. Prefácio. *In*: SHEPHERD, Tania M. G.; BERBER SARDINHA, Tony; PINTO, Marcia V. (org). **Caminhos da linguística de *corpus***. Campinas, SP: Mercados das Letras, 2012. p. 11-14.

FREITAS, Claudia; ROCHA, Paulo; BICK, Eckhard. Um mundo novo na Floresta Sintá(c)tica – o treebank do Português. **Calidoscópio**, Vol. 6, n. 3, p. 142-148, set/dez 2008. Disponível em: <http://www.linguateca.pt/documentos/FreitasetAl2008Calidoscopio.pdf>. Acesso em: 07 jan. 2013.

FREITAS, Deise J. T. **A composição do estilo do contista Machado de Assis**. 2007. 211f. Tese (doutorado). Programa de Pós-Graduação em Literatura – Universidade Federal de Santa Catarina, Florianópolis, 2007.

FRIZON, M. Morte e Vida Severina e o Super-Regionalismo. **Revista Terceira Margem**. Programa de pós-graduação em ciência da literatura. Ano IX, nº 12, janeiro-junho/2005. Disponível em: <http://www.letras.ufrj.br/ciencialit/terceiramargemonline/numero12/x.html>. Acesso em: 20 jan. 2013. Não paginado.

GALVES, Charlotte M. C. **Novos rumos para a pesquisa linguística no Brasil**. 2003. Disponível em: <http://www.ime.usp.br/~tycho/participants/c_galves/uerj_rev.html>. Acesso em: 16 set. 2013. Não paginado.

GALVES, Charlotte M. C.; BRITTO, Helena. **A Construção do *Corpus* Anotado do Português Histórico Tycho Brahe**: o sistema de anotação morfológica. Campinas: UNICAMP, 1999. Disponível em: <http://www.tycho.iel.unicamp.br/~tycho/old/papers/cgalves_britto.pdf>. Acesso em: 29 set. 2013.

GALVES, Charlotte M. C.; FARIA, Pablo. **Tycho Brahe Parsed *Corpus* of Historical Portuguese** (CHPTB). Campinas: Universidade Estadual de Campinas, 2012. Disponível em:<http://www.tycho.iel.unicamp.br/~tycho/corpus/>. Acesso em: 16 set. 2013. Não paginado.

GONÇALVES, Patricia N.; BRANCO, António. **Buscador online do CINTIL-Treebank.** In: XXV Encontro Nacional da Associação Portuguesa de Linguística, Porto, APL, 2010, pp. 465-473. Disponível em: <http://www.apl.org.pt/docs/25-textos-seleccionados/33-Patricia%20Nunes.pdf>. Acesso em: 27 set. 2013.

GURGEL, Rodrigo. Perseguido, mas brilhante. **Rascunho,** set. 2012. Disponível em: <http://rascunho.gazetadopovo.com.br/perseguido-mas-brilhante/>. Acesso em: 23 abr. 2014.

HOPCROFT, J. E.; MOTWANI, R.; ULLMAN, J. D. I**ntrodução à Teoria dos Autômatos, Linguagens e Computação.** Rio de Janeiro: Campus, 2002.

HUNSTON, Susan. ***Corpora* in Applied Linguistics**. London: Cambridge University Press, 2002. Disponível em: <http://books.google.com.br/books?id=BUIdMNgugGEC&printsec=frontcover&hl=pt-

BR&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false>. Acesso em: 29 set. 2013.

JOÃO DO RIO. **O momento literário**. Rio de Janeiro: Garnier, 1904. Disponível em: <http://www.dominiopublico.gov.br/pesquisa/DetalheObraForm.do?select_action=&co_obra=2144>. Acesso em: 22 abr. 2014.

JORGE, Henrique. **A academia do fardão e da confusão**: a academia brasileira de letras e seus "imortais" mortais. São Paulo: Geração Editorial, 1999. Disponível em: <http://books.google.com.br/books?id=64jKNYmXB8UC&printsec=frontcover&hl=es&source=gbs_ge_summary_r&cad=0#v=onepage&q=Coelho%20neto&f=false>. Acesso em: 02 out. 2013.

JURAFSKY, Daniel; MARTIN, James H. **Speech and Language Processing**. 2.ed. New Jersey: Pearson Prentice Hall, 2009.

KENNEDY, Graeme. **An introduction to corpus linguistics**. Nova Yoik, Longman, 1998. Disponível em: <acl.ldc.upenn.edu/J/J99/J99-2010.pdf>. Acesso em: 23 abr. 2014.

LAKATOS, Eva M.; MARCONI, Marina. de A. **Metodologia do trabalho científico**. 3. ed. São Paulo: Atlas, 1991.

LEECH, Geoffrey. Adding Linguistic Annotation. *In:* Wynne, Martin (org.). **Developing Linguistic *Corpora*:** a Guide to Good Practice**.** Oxford: Oxbow Books, 2004. Disponível em: <http://www.ahds.ac.uk/creating/guides/linguistic-*corpora*/chapter2.htm>. Acesso em: 29 set. 2013. Não paginado.

LIMA, Herman. Coelho Netto: as duas faces do espelho. *In*: **Coelho Netto: obra seleta**: romances. V. 01. Biblioteca Luso-brasileira. Rio de Janeiro: José Aguilar, 1958. p. XI-LXXXII.

LINGUATECA, 2012. Disponível em: <http://www.linguateca.pt/> Acesso em: 28 jan. 2013.

LOPES, Reginaldo J. A língua do Brasil, palavra por palavra. **Revista Unesp Ciência** (*online*), setembro de 2009, ano 1, n. 1, p. 26-29. Disponível em: <http://www.unesp.br/aci_ses/revista_unespciencia/acervo/01/novo-dicionario>. Acesso em: 31 jan. 2013.

MACHADO, José B. Estudo léxico-informático de 10 canções de Camões. *In:* **Fim do Milénio – VII e VIII**. Fóruns Camonianos, ed. Manuela de Azevedo. Lisboa: Edições Colibri, 2001. p.145 - 152. Disponível em: <alfarrabio.di.uminho.pt/vercial/zips/machad03.rtf>. Acesso em: 12 abr. 2014.

MACIEL, Carlos A. A. **Richesse et evolution du vocabulaire d'Érico Veríssimo** (1905-1975 – Porto Alegre, Brésil). Paris: Champion; Genéve: Slaktine, 1986.

MARTIN, Robert. **Para entender a Linguística**: epistemologia elementar de uma disciplina. Trad. Marcos Bagno. São Paulo: Parábola, 2003.

MATTOS E SILVA, Rosa V. **Sobre o programa para a história da Língua Portuguesa (PROHPOR) e sua inserção no projeto nacional para a história do português brasileiro (PHPB):** dados atualizados. UFBA, 2007. Disponível em: <http://www.prohpor.ufba.br/sobreprohpor.pdf>. Acesso em: 02 out. 2013.

McENERY, T.; WILSON, A. ***Corpus* linguistics**: an introduction. 2.ed. Edinburgh University Press Ltd, 2001. Disponível em:

<http://books.google.hu/books?id=nwmgdvN_akAC&printsec=frontcover&hl=hu&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false>. Acesso em: 20 jan. 2013.

MEIRELES, Mário; FERREIRA, Arnaldo de J.; VIEIRA FILHO, Domingos. **Antologia da Academia Maranhense de Letras**. São Luís: Gráfica Taveira, 1958.

MENDES, Rafael F. C; IGNÁCIO, Ewerton de F. Entre dois modus vivendi: arcaísmo e modernidade em Turbilhão, de Coelho Netto. *In*: VIII Seminário de Iniciação Científica e V Jornada de Pesquisa e Pós-Graduação. **Anais...** Universidade Estadual de Goiás, 2010. Disponível em: <http://www.prp.ueg.br/sic2010/apresentacao/trabalhos/pdf/linguistica/seminario/entre_dois_modus_vivendi.pdf>. Acesso em: 15 jan. 2013.

MESSINA, Graciela. Investigación en o investigación acerca de la formación docente: un estado del arte en los noventa. **Revista Iberoamericana de Educación**, n. 19, Formación Docente, Enero – Abril, 1999. Disponível em: <http://www.rieoei.org/oeivirt/rie19a04.htm>. Acesso em: 20 jul. 2013.

MOISÉS, Carlos F. Camilo Castelo Branco: a vida imita a arte, a arte imita a vida. *In:* BRANCO, Camilo C. **Amor de Perdição**. São Paulo, 1997.

MORAES PINTO, Deise C. de. **Gramaticalização e ordenação nos advérbios qualitativos e modalizadores em *-mente***. 2008, 145 p. Tese (Doutorado em Linguística) UFRJ. Rio de Janeiro, 2008. Disponível em: <www.letras.ufrj.br/poslinguistica/wp-content/.../deise-cristina-pinto.pdf>. Acesso em: 23 abr. 2014.

MORAES, Jomar. **Apontamentos de literatura maranhense**. São Luís: Edições SIOGE, 1976.

MOREIRA FILHO, José L. **Desenvolvimento de um software para preparação de aulas de inglês com *corpora*.** Dissertação (Mestrado de Linguística aplicada e estudos da linguagem). São Paulo: Pontifícia Universidade Católica de São Paulo (PUC), 2007. Disponível em: <http://www4.pucsp.br/pos/lael/lael-inf/teses/jose_lopes_moreira_filho.pdf>. Acesso em: 29 set. 2013.

NAMIUTI, Cristiane. O corpus anotado do português histórico: um avanço para as pesquisas em linguística histórica do português. **Revista Virtual de Estudos da Linguagem – ReVEL**. V. 2, n. 3, agosto de 2004. Disponível em: <http://www.revel.inf.br/files/artigos/revel_3_o_corpus_anotado_do_portugues_historico.pdf>. Acesso em: 03 out. 2013.

NERES, J. *Hem-hem*: uma marca polissêmica do falar maranhense. **Língua Portuguesa Conhecimento Prático**. São Paulo, 2010, nº 25. P. 30-33. Disponível em: <http://conhecimentopratico.uol.com.br/linguaportuguesa/gramatica-ortografia/25/artigo185982-1.asp>. Acesso em: 29 set. 2013.

NETTO, Coelho. Mandovi. *In*: NETTO, Coelho. **Sertão**. Porto: Chardron, 1926c. p. 208-225.

_____. O enterro. *In*: NETTO, Coelho. **Sertão**. Porto: Chardron, 1926a. p. 63-70.

_____. Turbilhão. *In*: NETTO, Coelho. **Obra seleta**: romances. V. 01. Biblioteca Luso-brasileira. Rio de Janeiro: José Aguilar, 1958. p. 827-1067.

_____. Firmo, o Vaqueiro. *In*: NETTO, Coelho. **Sertão**. Porto: Chardron, 1926b. p. 121-129.

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

_____. Introdução geral: imagem de uma vida. *In:* **Coelho Netto:** obra seleta: romances. V. 01. Biblioteca Luso-brasileira. Rio de Janeiro: José Aguilar, 1958. p. LXXXIII-CVI.

NEVES, Maria H. M. **Gramática de usos do português**. São Paulo: UNESP, 1999.

NILC/USP/LACIO-*WEB*. Compilação de *Corpus* do Português do Brasil e Implementação de Ferramentas para Análises Linguísticas, 2002. Disponível em: <http://www.nilc.icmc.usp.br/lacio*WEB*/index.htm>. Acesso em: 06 jan. 2013.

NISKIER, A. Coelho neto e a modernidade**. Artigos ABL/Jornal do Comércio**, 2010. Disponível em: <http://www.academia.org.br/abl/cgi/cgilua.exe/sys/start.htm?infoid=10142&sid=679>. Acesso em: 15 jan. 2013. Não paginado.

NLTK.ORG. Natural Language Toolkit. Disponível em: <http://nltk.org/>. Acesso em: 28 jan. 2013.

NUNES, M. das G. V. *et al*. Desafios na construção de recursos linguístico-computacionais para o processamento automático do português do Brasil. *In*: BERBER SARDINHA, T. (org.). **A língua portuguesa no computador**. Campinas, SP: Mercado de Letras, 2005. p. 33-70.

OLIVEIRA, Claudia; FREITAS, M. Cláudia de. Classes de palavras e etiquetagem na Linguística Computacional. **Calidoscópio**, Vol. 4, n. 3, p. 179-188, set/dez 2006.

OLIVEIRA, Lúcia P. de. Linguística de *corpus*: teoria, interfaces e aplicações. **Matraga**, Rio de Janeiro, v.16, n.24, jan./jun. 2009. Disponível em: <http://www.pgletras.uerj.br/matraga/matraga24/arqs/matraga24a02.pdf>. Acesso em: 29 set. 2013.

OLIVEIRA, Lúcia P. de; DIAS, Maria C. P. Compilação de *corpus*: representatividade e o CORPOBRAS. **Calidoscópio**, Vol. 7, n. 3, p. 192-198, set/dez 2009. Disponível em: < http://revistas.unisinos.br/index.php/calidoscopio/article/view/4872>. Acesso em: 29 set. 2013.

OTHERO, Gabriel de A.; MARTINS, Ronaldo T. *Parsing* do português. *In:* ALENCAR, Leonel F. de; OTHERO, Gabriel. de A. **Abordagens computacionais da teoria da gramática**. Campinas, SP: Mercado de Letras, 2011. p. 99-125.

OTHERO, Gabriel de A.; MENUZZI, Sérgio de M. **Linguística Computacional**: teoria & prática. São Paulo: Parábola, 2005.

PACHECO, W. Lúcio; LAPORTE, Éric. Descrição do verbo cortar para processamento automático de linguagem natural*. In:* LAPORTE, Éric (org). **Dialogar é preciso**: linguística para o processamento de línguas. Vitória-ES: PPGEL/UFES, 2013.

PADILHA, Tarcísio. **Discurso de posse na Academia Brasileira de Filosofia**, 2005. Disponível em: <http://www.academia.org.br/abl/cgi/cgilua.exe/sys/start.htm?infoid=11840&sid=98>. Acesso em: 07 out. 2013.

PADILHA, Thereza P. P.; LACERDA, Adriana N. Reconhecimento de Textos para Construção de Mapas Conceituais em Ambientes Colaborativos. *In:* BRAZILIAN

SYMPOSIUM ON COLLABORATIVE SYSTEMS, 2012. **Anais...** São Paulo: SBSC, 2012. Disponível em: <http://sws2012.ime.usp.br/sbsc/SBSC2012/data/4890a001.pdf>. Acesso em: 02 out. 2013.

PAIXÃO DE SOUSA, Mª Clara; KLEPER, Fábio N.; FARIA, Pablo P. F. de. E-dictor: novas perspectivas na codificação e edição de *corpora* de textos históricos. *In*: SHEPHERD, Tania M. G.; BERBER SARDINHA, Tony; PINTO, Marcia V. (org). **Caminhos da linguística de *corpus***. Campinas, SP: Mercados das Letras, 2012. p. 191-223 (Série Espaços da Linguística de *Corpus*).

PELLEGRINI, Jaqueline R.; PETKOFF, Juliana V. Estatística lexical em obras de Machado de Assis: um exercício exploratório. *In:* Salão de Iniciação Científica (21: 2009 out. 19-23: Porto Alegre, RS). **Apresentação...** Porto Alegre: UFRGS, 2009. Disponível em: <http://www.ufrgs.br/textecc/textquim/arquivos/Jaqueline_e_Juliana.pdf>. Acesso em: 04 out. 2013.

PINHEIRO, Gisele M.; ALUÍSIO, Sandra M. ***Corpus* Nilc**: Descrição e análise crítica com vistas ao projeto Lacio-*WEB (NILC-TR-0303).* São Carlos, SP: Universidade Federal de São Carlos, 2003. Disponível em: <http://www.nilc.icmc.usp.br/lacio*WEB*/downloads/NILC-TR-03-04.zip>. Acesso em: 29 set. 2013.

PINHO, Adeíltalo M. O sistema literário de A Conquista: nomes, leitura e números para um romance de Coelho Neto. **Revista Literatura em Debate** V.3, n.4, p. 109-128, 2009. Disponível em: <http://www.fw.uri.br/publicacoes/literaturaemdebate/artigos/n4_9.pdf>. Acesso em: 20 jan. 2013.

PYTHON.ORG. Python Programming Language – Official *WEB*site, 2013. Disponível em: <http://www.python.org/> Acesso em: 28 jan. 2013.

ROCHA, Marco. Relações anafóricas no português falado: uma abordagem baseada em corpus. **DELTA** vol.16 nº 2, 229-261, São Paulo, 2000. Disponível em: <http://www.scielo.br/pdf/delta/v16n2/a02v16n2.pdf>. Acesso em: 05 out. 2013.

ROMANOWSKI, Joana P.; ENS, Romilda T. As pesquisas denominadas do tipo 'estado da arte' em educação. **Diálogo Educacional**, Curítiba, v.6, n.19, p.37-50, set./dez., 2006. Disponível em: <http://www2.pucpr.br/reol/index.php/DIALOGO?dd1=237&dd99=view>. Acesso em: 20 jul. 2013.

SANTOS, Diana. Corporizando algumas questões. In: TAGNIN, Stela E.O.; VALE, Oto. A. (Org). **Avanços da linguística de *corpus* no Brasil**. São Paulo: Humanitas, 2008. P. 41-66.

_____. **Processamento computacional da língua portuguesa**: documento de trabalho. Lisboa: Ministério da Ciência e Tecnologia, 1999. Disponível em: <http://www.linguateca.pt/branco/proc_comp_port.html>. Acesso em: 16 set. 2013. Não paginado.

SARMENTO, Simone. **O uso dos verbos modais em manuais de aviação em inglês:** um estudo baseado em *corpus*. Tese (Doutorado em Estudos da Linguagem – Instituto de Letras), Universidade Federal do Rio Grande do Sul, Porto Alegre,

2008. Disponível em: <http://www.lume.ufrgs.br/bitstream/handle/10183/15568/000685529.pdf?sequence=1>. Acesso em: 29 set. 2013.

SCHER, A. P. **As construções com o verbo leve DAR e nominalizações em –ada no português do Brasil**. Tese (Doutorado em Letras). UNICAMP, Campinas, são Paulo, 2004.

SHEPHERD, Tania M. G. Panorama da linguística de *corpus*. *In*: SHEPHERD, Tania M. G.; BERBER SARDINHA, Tony; PINTO, Marcia V. (org). **Caminhos da linguística de *corpus***. Campinas, SP: Mercados das Letras, 2012. P. 14-30 (Série Espaços da Linguística de *Corpus*).

SHEPHERD, Tania M. G.; BERBER SARDINHA, Tony; PINTO, Marcia V. (org). **Caminhos da linguística de *corpus***. Campinas, SP: Mercados das Letras, 2012. (Série Espaços da Linguística de *Corpus*).

SILVEIRA, Felipe P. da. **Integração de ferramentas para compilação e exploração de *corpora***. Dissertação (Mestrado em Ciência da Computação). Fac. de Informática, PUCRS, 2008. Disponível em: <http://tede.pucrs.br/tde_busca/arquivo.php?codArquivo=2211>. Acesso em: 02 out. 2013.

SINCLAIR, John. *Corpus* and Text - Basic Principles. *In*: M. WYNNE (ed.). Developing linguistic *corpora*: a guide to good practice. Oxford, **Oxbow Books**, p. 1-16, 2005. Disponível em: http://ahds.ac.uk/linguistic-*corpora*/. Acesso em: 30 out. 2006.

_____. Tipologia Textual EAGLES. *In:* XI Encontro Nacional da Associação Portuguesa de Linguística**. Actas...** Lisboa: Faculdade de Letras da Universidade de Lisboa, 1995, v. 1. p. 39-91. Disponível em: <http://www.apl.org.pt/docs/actas-11-encontro-apl-1995_vol1.pdf>. Acesso em: 29 set. 2013.

SOUZA, Eurides A. de; GATTI, Iris K. Linguística de *corpus*: conceito, noções gerais e aplicação*.* **Pandaemonium germanicum,** 6/2002, 237-251. Revista de estudos germanísticos. Disponível em: <http://www.fflch.usp.br/dlm/alemao/pandaemoniumgermanicum/site/images/pdf/ed2002/14.pdf>. Acesso em: 07 fev. 2013.

TAGNIN, Stela E.O.; VALE, Oto. A. (Org). **Avanços da linguística de *corpus* no Brasil**. São Paulo: Humanitas, 2008.

TAGNIN, Stella E. O. Disponibilização de *corpora online*: os avanços do Projeto COMET. *In*: TAGNIN, Stela E.O.; VALE, Oto. A. (Org). **Avanços da linguística de *corpus* no Brasil**. São Paulo: Humanitas, 2008. p. 101-121.

_____. Glossário de linguística de *corpus. In:* ***Corpora*** **no Ensino de Línguas Estrangeiras**. São Paulo: HUB Editorial, 2010. p. 357-360. Disponível em: <http://www.hubeditorial.com.br/site/recursos/5_glossario/glossario_423.pdf>. Acesso em: 29 set. 2013.

TOGNINI BONELLI, E. Theoretical overview of the evolution of *corpus* linguistics. *In*: O´KEEFFE, A. e McCARTHY, M. (eds.) **The Routledge Handbook of *Corpus* Linguistics***.* Londres e Nova York: Routledge, 2010. Disponível em: <http://books.google.com.br/books?id=DrTsXGY_NaEC&pg=PA14&lpg=PA14&dq=theoretical+overview+of+the+evolution+of+corpus+linguistics&source=bl&ots=e_kWew

GXM3&sig=7WKuNvdBUBT73kV_Ifve5_WSdUM&hl=pt-BR&sa=X&ei=nVA7UsfKH5DK9gSTqYDgCQ&ved=0CFcQ6AEwBA#v=snippet&q=theoretical&f=false>. Acesso em: 19 set. 2013.

TOMAZELA, Élen C.; BARROS, Cláudia D.; RINO, Lucia H. M. Avaliação da anotação semântica do PALAVRAS e sua pós-edição manual para o *Corpus* Summ-it. **Linguamática** 2.3 (2010), pp. 29-42. ISSN: 1647—0818. Disponível em: <http://linguamatica.com/index.php/linguamatica/article/download/74/99>. Acesso em: 07 jan. 2013.

TOSH, Wayne. Linguística Computacional. *In:* HILL, Archibald A. **Aspectos da linguística moderna**. São Paulo: Cultrix, 1972.

UNIVERSIDADE FEDERAL DO CEARÁ. Guia de normalização de trabalhos acadêmicos da UFC, 2012. Disponível em: <http://www.biblioteca.ufc.br/images/stories/arquivos/bibliotecauniversitaria/guia_normalizacao_ufc_2012.pdf>. Acesso em: 14 jan. 2013.

UFC/TEDE. Biblioteca Digital de Teses e Dissertações. Disponível em: http://www.teses.ufc.br/. Acesso em 28 abr. 2014.

VIEIRA, Renata; LOPES, Lucelene. Processamento de linguagem natural e o tratamento computacional de linguagens científicas. *In:* PERNA, Cristina L.; DELGADO, Heloísa K.; FINATTO, Mª José (orgs.). **Linguagens especializadas em *corpora***: modos de dizer e interfaces de pesquisa (recurso eletrônico). Porto Alegre: EDIPUCRS, 2010. Disponível em: < www.pucrs.br/edipucrs/linguagensespecializadasem*corpora*.pdf>. Acesso em: 02 out. 2013.

VIEIRA, Renata; STRUBE DE LIMA, V. L. Linguística Computacional: princípios e aplicações. *In*: NEDEL, Luciana P. Nedel. (Org.). IX ESCOLA DE INFORMÁTICA DA SBC-Sul. **Anais...** Porto Alegre - RS: UFRGS, 2001, v. 1, p. 27-61. Disponível em: <http://www.inf.unioeste.br/~jorge/MESTRADOS/LETRAS%20-%20MECANISMOS%20DO%20FUNCIONAMENTO%20DA%20LINGUAGEM%20-%20PROCESSAMENTO%20DA%20LINGUAGEM%20NATURAL/ARTIGOS%20INTERESSANTES/lingu%EDstica%20computacional.pdf>. Acesso em: 28 jan. 2013.

VLCEK, Nathalie P. **A gramaticalização de advérbios em *-mente*/ment no português e no francês**: uma análise histórica. 2011. 107 f. Dissertação (Mestrado em Linguística da Faculdade de Letras). Programa de Pós-Graduação em Linguística da Universidade Federal do Rio de Janeiro. Rio de Janeiro: UFRJ, 2011.

ZYNGIER, Sônia. Polifonia de discursos: análise computacional de um *corpus* literário. **Texto Digital**, v. 1, n. 1, 2004. Universidade Federal de Santa Catarina, Florianópolis, Santa Catarina. Disponível em: <https://periodicos.ufsc.br/index.php/textodigital/article/view/1274/985>. Acesso em: 29 set. 2013.

ZYNGIER, Sônia; VIANA, Vander; SILVEIRA, Natália G. Discurso literário e linguística de *corpus*: uma visão empírica. **Cadernos de Letras (UFRJ)** n.28 – jul. 2011. Disponível em: <http://www.letras.ufrj.br/anglo_germanicas/cadernos/numeros/072011/textos/cl2831072011zyngier.pdf>. Acesso em: 14 abr. 2014.

APÊNDICE A – *CORPORA* DO PORTUGUÊS BRASILEIRO

1. **Banco de Português do Direct**[44] (grupo de pesquisa da PUCSP), criado em 1991, possui 750 milhões de palavras, de fala e escrita, de português contemporâneo. Atualmente, 2,2 milhão de palavras estão disponíveis para consulta *online* no site do CEPRIL/LAEL[45] (Centro de Pesquisa, Recursos e Informação em Linguagem);
2. **O projeto Corpus Brasileiro (GELC/PUCSP)**[46] é uma coletânea de aproximadamente um bilhão de palavras de português brasileiro, de vários tipos de linguagem, e é resultado de projeto coordenado por Tony Berber Sardinha. O *corpus* está disponível uma parte pela Linguateca, ou acessado integralmente via SketchEngine[47] (sistema de consulta de *corpus*). De textos literários constam contos, crônicas e biografias, mas não especificam quais. É um banco de palavras não de textos;
3. **COMET** (*Corpus* Multilíngue para Ensino e Tradução), *corpus* organizado pelo projeto COMET[48] (grupo de pesquisa da USP) apresentado oficialmente em 2001 e compreendendo três *subcorpora*: o CorTec que tem 200.000 palavras em inglês e português de textos técnicos; o CorAprend, constituído de redações de alunos de graduação dos cursos do Deptº de Letras Modernas da USP e o CorTrad, composto de textos originais de inglês e português (com suas respectivas traduções em ambas as línguas). O CorTrad "prevê textos de literatura estrangeira traduzida para o português e de literatura brasileira traduzida para línguas estrangeiras, além de um *subcorpus* de textos técnicos-científicos" (TAGNIN, 2008, p. 99). Atualmente, os textos literários que fazem parte desse *corpus* são somente 28 contos australianos traduzidos para o português;
4. ***Corpus* de Araraquara** (UNESP), com cerca de 220 milhões de ocorrências de palavras, que vão desde o século XVI até hoje. Com variedade ampla de textos, é composto de:

[44] *Corpus* Direct - http://www2.lael.pucsp.br/direct/
[45] CEPRIL/LAEL - http://www2.lael.pucsp.br/*corpora*/bp/
[46] Corpus Brasileiro - http://corpusbrasileiro.pucsp.br/cb/Inicial.html
[47] SketchEngine - http://www.sketchengine.co.uk/
[48] Projeto COMET - http://www.fflch.usp.br/dlm/comet/

> [...] poemas, romances, peças de teatro, crônicas, oratória (tanto os célebres Sermões do padre Antônio Vieira quanto discursos de políticos contemporâneos), textos jornalísticos, propaganda veiculada em jornais e revistas, periódicos técnicos e científicos, textos traduzidos e letras de música (LOPES, 2009, p. 27).

Sem o acesso a esse *corpus* não é possível detectar quais textos literários além de Vieira estão compilados. Borba (2007) fazendo uma referência sobre o *corpus* afirma ter extraído frases de uma obra de Coelho Netto: *Cidade Maravilhosa,* livro de contos de 1928. Este *Corpus* foi ponto de partida para a geração do Dicionário de Usos do Português do Brasil, de Borba. E como extensão desse trabalho, estava previsto para 2013 a publicação do Grande Dicionário do Português do Brasil, pela Editora Unesp, inclusive como edição *on-line*;

5. ***Corpus* NILC**, desenvolvido dentro do projeto ReGra do Núcleo Interinstitucional de Linguística Computacional, é composto por mais de 35 milhões de palavras de textos em prosa (corrigidos – livros, jornais e revistas; não corrigidos – redações, monografias e textos de publicidade; e semicorrigidos – contratos, relatórios, dissertações e teses acadêmicas). É um dos primeiros *corpora* que contêm vários gêneros de texto para o português brasileiro, mas predominam os textos jornalísticos, sobretudo dos textos corrigidos. Serviu "como *corpus* de base para diversos aplicativos desenvolvidos pelo NILC" (PINHEIRO; ALUISIO, 2003, p. 5);

6. **Lácio-*WEB***[49], originado do *corpus* NILC, iniciado em 2002, consiste na construção de seis tipos de *corpus*: Lácio-Ref, Lácio-Dev, Par-C, Comp-C, Mac-Morpho e Lácio-Sint. Contém 10 milhões de palavras do português contemporâneo. Dois deles foram disponibilizados livremente: uma versão do Lácio-Ref para pesquisa e geração de *subcorpus* e o Mac-Morpho para *download*, que é constituído somente por artigos jornalísticos retirados da Folha de São Paulo, do ano de 1994. O *Corpus* Lácio-Ref é um *corpus* aberto composto de textos em PB, caracterizado por serem escritos respeitando a norma culta. Não é anotado, portanto não possibilita informações morfossintáticas para análise. A maioria dos textos está disponibilizada na íntegra, mas não foi possível visualizá-los pelo *site*. Só estarão disponíveis após

[49] *Corpus* Lácio-WEB - http://www.nilc.icmc.usp.br/lacioweb/descricao.htm

solicitação de autorização de uso, inclusive os literários (NILC/USP/LACIO-*WEB*, 2002). Portanto, a princípio, não há como saber que textos fazem parte deste *corpus;*

7. **CORPOBRAS** (*Corpus* Representativo do POrtuguês do BRAsil) é um *corpus* fechado, contendo 27 (vinte e sete) gêneros discursivos: 20 (vinte) gêneros do discurso escrito, 5 (cinco) gêneros do discurso oral, e 2 (dois) gêneros do discurso escrito para ser falado. Em 2009 continha mais de um milhão de palavras, ultrapassando sua meta inicial. Contém 14 contos e 28 romances, não especificados nas informações. Não é um *corpus* anotado, tarefa a ser executada nas próximas etapas do projeto (OLIVEIRA; DIAS, 2009);
8. ***CORPUS* compartilhado diacrônico**[50] é composto de cartas e bilhetes pessoais (UFRJ) - edição em *fac-símile* de cartas particulares escritas no Rio de Janeiro por brasileiros e por portugueses, a partir dos séculos XVIII-XIX, localizadas em acervos cariocas, como é o caso do Arquivo Nacional e do Instituto Histórico Geográfico Brasileiro (digitados);
9. **Projeto CE-DOHS,** *corpus* eletrônico de documentos históricos do sertão (PHPB-BA)**:** textos manuscritos, impressos e orais. O acervo desse *corpus* é composto de 1.037 cartas particulares (1808-2000), escritas por 418 remetentes (nascidos entre 1724 e 1980), além de 25 cartões. “Embora os acervos, em sua maioria, sejam originados da grande área do semiárido baiano, há acervos de outras áreas da Bahia e também de diversas regiões do Brasil” [51];
10. **O *corpus* C-ORAL-BRASIL** (UFMG) - é constituído de textos da fala espontânea do PB (UFMG): textos orais produzidos em contexto natural;
11. **CORPTEXLIT** (*Corpus* de Língua Portuguesa de Textos Literários do Século XIX) é uma experiência do grupo de estudos CompLin (Computação e Linguagem Natural)[52] da Universidade Federal do Ceará (UFC), que anotará, morfossintaticamente, “textos de literatura brasileira do século XIX que compreenderá 40 obras do período [..] cerca de 2.500.000 *tokens*[53], com 10% a serem revistos manualmente” (ALENCAR, 2010b), *corpus* em construção de textos literários do século XIX de 40 (quarenta) obras de autores desse período,

[50] Disponível em: http://www.letras.ufrj.br/laborhistorico/*corpora*.htm
[51] Texto disponível em CE-DOHS – http://www.tycho.iel.unicamp.br/cedohs/*corpora*.html
[52] Homepage do Grupo CompLin - http://complin.blogspot.com.br/
[53] *Tokens* compreendido como conjuntos de palavras, números, sinais de pontuação, espaço, entre outros encontrados nos *corpora* (CANDIDO JR; ALUÍSIO, 2008)

todos anotados morfossintaticamente, totalizando 2,5 milhões de *tokens*. Pretende enriquecer o acervo do CHPTB, que dispõe de poucos textos brasileiros do século XIX. O projeto teve início em 2010, com previsão de término para 2014. Até o momento, o grupo realizou a anotação do romance *Luzia-Homem*, de Domingos Olímpio, publicado em 1903, dos oito primeiros capítulos manualmente corrigidos, praticamente 25% do total do romance (ALENCAR, 2010b).

É um projeto sem financiamento direto das principais agências de fomentos à pesquisa no Brasil, como CAPES e CNPQ, e conta somente com a contribuição dos participantes do grupo CompLin, mas como frisa Alencar (2010b), o "projeto está aberto, igualmente, à participação de quem quer que, imbuído dessa filosofia, se disponha a colaborar. Especialmente bem-vinda é a colaboração na revisão dos textos". Além disso, o CORPTEXLIT está gradualmente sendo distribuído livremente à comunidade de estudantes e pesquisadores, desde que seu uso seja feito sem pretensões comerciais (ALENCAR, 2010b).

O *Corpus* serve de base para geração de versões mais robustas do etiquetador automático Aelius (ALENCAR, 2010b), adotando o mesmo sistema de anotação morfossintática do CHPTB, contribuindo para o enriquecimento do acervo deste *Corpus*. Uma visualização de uma amostra do trabalho já realizado, de anotação e correção manual do *Corpus*, pode ser visto no *site* do Grupo CompLin[54].

12. **PROHPOR-UFBa** (Programa para a História da Língua Portuguesa), organizado por um grupo de pesquisa vinculado ao Departamento de Letras Vernáculas do Instituto de Letras da Universidade Federal da Bahia, que utiliza o Tycho na produção de *corpora* de português antigo e clássico, composto por textos escritos nas diversas fases da história da Língua Portuguesa, especificamente do Português Arcaico (séc. XIII a XVI), e basicamente de textos literários (MATTOS E SILVA, 2007).
13. **CETENFolha**, um *corpus* de textos eletrônicos da Folha de São Paulo, organizado pelo NILC da Universidade Federal de São Carlos, constituinte do projeto Lácio-*WEB*;
14. ***Corpus* Brasileiro** é uma coletânea de aproximadamente um bilhão de palavras, resultado de projeto coordenado por Tony Berber Sardinha (PUC-SP),

[54] ComPlin/CORPTEXLIT – disponível em: http://complin.blogspot.com.br/search?updated-max=2012-04-05T08:36:00-07:00&max-results=7&start=7&by-date=false

contendo diversos tipos de textos. Apesar de este *corpus* conter textos literários do PB, estão disponíveis somente através de busca por resultados e busca de frequência de ocorrência dos termos, pois estão divididos em frases. Não se tem acesso ao texto integral dos autores.

APÊNDICE B – *CORPORA* DO PORTUGUÊS BRASILEIRO E PORTUGUÊS EUROPEU

1. ***Corpus* Histórico do Português Tycho Brahe (CHPTB)**[55], *corpus* eletrônico parcialmente anotado, com livre acesso pela internet, é basicamente composto de textos em português escritos por autores, inicialmente europeus, nascidos entre 1380 e 1845, e de textos do português brasileiro (PB), cujos textos disponíveis são: Atas dos Brasileiros, Cartas Brasileiras, Atas dos Africanos; Cartas para Cícero Dantas Martins (Barão de Jeremoabo), Cartas para vários destinatários, Maria ou a menina roubada, e Iracema: destes só os dois primeiros estão anotados morfossintaticamente. Atualmente, contém 62 textos literários com 2.698.928 palavras, com anotação linguística realizada de duas formas: anotação morfológica (aplicada em 33 textos, num total de 1.485.943 palavras); e anotação sintática (aplicada em 16 textos, num total de 671.694 palavras) (GALVES; FARIA, 2012): está vinculado ao projeto Padrões Rítmicos, Fixação de Parâmetros & Mudança Linguística da UNICAMP[56], e associado ao Projeto Matemática, Computação, Linguagem e Cérebro (MaCLinC) da USP[57].

   Os textos do CHPTB, anotados automaticamente, servem de base de estudo para a produção de teses, dissertações e outros trabalhos sobre a morfologia e a sintaxe do português clássico (PAIXÃO DE SOUSA; KLEPER; FARIA, 2012), como, por exemplo, o CORPTEXLIT (*Corpus* de Língua Portuguesa de Textos Literários do Século XIX) e o PROHPOR-UFBa (Programa para a História da Língua Portuguesa).

   O Tycho utiliza o eDictor[58], ferramenta que combina um editor de XML e um etiquetador morfossintático, gerando versões editadas em *html*, e está voltado ao trabalho filológico e à análise linguística automática (PAIXÃO DE SOUSA; KLEPER, FARIA, 2012);

---

[55] CHPTB – http://www.tycho.iel.unicamp.br/~tycho/*corpus*/en/index.html

[56] Projeto Padrões Rítmicos, Fixação de Parâmetros & Mudança Linguística – http://www.tycho.iel.unicamp.br/~tycho/prfpml/fase2/index.html

[57] Projeto MaCLinC – http://numec.prp.usp.br/MaCLinC/apresentacao?set_language=pt-br

[58] eDictor – http://humanidadesdigitais.org/edictor/. Apesar de o site disponibilizar o link para utilização do eDictor por download, o mesmo não deu acesso, nem a versão mais recente em *web-based*.

2. **O *Corpus* do Português**[59] é um *corpus* de textos da língua portuguesa que contém mais de 45 milhões de palavras de quase 57.000 textos em português do século XIV ao século XX. A interface do *Corpus* permite pesquisar palavras exatas ou frases, curingas, lemas, classes gramaticais, ou qualquer combinação destes. É possível fazer comparações da frequência e distribuição de palavras, frases e construções gramaticais através de textos, diferenciando o Registro, o Dialeto e ou o Período Histórico. Foi compilado e é mantido pelos pesquisadores Mark Davies, professor de Linguística da Universidade Brigham Young e Michael J. Ferreira (Universidade de Georgetown), com suporte financeiro proveniente do *U.S. National Endowment for the Humanities*. É um *corpus* de banco de palavras, não especificando os textos utilizados para sua composição;
3. **CINTIL**[60] (*Corpus* Internacional do Português) é um *corpus* de treino do PE, com uma rica variedade de gêneros e tipos de textos, que contém atualmente um milhão de palavras anotadas e corrigidas manualmente por especialistas. Mais de um terço do *corpus* é constituído por transcrições de gravações orais, sendo que metade delas consiste em transcrições de conversas informais. Dos textos escritos, a maioria inclui artigos de jornais e revistas portugueses, e o restante é constituído por textos literários (16,80%): o *site* não especifica quem são os autores desses textos. A versão atual do CINTIL foi desenvolvida e é mantida na Universidade de Lisboa pelo Grupo REPORT[61], parceria entre o Centro de Linguística e o Departamento de Informática da Universidade, entre 2004 e 2006. Disponibiliza no site uma interface de seu Concordanciador[62], que é um serviço *online* gratuito de extração de concordâncias para a pesquisa de linguística do *Corpus* do CINTIL e aberto para outras pesquisas;
4. **CRPC**[63] **-** O Corpus de Referência do Português Contemporâneo é um corpus composto maioritariamente da variedade europeia do Português, mas encontram-se também representadas no *corpus* outras variedades nacionais, como o Português do Brasil, de África (Angola, Cabo Verde, Guiné-Bissau, Moçambique e São Tomé e Príncipe) e da Ásia (Macau, Goa e Timor-Leste).

---

[59] Corpus do Português - http://www.corpusdoportugues.org/
[60] ***Corpus* CINTIL** - http://cintil.ul.pt/pt/cintilfeatures.html
[61] Coordenado por António Branco e Maria Fernanda Bacelar do Nascimento.
[62] Concordanciador CINTIL – http://cintil.ul.pt/pt/index.jsp (ferramenta que permite a busca de padrões através de expressões regulares).
[63] CRPC - http://www.clul.ul.pt/pt/recursos/183-reference-corpus-of-contemporary-portuguese-crpc

Contém 311,4 milhões de palavras, com diferentes tipos de textos escritos (literário, jornalístico, técnico, etc.) e de registos orais (formal e informal), da segunda metade do século XIX até 2006, embora a maioria dos textos seja posterior ao ano de 1970. Dos textos literários, encontramos 417 textos literários do PB e PE. Na relação disponível no *site* encontramos relacionados os textos de Coelho Neto (*A Conquista*), Aluísio Azevedo (*O Cortiço*) e Camilo Castelo Branco (*Amor de Perdição*). Os textos foram automaticamente toquenizados com o toquenizador LX Tagger. Todos os textos são anotados morfossintaticamente, pelo etiquetador adaptado da parte escrita do *corpus* CINTIL. O sistema de anotação usado contém um conjunto de 80 etiquetas. Em uma busca rápida, observarmos que os textos não estão disponíveis de forma integral, o acesso aos dados só é feito via concordanciador.

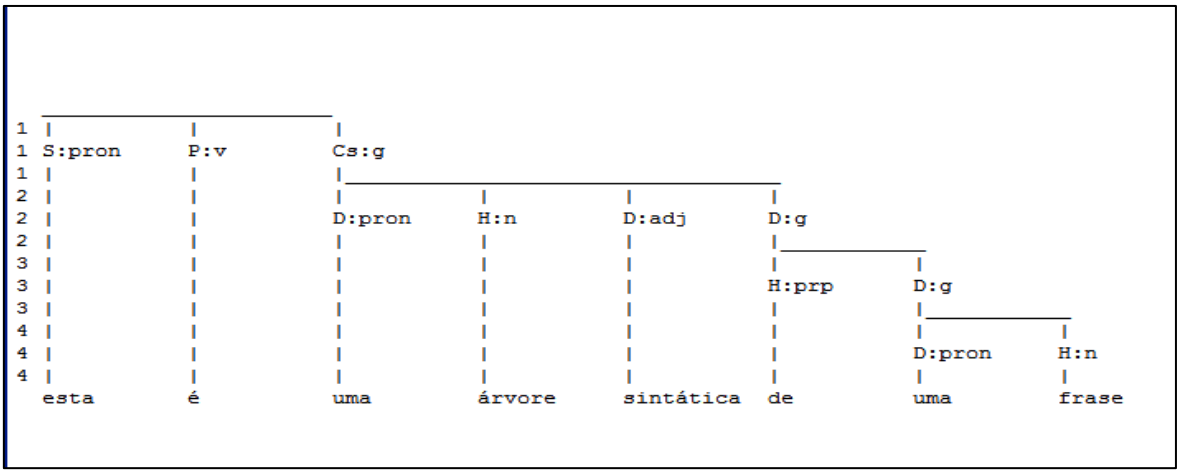
5. **Floresta Sintática**, disponível no Linguateca, contém textos do PE e PB e são todos anotados automaticamente pelo analisador sintático PALAVRAS[64] (BICK, 2005). Gera um conjunto de árvores sintáticas usadas no processamento computacional da língua portuguesa: são frases analisadas morfossintaticamente explicitando hierarquicamente informações relativas à estrutura de seus constituintes: uma frase sintaticamente analisada se parece com uma árvore, donde um conjunto de árvores constitui uma floresta sintática, ou seja, “nada mais é que um conjunto de frases analisadas sintaticamente e com informação relativa aos níveis de constituintes” (FREITAS; ROCHA; BICK, 2008, p. 142).

   Está vinculado ao projeto VISL (*Visual Interactive Syntax Learning*)[65] do Instituto de Linguagem e Comunicação da Universidade do Sul da Dinamarca que teve início em 1996, que constitui-se "em um *parser* que fornece etiquetagem sintática e morfossintática" (BERBER SARDINHA, 2004, p. 136), baseado em análise computacional automática, e contempla um núcleo de ferramentas computacionais e bancos de dados linguísticos para serem utilizados *on-line*. As mais recentes atividades do projeto concentram-se nas áreas de semântica e tradução automática, e a compilação e etiquetagem de *corpora* (AFONSO, 2002).

[64] O PALAVRAS é um analisador automático (*tagger-parser*) para português que foi desenvolvido por Eckhard Bick no contexto dum projeto de doutoramento, na Dinamarca, em 2000. Fornece análise sintática e morfológica. Texto disponível em: http://www.linguateca.pt/treebank/InicialFloresta.html.
[65] VISL - http://beta.visl.sdu.dk/visl/pt/

A Figura 28 é uma demonstração de uma árvore sintática utilizando a frase: “Esta é uma árvore sintática de uma frase”, gerada pelo VISL.

Figura 26 – Representação de uma árvore sintática gerada pelo Machine *Analysis* do VISL.



Fonte: http://beta.visl.sdu.dk/visl/pt/parsing/automatic/trees.php, em 30 abr. 2014.

O Floresta Sintática é dividido em quatro partes:

- **Floresta Virgem**: “contém cerca de 1.6 milhões de palavras (95 mil frases)” coletadas do CETENFolha (parte do *corpus*NILC/São Carlos: textos da Folha de São Paulo de 1994) e do CETEMPúblico (textos do jornal português PÚBLICO): anotadas automaticamente, mas sem revisão linguística (FREITAS, ROCHA, BICK, 2008, p. 146);
- **Amazônia**: Possui somente textos brasileiros e todas as palavras são anotadas e também, como a Floresta Virgem, não foi revisto por linguistas (FREITAS, ROCHA, BICK, 2008, p. 146): “contém 3.8 milhões de palavras (cerca de 194 mil frases), com 4.070 textos e 1.303 autores”;

- **Selva**: “contém cerca 300 mil palavras divididas entre diferentes modalidades (escrita e falada), gêneros, domínios e as variantes portuguesa e brasileira do Português”. A Selva é subdividida em três partes: A Selva falada; A Selva literária; A Selva científica. A Selva Literária é a única parte do Floresta que “contém textos literários brasileiros e portugueses do final do século XIX e do início do século XX - cerca de 10.000 palavras por autor - recolhidos na *Wikisource*[66], e ainda cerca de 10.000 palavras de literatura contemporânea” (FREITAS, ROCHA, BICK, 2008, p. 146). Os textos contemporâneos são dois contos da autoria de Luísa Coheur. Os textos anotados são dos seguintes escritores:

  José de Alencar, *O Guarani*, parte 4, cap XII
  Machado de Assis, *Memórias Póstumas de Brás Cubas*, cap. I a XV
  Lima Barreto, *Clara dos Anjos*, cap. I e II
  Euclides da Cunha, *Sertões*, parte 2, cap I e II
  Bernardo Guimarães, *A Escrava Isaura*, caps. X a XII
  Raúl Brandão, *Os Pobres*, caps. IV a X
  Camilo Castelo Branco, *Amor de Perdição*, caps. XIV a XIX
  Júlio Dinis, *As Pupilas do Senhor Reitor*, caps. I a VI
  Alexandre Herculano, *Arras por Foros de Espanha*, caps. I e II
  Eça de Queiróz, *Primo Basílio*, caps. IX e X (ROCHA e FREITAS, 2008).

- **Bosque**: Parte linguisticamente revista do Floresta. Contém 190 mil palavras, 9.368 frases, retiradas dos *corpora* CETENFolha (textos do jornal brasileiro Folha de S. Paulo) e CETEMPúblico (textos do jornal português Público)” (FREITAS; ROCHA; BICK, 2008, p. 146).

O Floresta não possui textos literários integrais para análise. As anotações da Selva e do Bosque estão sendo revistas por linguistas com o auxílio do Milhafre, que é uma ferramenta de busca desenvolvida especialmente para o Floresta Sintática (FREITAS; ROCHA; BICK, 2008).

[66] O *Wikisource* é uma biblioteca livre digital, de forma colaborativa, que abriga obras literárias e científicas todas de domínio público. Disponível em: https://pt.*Wikisource*.org/wiki/*Wikisource*:P%C3%A1gina_principal

APÊNDICE C – TEXTO DIGITAL ORIGINAL

ROMANCE: A CONQUISTA (1899) – CAPÍTULO I (QUATRO PARÁGRAFOS INICIAIS)

Revistas as últimas provas do conto de Aurélio Mendes o Anacharsis dos "Idílios pagãos", Paulo Jove arredou a cadeira e pôs-se de pé, desabafando. Doía-lhe a espinha e, como havia fumado quase todo o maço de cigarros, tinha a boca amarga e áspera, os olhos ardidos, não só do fumo e da claridade intensíssima das lâmpadas elétricas, como da fixidez atenta em que os mantinha desde as sete e meia até àquela hora alta da noite.

Curvou-se de mãos nas ilhargas, d'ímpeto esticou os braços, arrojou-os à frente com um ahn! surdo de atleta que exercita os músculos entorpecidos e desabou-os depois, com força, sacudindo-se todo, virando, revirando a cabeça, como em ânsia angustiosa. 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. A caldeira reboava num retroar soturno de caverna que repercutisse, sem descontinuar, o gorgorejo possante de águas encachoeiradas.

Na sala da revisão, estreita e abafada, mal comportando as quatro mesas de serviço, os revisores repousavam; apenas o Brites, esgalgado e míope, lia o antigo de fundo, todo em períodos lamentosos augurando fome e lutas; e o Amaro, conferente, acusando a pontuação de quando em quando batia na mesa pancadas secas com um lápis ou dizia claramente uma palavra, repetindo-a devagar, sílaba a sílaba, enquanto o Brites, debruçado sobre a prova, fazia a emenda resmungando.

APÊNDICE D – TEXTO COMPARADO, EDITADO E ATUALIZADO CONFORME NORMA ORTOGRÁFICA VIGENTE – em negrito
ROMANCE: TURBILHÃO (1899) – CAPÍTULO I (QUATRO PARÁGRAFOS INICIAIS)

Revistas as últimas provas do conto de Aurélio Mendes**, (acrescentado vírgula)** o Anacharsis dos "Idílios pagãos", Paulo Jove arredou a cadeira e pôs-se de pé, desabafando. Doía-lhe a espinha e, como havia fumado quase todo o maço de cigarros, tinha a boca amarga e áspera, os olhos ardidos, não só do fumo e da claridade intensíssima das lâmpadas elétricas, como da fixidez atenta em que os mantinha desde as sete e meia até àquela hora alta da noite.

Curvou-se de mãos nas ilhargas, **de ímpeto** esticou os braços, arrojou-os à frente com um ahn! surdo de atleta que **exercia** os músculos entorpecidos e desabou-os depois, com força, sacudindo-se todo, virando, revirando a cabeça, como em ânsia angustiosa. Levantou-os, de novo, acima da cabeça, as mãos juntas, estrincando os dedos enclavinhados e bocejou, espichando-se nas pontas dos pés caindo depois**, (acrescentado vírgula)** rijamente, sobre os tacões.

Já as primeiras páginas haviam descido para a clichagem. Embaixo, martelavam pancadas crebas, como de matracas. A caldeira reboava num retroar soturno de caverna que repercutisse, sem descontinuar, o gorgorejo possante de águas encachoeiradas.

Na sala da revisão, estreita e abafada, mal comportando as quatro mesas de serviço, os revisores repousavam: **(substituição do ponto e vírgula pelos dois pontos)** apenas o Brites, esgalgado e míope, lia o **artigo** de fundo, todo em períodos lamentosos**, (acrescentado vírgula)** augurando fome e lutas; e o Amaro, conferente, acusando a pontuação**, (acrescentado vírgula)** de quando em quando batia na mesa pancadas secas com um lápis ou dizia claramente uma palavra, repetindo-a devagar, sílaba a sílaba, enquanto o Brites, debruçado sobre a prova, fazia a emenda**, (acrescentado vírgula)** resmungando.

1. Os grifos, nosso, indicam as mudanças na grafia e correções de texto ao se comparar o impresso com o digital.
2. A correção do texto foi feita comparando o texto disponível no link acima e o texto impresso publicado pela Editora José Aguilar Ltda, Porto-PT, em 1958.

APÊNDICE E – TRECHO DE TEXTO ANOTADO MORFOSSINTATICAMENTE, APÓS EDIÇÃO.
ROMANCE: TURBILHÃO (1899) – CAPÍTULO I (QUATRO PARÁGRAFOS INICIAIS)

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 ./.
Doía/VB-D -/+ lhe/CL a/D-F espinha/N e/CONJ ,/, como/CONJS havia/HV-D fumado/N quase/ADV todo/Q o/D maço/N de/P cigarros/N-P ,/, tinha/TR-D a/D-F boca/N amarga/ADJ-F e/CONJ áspera/ADJ-F ,/, os/D-P olhos/N-P ardidos/VB-AN-P ,/, não/NEG só/FP do/P+D fumo/N e/CONJ da/P+D-F claridade/N intensíssima/ADJ-S-F das/P+D-F-P lâmpadas/N-P elétricas/ADJ-F-P ,/, como/CONJS da/P+D-F fixidez/N atenta/ADJ-F em/P que/WPRO os/CL mantinha/VB-D desde/P as/D-F-P sete/NUM e/CONJ meia/ADJ-F até/P àquela/P+D-F hora/N alta/ADJ-F da/P+D-F noite/N ./.

Curvou/VB-P -/+ se/SE de/P mãos/N-P nas/P+D-F-P ilhargas/N-P ,/, de/P ímpeto/N esticou/VB-D os/D-P braços/N-P ,/, arrojou/VB-D -/+ os/CL à/P+D-F frente/ADV com/P um/D-UM ahn/N !/. surdo/ADJ de/P atleta/N que/WPRO exercia/VB-D os/D-P músculos/N-P entorpecidos/VB-AN-P e/CONJ desabou/VB-D -/+ os/CL depois/ADV ,/, com/P força/N ,/, sacudindo/VB-G -/+ se/SE todo/Q ,/, virando/VB-G ,/, revirando/VB-G a/D-F cabeça/N ,/, como/CONJS em/P ânsia/N angustiosa/ADJ-F ./.
Levantou/VB-D -/+ os/CL ,/, de/P novo/ADJ ,/, acima/ADV da/P+D-F cabeça/N ,/, as/D-F-P mãos/N-P juntas/ADJ-F-P ,/, estrincando/VB-G os/D-P dedos/N-P enclavinhados/VB-AN-P e/CONJ bocejou/VB-D ,/, espichando/VB-G -/+ se/SE nas/P+D-F-P pontas/N-P dos/P+D-P pés/N-P caindo/VB-G depois/ADV ,/, rijamente/ADV ,/, sobre/P os/D-P tacões/N-P ./.

Já/ADV as/D-F-P primeiras/ADJ-F-P páginas/N-P haviam/HV-D descido/VB-PP para/P a/D-F clichagem/N ./. Embaixo/ADJ ,/, martelavam/VB-D pancadas/N-P crebas/ADJ-F-P ,/, como/CONJS de/P matracas/N-P ./. A/D-F caldeira/N reboava/VB-D num/P retroar/VB soturno/N de/P caverna/N que/C repercutisse/VB-

SD ,/, sem/P descontinuar/VB ,/, o/D gorgorejo/N possante/ADJ-G de/P águas/N-P encachoeiradas/VB-AN-F-P ./.

Na/P+D-F sala/N da/P+D-F revisão/N ,/, estreita/ADJ-F e/CONJ abafada/VB-AN-F ,/, mal/ADV comportando/VB-G as/D-F-P quatro/NUM mesas/N-P de/P serviço/N ,/, os/D-P revisores/N-P repousavam/VB-D :/. apenas/ADV o/D Brites/NPR ,/, esgalgado/VB-AN e/CONJ míope/NPR ,/, lia/VB-D o/D artigo/N de/P fundo/N ,/, todo/Q em/P períodos/N-P lamentosos/ADJ-P ,/, augurando/VB-G fome/N e/CONJ lutas/N-P ;/. e/CONJ o/D Amaro/NPR ,/, conferente/N ,/, acusando/VB-G a/D-F pontuação/N ,/, de/P quando/CONJS em/P quando/CONJS batia/VB-D na/P+D-F mesa/N pancadas/N-P secas/ADJ-F-P com/P um/D-UM lápis/N ou/CONJ dizia/VB-D claramente/ADV uma/D-UM-F palavra/N ,/, repetindo/VB-G -/+ a/CL devagar/ADV ,/, sílaba/N a/D-F sílaba/N ,/, enquanto/CONJS o/D Brites/NPR ,/, debruçado/VB-AN sobre/P a/D-F prova/N ,/, fazia/VB-D a/D-F emenda/N ,/, resmungando/VB-G ./.

APÊNDICE F – TRECHO DE TEXTO ANOTADO COM CORREÇÃO MANUAL (DOIS PARÁGRAFOS INICIAIS)

ROMANCE: TURBILHÃO (1899) – CAPÍTULO I (QUATRO PARÁGRAFOS INICIAIS)

Revistas/N-P@**VB-AN-F-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 ./.
Doía/VB-D -/+ lhe/CL a/D-F espinha/N e/CONJ ,/, como/CONJS havia/HV-D fumado/N@**VB-PP** quase/ADV todo/Q o/D maço/N de/P cigarros/N-P ,/, tinha/TR-D a/D-F boca/N amarga/ADJ-F e/CONJ áspera/ADJ-F ,/, os/D-P olhos/N-P ardidos/VB-AN-P ,/, não/NEG só/FP do/P+D fumo/N e/CONJ da/P+D-F claridade/N intensíssima/ADJ-S-F das/P+D-F-P lâmpadas/N-P elétricas/ADJ-F-P ,/, como/CONJS da/P+D-F fixidez/N atenta/ADJ-F em/P que/WPRO os/CL mantinha/VB-D desde/P as/D-F-P sete/NUM e/CONJ meia/ADJ-F até/P àquela/P+D-F hora/N alta/ADJ-F da/P+D-F noite/N ./.

Curvou/VB-P -/+ se/SE de/P mãos/N-P nas/P+D-F-P ilhargas/N-P ,/, de/P ímpeto/N esticou/VB-D os/D-P braços/N-P ,/, arrojou/VB-D -/+ os/CL à/P+D-F frente/ADV com/P um/D-UM ahn/N@**INT** !/. surdo/ADJ de/P atleta/N que/WPRO exercia/VB-D os/D-P músculos/N-P entorpecidos/VB-AN-P e/CONJ desabou/VB-D -/+ os/CL depois/ADV ,/, com/P força/N ,/, sacudindo/VB-G -/+ se/SE todo/Q ,/, virando/VB-G ,/, revirando/VB-G a/D-F cabeça/N ,/, como/CONJS em/P ânsia/N angustiosa/ADJ-F ./. Levantou/VB-D -/+ os/CL ,/, de/P novo/ADJ ,/, acima/ADV da/P+D-F cabeça/N ,/, as/D-F-P mãos/N-P juntas/ADJ-F-P ,/, estrincando/VB-G os/D-P dedos/N-P enclavinhados/VB-AN-P e/CONJ bocejou/VB-D ,/, espichando/VB-G -/+ se/SE nas/P+D-F-P pontas/N-P dos/P+D-P pés/N-P caindo/VB-G depois/ADV ,/, rijamente/ADV ,/, sobre/P os/D-P tacões/N-P ./.

Já/ADV as/D-F-P primeiras/ADJ-F-P páginas/N-P haviam/HV-D descido/VB-PP para/P a/D-F clichagem/N ./. Embaixo/ADJ@**ADV** ,/, martelavam/VB-D pancadas/N-P crebas/ADJ-F-P ,/, como/CONJS de/P matracas/N-P ./. A/D-F caldeira/N

reboava/VB-D num/P retroar/VB soturno/N@**ADJ** de/P caverna/N que/C repercutisse/VB-SD ,/, sem/P descontinuar/VB ,/, o/D gorgorejo/N possante/ADJ-G de/P águas/N-P encachoeiradas/VB-AN-F-P ./.

Na/P+D-F sala/N da/P+D-F revisão/N ,/, estreita/ADJ-F e/CONJ abafada/VB-AN-F ,/, mal/ADV comportando/VB-G as/D-F-P quatro/NUM mesas/N-P de/P serviço/N ,/, os/D-P revisores/N-P repousavam/VB-D :/. apenas/ADV o/D Brites/NPR ,/, esgalgado/VB-AN e/CONJ míope/NPR@**ADJ** ,/, lia/VB-D o/D artigo/N de/P fundo/N ,/, todo/Q em/P períodos/N-P lamentosos/ADJ-P ,/, augurando/VB-G fome/N e/CONJ lutas/N-P ;/. e/CONJ o/D Amaro/NPR ,/, conferente/N ,/, acusando/VB-G a/D-F pontuação/N ,/, de/P quando/CONJS em/P quando/CONJS batia/VB-D na/P+D-F mesa/N pancadas/N-P secas/ADJ-F-P com/P um/D-UM lápis/N ou/CONJ dizia/VB-D claramente/ADV uma/D-UM-F palavra/N ,/, repetindo/VB-G -/+ a/CL devagar/ADV ,/, sílaba/N a/D-F sílaba/N ,/, enquanto/CONJS o/D Brites/NPR ,/, debruçado/VB-AN sobre/P a/D-F prova/N ,/, fazia/VB-D a/D-F emenda/N ,/, resmungando/VB-G ./.

NOTA: A correção manual é feita inserindo o arroba (@) depois da etiqueta e em outro arquivo procedemos a correção com a etiqueta adequada.

APÊNDICE G – TRECHO DE TEXTO ANOTADO COM PADRÃO *GOLD*: APÓS CORREÇÃO MANUAL E CORREÇÃO AUTOMÁTICA POR PYTHON.

ROMANCE: TURBILHÃO (1899) – CAPÍTULO I (DOIS PARÁGRAFOS INICIAIS)

Revistas/**VB-AN-F-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 ./.
Doía/VB-D -/+ lhe/CL a/D-F espinha/N e/CONJ ,/, como/CONJS havia/HV-D fumado/**VB-PP** quase/ADV todo/Q o/D maço/N de/P cigarros/N-P ,/, tinha/TR-D a/D-F boca/N amarga/ADJ-F e/CONJ áspera/ADJ-F ,/, os/D-P olhos/N-P ardidos/VB-AN-P ,/, não/NEG só/FP do/P+D fumo/N e/CONJ da/P+D-F claridade/N intensíssima/ADJ-S-F das/P+D-F-P lâmpadas/N-P elétricas/ADJ-F-P ,/, como/CONJS da/P+D-F fixidez/N atenta/ADJ-F em/P que/WPRO os/CL mantinha/VB-D desde/P as/D-F-P sete/NUM e/CONJ meia/**NUM** até/P àquela/P+D-F hora/N alta/ADJ-F da/P+D-F noite/N ./.

Curvou/VB-P -/+ se/SE de/P mãos/N-P nas/P+D-F-P ilhargas/N-P ,/, de/P ímpeto/N esticou/VB-D os/D-P braços/N-P ,/, arrojou/VB-D -/+ os/CL à/P+D-F frente/ADV com/P um/D-UM ahn/**INT** !/. surdo/ADJ de/P atleta/N que/WPRO exercia/VB-D os/D-P músculos/N-P entorpecidos/VB-AN-P e/CONJ desabou/VB-D -/+ os/CL depois/ADV ,/, com/P força/N ,/, sacudindo/VB-G -/+ se/SE todo/Q ,/, virando/VB-G ,/, revirando/VB-G a/D-F cabeça/N ,/, como/CONJS em/P ânsia/N angustiosa/ADJ-F ./. Levantou/VB-D -/+ os/CL ,/, de/P novo/ADJ ,/, acima/ADV da/P+D-F cabeça/N ,/, as/D-F-P mãos/N-P juntas/ADJ-F-P ,/, estrincando/VB-G os/D-P dedos/N-P enclavinhados/VB-AN-P e/CONJ bocejou/VB-D ,/, espichando/VB-G -/+ se/SE nas/P+D-F-P pontas/N-P dos/P+D-P pés/N-P caindo/VB-G depois/ADV ,/, rijamente/ADV ,/, sobre/P os/D-P tacões/N-P ./.

Já/ADV as/D-F-P primeiras/ADJ-F-P páginas/N-P haviam/HV-D descido/VB-PP para/P a/D-F clichagem/N ./. Embaixo/**ADV** ,/, martelavam/VB-D pancadas/N-P crebas/ADJ-F-P ,/, como/CONJS de/P matracas/N-P ./. A/D-F caldeira/N reboava/VB-D num/P retroar/VB soturno/**ADJ** de/P caverna/N que/C

repercutisse/VB-SD ,/, sem/P descontinuar/VB ,/, o/D gorgorejo/N possante/ADJ-G de/P águas/N-P encachoeiradas/VB-AN-F-P ./.

Na/P+D-F sala/N da/P+D-F revisão/N ,/, estreita/ADJ-F e/CONJ abafada/VB-AN-F ,/, mal/ADV comportando/VB-G as/D-F-P quatro/NUM mesas/N-P de/P serviço/N ,/, os/D-P revisores/N-P repousavam/VB-D :/. apenas/ADV o/D Brites/NPR ,/, esgalgado/VB-AN e/CONJ míope/**ADJ** ,/, lia/VB-D o/D artigo/N de/P fundo/N ,/, todo/Q em/P períodos/N-P lamentosos/ADJ-P ,/, augurando/VB-G fome/N e/CONJ lutas/N-P ;/. e/CONJ o/D Amaro/NPR ,/, conferente/N ,/, acusando/VB-G a/D-F pontuação/N ,/, de/P quando/CONJS em/P quando/CONJS batia/VB-D na/P+D-F mesa/N pancadas/N-P secas/ADJ-F-P com/P um/D-UM lápis/N ou/CONJ dizia/VB-D claramente/ADV uma/D-UM-F palavra/N ,/, repetindo/VB-G -/+ a/CL devagar/ADV ,/, sílaba/N a/D-F sílaba/N ,/, enquanto/CONJS o/D Brites/NPR ,/, debruçado/VB-AN sobre/P a/D-F prova/N ,/, fazia/VB-D a/D-F emenda/N ,/, resmungando/VB-G ./.

NOTA: A correção automática é feita via Python que através de comando específico gera arquivo padrão *gold*: anotado e corrigido automaticamente.

APÊNDICE H – RELAÇÃO SIMPLIFICADA DE ETIQUETAS DO AELIUS, MODELO AELIUSHUNPOS, UTILIZADOS NA ANOTAÇÃO MORFOSSINTÁTICA DO *CORPUS* (BASEADO NAS ETIQUETAS DO *CORPUS* TYCHO BRAHE)

TYPE	TAG	CLASSIFICAÇÃO
VERBO em geral	VB	Infinitivo
	VB-I	Imperativo
	VB-P	Presente
	VB-SP	Presente do Subjuntivo
	VB-D	Pretérito Passado
	VB-SD	Pretérito do Subjuntivo
	VB-R	Futuro
	VB-SR	Futuro do Subjuntivo
	VB-G	Gerúndio
	VB-PP	Particípio perfeito irregular
	VB-AN	Particípio passivo
Verbo SER	SR-	Basicamente, seguem a mesma lógica de anotação dos verbos em geral.
Verbo ESTAR	ET-	Basicamente, seguem a mesma lógica de anotação dos verbos em geral.
Verbo TER	TR-	Basicamente, seguem a mesma lógica de anotação dos verbos em geral.
Verbo HAVER	HV-	Basicamente, seguem a mesma lógica de anotação dos verbos em geral.
GÊNERO	Nenhuma	Masculino
	F-	Feminino
NÚMERO	Nenhuma	Singular
	P-	Plural
NOME	N	Substantivo
NOME	NPR	Substantivo próprio
PRONOME	PRO	Pronome pessoal, forma oblíqua
PRONOME POSSESSIVO	PRO$	Seus, teus, meus
CLÍTICOS	CL	Clíticos em geral (me, te, lhe, lo...),
	SE	Clítico “se”
ARTIGOS DEFINIDOS e DEMONSTRATIVOS	D	Masculino singular
	D-F	Feminino singular
	D-P	Masculino plural
	D-F-P	Feminino plural
ARTIGOS INDEFINIDOS e NÚMERO CARDINA “UM”	D-UM	Masculino singular
	D-UM-F	Feminino singular
	D-UM-P	Masculino plural
	D-UM-F-P	Feminino plural
OUTROS DEMONSTRATIVOS	DEM	Demonstrativos invariáveis
ADJETIVOS	ADJ	Forma geral / segundo / Meio / mesmo / número ordinal
ADJETIVOS	ADJ-G	Comum de dois gêneros / demais
ADJETIVOS COMPARATIVOS	ADJ-R	Maior, menor
ADJETIVOS SUPERLATIVOS	ADJ-S	-íssima
ADVÉRBIO	ADV	Advérbio em –mente, de lugar, de tempo

ADVÉRBIOS COMPARATIVOS	ADV-R	Mais/menos/demais/bastante/melhor/pior/tão/tanto/tal/antes
ADVÉRBIOS SUPERLATIVOS	ADV-S	
ADVÉRBIOS QUANTIFICADORES	Q	muito/pouco/todo
ADVÉRBIO NEGAÇÃO	ADV-NEG	nunca
QUANTIFICADORES EM GERAL	Q	Tudo, alguém, outrem, muito, pouco, um pouco, algum, alguma, ambos, cada, qualquer
QUANTIFICADORES NEGATIVO	Q-NEG	Nada, ninguém, nenhum
CONJUNÇÃO COORDENADA	CONJ	Aditiva e Alternativa
CONJUNÇÃO SUBORDINADA	CONJS	
PREPOSIÇÃO	P	
CONTRAÇÃO	P+	
PRONOME RELATIVO	WPRO	
OUTRO	OUTRO	
INTERJEIÇÕES	INTJ	
NUMERAL CARDINAL	NUM	1, uno, 100, 153, milhão
PALAVRAS ESTRANGEIRAS	FW	
PONTUAÇÃO	.	Período ou reticências, interrogação, exclamação, ponto e vírgula, dois-pontos,
	,	Vírgula
	QT	Aspas
	(	Parênteses, travessão

Fonte: Organização própria com base no manual de etiquetas morfológicas do Tycho Brahe, disponível em http://www.tycho.iel.unicamp.br/~tycho/corpus/manual/*tags*.html

APÊNDICE I – RELAÇÃO DE GRUPOS DE PESQUISA DE LINGUÍSTICA DE CORPUS

Fonte: Diretório de Grupos de Pesquisas do Brasil/CNPQ – disponível em: dgp.cnpq.br/buscaoperacional/. Acesso em: 28 de jan. 2014.

1.	**Gr:** A história social do falar mineiro - UFJF **Li:** Patrícia Fabiane Amaral da Cunha Lacerda **AP:** Linguística
2.	**Gr:** A investigação de relações anafóricas a partir de corpus: uma abordagem integrada - UFSC **Li:** Marco Antonio Esteves da Rocha **AP:** Linguística
3.	**Gr:** A oralidade na ficção literária brasileira - PUC/SP **Li:** Dino Fioravante Preti **AP:** Letras
4.	**Gr:** COMET - Corpus Multilíngue para Ensino e Tradução - USP **Li:** Stella Esther Ortweiler Tagnin **AP:** Linguística
5.	**Gr:** Computação e Linguagem Natural - UFC **Li:** Leonel Figueiredo de Alencar Araripe **AP:** Linguística
6.	**Gr:** DIRECT- Em Direção à Linguagem do Trabalho - PUC/SP **Li:** Leila Barbara **AP:** Linguística
7.	**Gr:** ELO - Emergência da Linguagem Oral - UFPEL **Li:** Mirian Rose Brum-de-Paula **AP:** Linguística
8.	**Gr:** Estudos do léxico e da tradução-GELTra - UNESP **Li:** Lidia Almeida Barros **AP:** Linguística
9.	**Gr:** Estudos em Linguística de Corpus - UFU **Li:** Guilherme Fromm **AP:** Linguística
10.	**Gr:** Estudos em Língua Estrangeira - UNINOVE **Li:** Luciana Latarini Ginezi **AP:** Letras
11.	**Gr:** FLEC - Formalismos Linguísticos e Computação - PUCRS **Li:** Ana Maria Tramunt Ibaños **AP:** Linguística
12.	**Gr:** FONOLOGIA COGNITIVA: Investigação de Padrões Sonoros Emergentes - UFMG **Li:** Thaïs Cristófaro Alves da Silva **AP:** Linguística
13.	**Gr:** GELCORP-SUL- Grupo de Estudos em Linguística de Corpus do Sul - UFRGS **Li:** Maria Jose Bocorny Finatto **AP:** Linguística
14.	**Gr:** GELVI - Grupo de Estudos Interdisciplinares sobre Linguagem e Vulnerabilidade Infantil - PUC Minas **Li:** Pedro Perini Frizzera da Mota Santos **AP:** Linguística
15.	**Gr:** GRUPO DE PESQUISAS E ESTUDOS EM ANÁLISE DE DISCURSO CRÍTICA E LINGUÍSTICA SISTÊMICO-FUNCIONAL - UFU **Li:** Maria Aparecida Resende Ottoni

	AP: Linguística
16.	**Gr:** Grupo Interdisciplinar de Pesquisas em Linguística Informática - USP **Li:** Zilda Maria Zapparoli **AP:** Linguística
17.	**Gr:** Grupo Multidisciplinar de Investigações Linguísticas - UFRRJ **Li:** Fernando Vieira Peixoto Filho **AP:** Linguística
18.	**Gr:** Grupo de Estudos Interdisciplinares em Tecnologia, Linguagens e Educação - IFRS **Li:** Viviane Diehl **AP:** Educação
19.	**Gr:** Grupo de Estudos Variacionistas - GEVAR - UFTM **Li:** Juliana Bertucci Barbosa **AP:** Linguística
20.	**Gr:** Grupo de Estudos da Tradução - GET - UERN **Li:** Edilene Rodrigues Barbosa **AP:** Linguística
21.	**Gr:** Grupo de Estudos de Linguística de Corpus - GELC - PUC/SP **Li:** Antonio Paulo Berber Sardinha **AP:** Linguística
22.	**Gr:** Grupo de Pesquisa Estudos do Léxico: descrição e ensino - UNESP **Li:** Odair Luiz Nadin da Silva **AP:** Linguística
23.	**Gr:** Grupo de Pesquisa em Gestão do Conhecimento, Processamento de Linguagens Especializadas e Tecnologia - ESPM **Li:** Ana Eliza Pereira Bocorny **AP:** Linguística
24.	**Gr:** Grupo de Pesquisas em Ciências Aeronáuticas - PUCRS **Li:** Éder Henriqson **AP:** Engenharia Aeroespacial
25.	**Gr:** História Social do Português - UFMG **Li:** Jania Martins Ramos **AP:** Linguística
26.	**Gr:** História, memória e literatura bíblica - PUC-Rio **Li:** Maria de Lourdes Corrêa Lima **AP:** História
27.	**Gr:** Inteligência Computacional - USP **Li:** Maria Carolina Monard **AP:** Ciência da Computação
28.	**Gr:** Interfaces linguagem, cognição e cultura - INCOGNITO - UFMG **Li:** Heliana Ribeiro de Mello **AP:** Linguística
29.	**Gr:** LINGUÍSTICA DINÂMICA - UFPE **Li:** Antonio Torre Medina **AP:** Linguística
30.	**Gr:** Linguística comparativa Português-Espanhol. Estudo da conexão pragmática/sintaxe - UFF **Li:** Paulo Antonio Pinheiro Correa **AP:** Linguística
31.	**Gr:** Linguística Computacional, Língua e Literatura - UNESP **Li:** Maurizio Babini **AP:** Linguística
32.	**Gr:** Linguística sistêmico-funcional, linguística de corpus e análise do discurso - PUC-Rio **Li:** Lúcia Pacheco de Oliveira **AP:** Linguística
33.	**Gr:** NILC - Núcleo Interinstitucional de Linguística Computacional - USP

	Li: Maria das Graças Volpe Nunes **AP:** Ciência da Computação
34.	**Gr:** Padrões rítmicos, fixação de parâmetros e mudança linguística - UNICAMP **Li:** Charlotte Marie Chambelland Galves **AP:** Linguística
35.	**Gr:** Práticas discursivas em contextos pedagógicos - PUC-Rio **Li:** Lúcia Pacheco de Oliveira **AP:** Linguística
36.	**Gr:** SEMANTEC - UNISINOS **Li:** Rove Luiza de Oliveira Chishman **AP:** Linguística
37.	**Gr:** TEL - Tecnologias nos Estudos da Linguagem - UNESP **Li:** Ana Mariza Benedetti **AP:** Linguística
38.	**Gr:** TERMISUL - Projeto Terminológico Cone Sul - UFRGS **Li:** Cleci Regina Bevilacqua **AP:** Linguística
39.	**Gr:** TRADUÇÃO, TERMINOLOGIA E CORPORA - UNESP **Li:** Diva Cardoso de Camargo **AP:** Linguística
40.	**Gr:** VARIUS - Variação e uso - UFRRJ **Li:** Viviane Conceição Antunes Lima **AP:** Letras
41.	**Gr:** Vozes da Amazônia - UFPA **Li:** Regina Celia Fernandes Cruz **AP:** Linguística

APÊNDICE J – RELAÇÃO DE GRUPOS DE PESQUISA EM LINGUÍSTICA COMPUTACIONAL

Fonte: Diretório de Grupos de Pesquisas do Brasil/CNPQ – disponível em: dgp.cnpq.br/buscaoperacional/. Acesso em: 28 de jan. 2014.

1.	**Gr:** A investigação de relações anafóricas a partir de corpus: uma abordagem integrada - UFSC **Li:** Marco Antonio Esteves da Rocha **AP:** Linguística
2.	**Gr:** Computação Científica na Pesquisa Agropecuária - CC-Agro - EMBRAPA **Li:** Sônia Ternes **AP:** Ciência da Computação
3.	**Gr:** Computação e Linguagem Natural - UFC **Li:** Leonel Figueiredo de Alencar Araripe **AP:** Linguística
4.	**Gr:** Educação a Distância - UNICAMP **Li:** Heloisa Vieira da Rocha **AP:** Ciência da Computação
5.	**Gr:** Estudos do léxico e da tradução-GELTra - UNESP **Li:** Lidia Almeida Barros **AP:** Linguística
6.	**Gr:** Estudos em Linguística Geral e Aplicada - UEAP **Li:** Fabio Xavier da Silva Araújo **AP:** Linguística
7.	**Gr:** FRAGMENTAÇÃO E UNIDADE EM MANIFESTAÇÕES ARTÍSTICAS E SUAS IMPLICAÇÕES URBANAS - UFES **Li:** Ernesto de Souza Pachito **AP:** Artes
8.	**Gr:** Ferramentas em Linguística Computacional - PUC-Rio **Li:** Violeta de San Tiago Dantas Barbosa Quental **AP:** Linguística
9.	**Gr:** GIA - Grupo de Pesquisa em Inteligência Artificial - UFFS **Li:** Ilson Wilmar Rodrigues Filho **AP:** Ciência da Computação
10.	**Gr:** GRAMÁTICA & COGNIÇÃO - UFJF **Li:** Maria Margarida Martins Salomão **AP:** Linguística
11.	**Gr:** Grupo Interdisciplinar de Mineração de Dados e Aplicações (GIMDA) - UFABC **Li:** Ronaldo Cristiano Prati **AP:** Ciência da Computação
12.	**Gr:** Grupo Multidisciplinar de Investigações Linguísticas - UFRRJ **Li:** Fernando Vieira Peixoto Filho **AP:** Linguística
13.	**Gr:** Grupo de Computação Inteligente e Autônoma - UFAM **Li:** Edjard de Souza Mota **AP:** Ciência da Computação
14.	**Gr:** Grupo de Estudos Interdisciplinares em Tecnologia, Linguagens e Educação - IFRS **Li:** Viviane Diehl **AP:** Educação
15.	**Gr:** Grupo de Inteligência Aplicada - UNIOESTE **Li:** Clodis Boscarioli **AP:** Ciência da Computação

16.	**Gr:** Grupo de Inteligência Artificial - USP **Li:** José de Jesús Pérez Alcázar **AP:** Ciência da Computação
17.	**Gr:** Grupo de Inteligência Computacional - UFSCAR **Li:** Heloisa de Arruda Camargo **AP:** Ciência da Computação
18.	**Gr:** Grupo de Pesquisa Multidisciplinar em Léxico e Dicionários - GLexD - UNESP **Li:** Claudia Zavaglia **AP:** Linguística
19.	**Gr:** Grupo de Pesquisa e Estudo de Linguística e Língua Portuguesa (GPELLP) - UFTM **Li:** Jauranice Rodrigues Cavalcanti **AP:** Linguística
20.	**Gr:** Grupo de Pesquisa em Gestão do Conhecimento, Processamento de Linguagens Especializadas e Tecnologia - ESPM **Li:** Ana Eliza Pereira Bocorny **AP:** Linguística
21.	**Gr:** Idealize: Inteligência Computacional Aplicada - UFOP **Li:** Alvaro Rodrigues Pereira Junior **AP:** Ciência da Computação
22.	**Gr:** Inteligência Computacional - USP **Li:** Maria Carolina Monard **AP:** Ciência da Computação
23.	**Gr:** LEA: Laboratório de Estudos Avançados em Computação - UNIPAMPA **Li:** Fabio Natanael Kepler **AP:** Ciência da Computação
24.	**Gr:** LaLiC: Laboratório de Linguística Computacional - UFSCAR **Li:** Helena de Medeiros Caseli **AP:** Ciência da Computação
25.	**Gr:** Laboratório de Inteligência Social e Sistemas Complexos - UFABC **Li:** Maria das Graças Bruno Marietto **AP:** Ciência da Computação
26.	**Gr:** Linguística Computacional - UFMS **Li:** Gedson Faria **AP:** Linguística
27.	**Gr:** Linguística Computacional, Língua e Literatura - UNESP **Li:** Maurizio Babini **AP:** Linguística
28.	**Gr:** Logprob: Lógica e Probabilidade - USP **Li:** Marcelo Finger **AP:** Ciência da Computação
29.	**Gr:** Léxico, Gramática e Terminologia - UFSCAR **Li:** Oto Araújo Vale **AP:** Linguística
30.	**Gr:** Lógica, Inteligência Artificial e Métodos Formais - USP **Li:** Renata Wassermann **AP:** Ciência da Computação
31.	**Gr:** NILC - Núcleo Interinstitucional de Linguística Computacional - USP **Li:** Maria das Graças Volpe Nunes **AP:** Ciência da Computação
32.	**Gr:** Núcleo de Ciência Cognitiva - USP **Li:** Marcio Lobo Netto **AP:** Ciência da Computação
33.	**Gr:** PIIM - Pesquisas Interdisciplinares em Informação Multimídia - CEFET/MG **Li:** Flávio Luis Cardeal Pádua

	AP: Ciência da Computação
34.	**Gr:** R.E.G.I.I.M.E.N.T.O., Arquitetura da Informação, Linguística Computacional e Multimodalidade, Mídias e Interatividade - UNB **Li:** Claudio Gottschalg Duque **AP:** Ciência da Informação
35.	**Gr:** ROBÓTICA INTELIGENTE E VISÃO ARTIFICIAL - UFRGS **Li:** Dante Augusto Couto Barone **AP:** Ciência da Computação
36.	**Gr:** SEMANTEC - UNISINOS **Li:** Rove Luiza de Oliveira Chishman **AP:** Linguística
37.	**Gr:** Sistema Interativos Inteligentes - UEM **Li:** Sérgio Roberto Pereira da Silva **AP:** Ciência da Computação
38.	**Gr:** Sistemas Inteligentes - UNIFOR **Li:** Joao Jose Vasco Peixoto Furtado **AP:** Ciência da Computação
39.	**Gr:** Teoria da Gramática e o Português Brasileiro - UFSC **Li:** Sandra Quarezemin **AP:** Linguística
40.	**Gr:** Texto Livre: Semiótica e Tecnologia - UFMG **Li:** Ana Cristina Fricke Matte **AP:** Linguística
41.	**Gr:** VARIUS - Variação e uso - UFRRJ **Li:** Viviane Conceição Antunes Lima **AP:** Letras

APÊNDICE K – LISTA DE *TAGS* INCORRETAMENTE ANOTADAS PELO ETIQUETADOR, OBTIDAS APÓS CORREÇÃO MANUAL (SÍNTESE).

TOKEN	ARQUIVO	TAG errada Anotação automática	TAG correta Anotação manual
dolorido/	Conq	ADJ	VB-AN
feminino/	Conq	ADJ	N
velho/	Conq	ADJ	N
vazio/	Turb	ADJ	N
elétrico/	Turb	ADJ	N
pequeno/	Turb	ADJ	N
amarelo/	Ent	ADJ	N
alto/	Ent	ADJ	N
demorado/	Ent	ADJ	VB-AN
alto/	Ent	ADJ	N
fazendeiro/	Cor5	ADJ	N
futuro/	Cor5	ADJ	N
preguiçoso/	Cor5	ADJ	N
superior/	Cor5	ADJ	N
amarelo/	Cor6	ADJ	N
cômodo/	Cor6	ADJ	N
pequeno/	Cor6	ADJ	N
alto/	Cor6	ADJ	N
octogenário/	Cor6	ADJ	N
cômodo/	Cor6	ADJ	N
estiva/	Ent	ADJ-F	N
lavadeira/	Ent	ADJ-F	N-F
sexta/	Ent	ADJ-F	NPR
bárbara/	Ent	ADJ-F	N
vizinha/	Cor5	ADJ-F	N
cômoda/	Cor5	ADJ-F	N
outra/	Cor5	ADJ-F	OUTRO-F
Marciana/	Cor5	ADJ-F	NPR

Nova/	Cor5	ADJ-F	NPR
meia/	Cor6	ADJ-F	N
Baiana/	Cor6	ADJ-F	NPR
meia/	Cor6	ADJ-F	N
direita/	Cor6	ADJ-F	N
esquerda/	Cor6	ADJ-F	N
bica/	Cor6	ADJ-F	N
mesmo/	Conq	ADJFP	FP
placas/	Conq	ADJ-F-P	N-P
modinhas/	Conq	ADJ-F-P	N-P
neblinas/	Cor5	ADJ-F-P	N-P
harmônicas/	Cor6	ADJ-F-P	N-P
brasileiras/	Cor6	ADJ-F-P	N-F-P
companheiras/	Cor6	ADJ-F-P	N-F-P
mirante/	Conq	ADJ-G	N
só/	Conq	ADJ-G	FP
dobre/	Ent	ADJ-G	N
pessoal/	Cor5	ADJ-G	N
simples/	Cor5	ADJ-G	N
espirais/	Ent	ADJ-G-P	N-P
Velhos/	Conq	ADJ-P	N-P
Borrifos/	Conq	ADJ-P	N-P
outros/	Turb	ADJ-P	OUTRO-P
rústicos/	Ent	ADJ-P	N-P
primeiros/	Cor5	ADJ-P	N-P
operários/	Cor6	ADJ-P	N-P
tanto/	Cor6	ADJ-R	ADV-R
mesmas/	Turb	ADJ-R-F-P	ADJ-F-P
mor/	Cor6	ADJ-R-G	N
fora/	Conq	ADV	SR-RA
querosene/	Cor5	ADV	N
fora/	Cor6	ADV	VB-RA
chique/	Cor6	ADV	ADJ-G

o/	Cor5	CL	D
se/	Cor6	CONJS	WQ
a/	Conq	D-F	P
A/	Ent	D-F	P
até/	Cor5	FP	P
sóbrio/	Conq	N	ADJ
discutia/	Conq	N	VB-D
Luiz/	Conq	N	NPR
carola/	Conq	N	ADJ
estética/	Conq	N	ADJ
morno/	Conq	N	ADJ
sinistro/	Conq	N	ADJ
direito/	Turb	N	ADV
Rócio/	Turb	N	NPR
surgia/	Ent	N	VB-D
soalheira/	Ent	N	ADJ-G
banzeiro/	Ent	N	ADJ
mugia/	Ent	N	VB-D
brandia/	Ent	N	VB-D
Ahu/	Ent	N	INTJ
outro/	Ent	N	OUTRO
cuidado/	Ent	N	VB-AN
Sucumbira/	Ent	N	VB-RA
Súbito/	Ent	N	ADJ
Machona/	Cor5	N	NPR
Farjão/	Cor5	N	NPR
Machucas/	Cor5	N	NPR
medida/	Cor5	N	VB-AN-F
penteado/	Cor6	N	VB-AN
Leonor/	Cor6	N	NPR
reza/	Cor6	N	VB-P
cerimoniosa/	Cor6	N	ADJ-F
carnaval/	Cor6	N	NPR

Libânia/	Cor6	N	NPR
lutulentas/	Conq	N-P	ADJ-F-P
curiosos/	Conq	N-P	ADJ-P
Lins/	Conq	N-P	NPR
ameaçadores/	Conq	N-P	ADJ-P
sertanejas/	Conq	N-P	ADJ-F-P
cais/	Turb	N-P	N
sertanejos/	Ent	N-P	ADJ-P
moribundas/	Ent	N-P	ADJ-F-P
bentas/	Cor5	N-P	ADJ-F-P
Maricas/	Cor5	N-P	NPR
ourives/	Cor5	N-P	N
réis/	Cor5	N-P	N
homicidas/	Cor5	N-P	ADJ-G-P
metes/	Cor6	N-P	VB-P
casas/	Cor6	N-P	VB-P
coitados/	Cor6	N-P	ADJ-P
Dores/	Cor6	N-P	NPR
alípedes/	Conq	N-P	ADJ
barítono/	Conq	NPR	N
mestre/	Conq	NPR	N
casarão/	Conq	NPR	N
Purpúreo/	Conq	NPR	ADJ
barítono/	Conq	*NPR*	*ADJ*
rútila/	Conq	NPR	ADJ-F
merencório/	Conq	NPR	ADJ
hálito/	Turb	NPR	N
mercúrio/	Ent	NPR	N
pagã/	Ent	NPR	ADJ-F
campanário/	Ent	NPR	N
Réquiem/	Ent	NPR	N
alcobaça/	Cor5	NPR	N
capilé/	Cor6	NPR	N

Olé/	Cor6	NPR	INTJ
manjericão/	Cor6	NPR	N
Cadê/	Cor6	NPR	ADV
Êh/	Cor6	NPR	INTJ
violão/	Cor6	NPR	N
Dores/	Cor6	NPR-P	NPR
a/	Conq	P	D
a/	Conq	P	D-F
Num/	Conq	P	P+D-UM
até/	Cor5	P	FP
a/	Conq	PD-F	D-F
Cantu/	Conq	PRO	NPR
todo/	Cor5	Q	N
nada/	Conq	Q-NEG	N
se/	Conq	SE	CONJS
olhar/	Turb	VB	N
crochê/	Cor6	VB	N
samburá/	Cor6	VB	N
olhar/	Cor6	VB	N
Toledo/	Conq	VB-AN	NPR
bêbado/	Conq	VB-AN	ADJ
trocado/	Conq	VB-AN	VB-PP
Obrigado/	Conq	VB-AN	ADJ
desconhecido/	Conq	VB-AN	N
cercado/	Ent	VB-AN	N
ordenado/	Cor5	VB-AN	N
empregado/	Cor5	VB-AN	N
achado/	Cor5	VB-AN	N
parado/	Cor6	VB-AN	VB-PP
mórbida/	Conq	VB-AN-F	ADJ-F
seguida/	Ent	VB-AN-F	N
morta/	Ent	VB-AN-F	N
lentas/	Conq	VB-AN-F-P	ADJ-F-P

úmidas/	Conq	VB-AN-F-P	ADJ-F-P
pernadas/	Conq	VB-AN-F-P	N-P
ávidas/	Cor6	VB-AN-F-P	ADJ-F-P
talhadas/	Cor6	VB-AN-F-P	N-F-P
petits/	Conq	VB-D	FW
Lou/	Ent	VB-D	INTJ
ehou/	Ent	VB-D	INTJ
Saciado/	Turb	VB-G	VB-AN
carrancudo/	Turb	VB-G	ADJ
encarquilhado/	Ent	VB-G	VB-AN
encurralado/	Cor5	VB-G	VB-NA
malgrado/	Cor5	VB-G	P
Teçai/	Ent	VB-I	NPR
exalavam/	Conq	VB-P	VB-D
toma/	Conq	VB-P	VB-I
Curvou/	Turb	VB-P	VB-D
troco/	Turb	VB-P	N
peco/	Ent	VB-P	ADJ
Eh/	Ent	VB-P	INTJ
paga/	Cor5	VB-P	VB-AN-F
enviara/	Cor5	VB-P	VB-RA
Irrequieta/	Cor6	VB-P	ADJ-F
mostra/	Cor6	VB-P	N
mudo/	Cor6	VB-P	ADJ
satisfeito/	Conq	VB-PP	VB-AN
chorado/	Cor6	VB-PP	N
ferrão/	Ent	VB-R	N
desse/	Conq	VB-SD	P+D
tomes/	Conq	VB-SP	VB-I
tênue/	Conq	VB-SP	ADJ-G
ouça/	Turb	VB-SP	VB-I
fales/	Cor6	VB-SP	VB-I
arribaste/	Cor6	VB-SP	VB-D

faça/	Cor6	VB-SP	VB-I
Diga/	Cor6	VB-SP	VB-I
como/	Conq	WADV	CONJS
que/	Conq	WPRO	C
se/	Cor6	WQ	SE

LEGENDA:

Conq = A Conquista

Turb = Turbilhão

Ent = O enterro

Cor5 = O cortiço cap. 5

Cor6 = O cortiço cap. 6

APÊNDICE L – SCRIPT CORRIGEPONTUAÇÃO.SH

```
#! /bin/bash
# CorrigePontuacao.sh

UPPER=ÁÉÍÓÚÀÈÌÒÙÂÊÔÃÕÜÇ
        export UPPER
LOWER=áéíóúàèìòùâêôãõüç
        export LOWER
L=[:alpha:]
        export L
W=[-${LOWER}${UPPER}$L]+
        export W
for i in $*
do
set $(echo "$i" | awk 'BEGIN {FS= "."} { print $1}')

# Correção de casos como . .. para ...
sed -E "s@(\.)[[:space:]]+(\.)@\1\2@g" "$i" | \
sed -E "s@(\.)[[:space:]]+(\.)@\1\2@g" | \

# Uniformização dos travessões para Unicode 2013 EN DASH
sed -E "s@–|—|—|・@–@g" | \

# Uniformização das aspas para "
sed -E "s@“|”|„|‘|’|«|»@\"@g" | \

# Uniformização de reticências para ...
sed -E "s@…@...@g" | \

# Substituição de hífens por travessão nos casos apropriados
sed -E "s@([[:space:]]+)\-@\1–@g" | \
sed -E "s@^\-@– @g" | \

# Eliminação de pontos indevidos antes de vírgula;
# esse comando pode eliminar o ponto de abreviaturas antes de vírgula!
sed -E "s@\.,@,@g" | \

# Eliminação de reticências com 4 pontos
sed "s@\.\.\.\.@...@g" | \
# esse comando pode eliminar o ponto de abreviaturas

# Substituição de quebra de linha por espaço em branco
# tr -s "\n" " " | \

# Eliminação de hífens no final de palavras
sed -E "s@\-[[:space:]]+@-@g" | \

# Inserção de ponto após reticências que terminam sentenças
# esse ponto deverá ser excluído da versão etiquetada

# Eliminação de ponto em casos como palavra. palavra
sed -E "s@($W)\.([[:space:]]+[a-z$LOWER]+)@\1\2@g" > "$1".pdt.txt
```