# COMPILAÇÃO E ANOTAÇÃO DO *CORPUS* COELHO NETTO: LUZES NO ENSINO E PESQUISAS LITERÁRIAS POR MEIO DA LINGUÍSTICA DE *CORPUS*

Francimary Macêdo Martins

Doutora em Linguística

*Universidade Federal do Maranhão*, fm.martins@ufma.br

**RESUMO**:

Este trata da compilação e anotação morfossintática do *Corpus* Coelho Netto, um *corpus* de textos literários dos séc. XIX e XX, dos romances *A Conquista* e *Turbilhão* e contos de *Sertão*, do escritor Coelho Netto. A pesquisa está na interface da Linguística Computacional e da Linguística de *Corpus* (BERBER SARDINHA, 2000, 2004, 2005a; ALUÍSIO; ALMEIDA, 2006; ALENCAR, 2010, 2013a, 2013b). A compilação consistiu nas etapas de seleção, coleta de textos e manipulação; nesta são realizadas a limpeza, edição e atualização dos textos; e a Anotação Morfossintática do *Corpus*, que foi realizada pelo etiquetador automático Aelius, modelo AeliusHunPos, um *software* livre em Python que utiliza a *Natural Language Toolkit* – NLTK. O *Corpus* está preparado para ser submetido a análises linguístico-computacionais, envolvendo o campo do ensino e estudo em áreas diversas. O *Corpus* está em processo de ajustes de adequação para envio e publicação no Repositório de Recursos da Linguateca-PT.

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

## 1. INTRODUÇÃO

Santos (1999) e Berber Sardinha (2005a) entendem 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, e o computador é sem dúvida uma ferramenta indispensável no mundo moderno, 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.

O processamento computacional da linguagem está vinculado à relação estreita entre a Linguística Aplicada (LA) com as Tecnologias de Informação e Comunicação (TICs). E a convergência das áreas de Engenharia de Computação, Engenharia Elétrica, Inteligência Artificial, Literatura e Linguística inauguram um cenário senão diferente, mas inovador em todas essas áreas, abrindo frente para novas linhas de pesquisa.

Nessa perspectiva computacional, a Linguística Computacional e a Linguística de *Corpus* (LCLC) vêm contribuindo significativamente para novas pesquisas geradas do cruzamento entre essas áreas, pois a 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. A compilação de um *corpus* literário a ser apresentado tende a enriquecer *corpora* desse gênero, como: *O Corpus* Histórico do Português Tycho Brahe (CHPTB), CORPTEXLIT, Linguateca, Floresta Sintática, *Corpus* de Araraquara, Lácio-*WEB*/NILC, CE-DOHS dentre outros (BERBER SARDINHA, 2000, 2004; ALUÍSIO; ALMEIDA, 2006).

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 confiança na análise manual. Essa influência positiva é evidenciada por ZYNGIER (2011, p. 102) ao destacar que a Linguística de *Corpus* "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".

Porquanto, a compilação e anotação do *Corpus* Coelho Netto e sua disponibilização digital para a comunidade científica constitui-se uma colaboração primordial para que os textos deste escritor possam ser explorados em análises linguístico-literárias com viés computacional, permitindo 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.

## 2. A LINGUÍSTICA COMPUTACIONAL E A LINGUÍSTICA DE CORPUS

A Linguística Computacional[1] (LCOMP), até meados de 1960, 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, sintaxe, semântica, pragmática, discurso, texto, aquisição de linguagem, entre outros (DIAS-DA-SILVA, 2006). Conforme Othero e Menuzzi (2005), a Linguística Computacional subdivide-se em duas subáreas: a Linguística de *Corpus* e o Processamento de Linguagem Natural (PLN). Por vezes, a LCOMP é considerada sinônima do PLN (BERBER SARDINHA, 2005a, DI FELIPPO; DIAS-DA-SILVA, 2009).

A Linguística de *Corpus* (LCORP) 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

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

variados tipos: *corpora* de linguagem falada, de linguagem literária, com textos de jornal, 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).

As possibilidades de análise em um *corpus* são diversas. 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).

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.

Ao se constituir o *Corpus* Coelho Netto, a pretensão é disponibilizar um *corpus* computadorizado e anotado morfossintaticamente permitindo variadas pesquisas sobre o fazer literário.

## 3. O CORPUS COELHO NETTO (CCN)

A escolha de textos literários para a compilação do *Corpus*, 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).

O CCN é caracterizado como um "*corpus*-amostra" porque que se trata da compilação somente de capítulos de dois romances e três contos do acervo da obra de Coelho Netto. 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. De modo geral, sempre quando se trata do gênero literário, o *corpus* é constituído por amostragens dos títulos selecionados (BACELAR DO NASCIMENTO, 2002).

O *Corpus* compreende textos de 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.* Contém 53.080 *tokens* (compreendem palavras e pontuações)

## 3.1 COMPILAÇÃO E ANOTAÇÃO MORFOSSINTÁTICA

Para proceder à compilação e anotação do *Corpus* Coelho Netto utilizamos o processador Python, na versão 2.7.3 de 2012 compatível com a versão 2.0.1rc1 do *Natural Language Toolkit (NLTK)* que é um conjunto (kit) de ferramentas utilizado na construção de programas de Python.

O etiquetador automático utilizado foi o Aelius, modelo AeliusHunPos, treinado no *Corpus* do Português Histórico Tycho Brahe, que utilizou textos literários para seus testes e demonstrou uma acurácia superior aos outros modelos (ALENCAR, 2010, 2013a). A versão do Aelius é a 0.9.7, atualizada em 25/02/2013[2] pelo Prof. Dr. Leonel Alencar da UFC (ALENCAR, 2013b), e faz parte do projeto *Aelius Brazilian Portuguese POS-Tagger*[3] (ALENCAR, 2013b), registrado no SourceForge.net[4], adotando uma abordagem híbrida, que se configura em mesclar as abordagens baseadas em regras e em *n-gramas*.Tem como tarefa "identificar as categorias gramaticais das palavras em sentenças" (DOMINGUES; FAVERO; MEDEIROS, 2008, p. 269) através da etiquetagem morfossintática ou anotação morfossintática.

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).

A compilação do CCN consistiu nas etapas de **Seleção e coleta de textos e Limpeza, geração e nomeação de arquivos.** Todos os arquivos de textos devem ser convertidos para o formato "*.txt*". 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

[2] AELIUS, versão 0.9.7 - http://sourceforge.net/projects/aelius/files/Aelius-February-25-2013.zip/download
[3] AELIUS BRAZILIAN PORTUGUESE POS-TAGGER - http://sourceforge.net/projects/aelius/files/
[4] SOURCEFOGE.NET - Maior hospedagem mundial de *software* de código aberto. http://sourceforge.net/

atualização). O processo de **edição dos textos** tem, sobretudo, a finalidade de conferir à edição a responsabilidade pela máxima fidelidade em relação aos textos originais. Paixão de Sousa, Kleper e Faria (2012) destacam a garantia de fidelidade às formas originais dos textos, no sentido de produzir *corpora* idôneos: isso reforça a importância da edição dos textos. A **atualização** requerida para estes textos é em relação à ortografia vigente à época da publicação dos livros, no período entre 1862 a 1906, a maioria pertencente ao séc. XIX.

## 4. CONCLUSÃO

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.

A compilação do CCN serve como produto científico-tecnológico para fins linguísticos, corroborando as pesquisas dos linguistas de *corpus*, especialmente no campo literário. 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. Além disso, o CCN pretende “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” (ALENCAR, 2013c).

Atualmente, o *Corpus* está em processo de ajustes de adequação para publicação no Repositório de Recursos da Linguateca, ficando disponível para acesso livre no link http://www.linguateca.pt/Repositorio/, com outros *corpora* já existentes.

## REFERÊNCIAS

ALENCAR, L. F. 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 2010. Disponível em: < https://goo.gl/iZ3LF6>. Acesso em: 26 set. 2016.

_____. 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.

ALUÍSIO, S. M.; ALMEIDA, G. 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: < https://goo.gl/pZ0Ofm>. Acesso em: 26 set. 2016.

BACELAR DO NASCIMENTO, Mª F. **O lugar do corpus na investigação linguística**. Centro de Linguística da Universidade de Lisboa, 2002. Disponível em: < https://goo.gl/aviR9x>. Acesso em: 27 jul. 2016.

BERBER SARDINHA, T. **Linguística de *Corpus***. São Paulo, Manole, 2004. Disponível em: <https://goo.gl/Jklq2J>. Acesso em: 26 set. 2016.

_____. Linguística de *Corpus*: histórico e problemática. **DELTA [online]**. 2000, vol. 16, n. 2, pp. 323-367. Disponível em < https://goo.gl/6Rgwpr>. Acesso em 26 set. 2016.

_____. Trazendo a língua portuguesa para o computador. In: BERBER SARDINHA, T. (org.). **A língua portuguesa no computador**. Campinas, SP: Mercado de Letras, 2005a. p.269-295.

CÚRCIO, V. 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, A.; DIAS-DA-SILVA, B. 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: <https://goo.gl/ePcu7x>. Acesso em 26 set. 2016.

DIAS-DA-SILVA, B. C. O estudo linguístico-computacional da linguagem. **Letras de Hoje**, Porto Alegre. v. 41(2): 103-138, junho 2006. Disponível em: <https://goo.gl/HyzXSS>. Acesso em 26 set. 2016.

DOMINGUES, M. L.; FAVERO, E. L.; MEDEIROS, I. 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.

GALVES, C. M. C.; BRITTO, H. **A Construção do *Corpus* Anotado do Português Histórico Tycho Brahe**: o sistema de anotação morfológica. Campinas: UNICAMP, 1999. Disponível em: <https://goo.gl/jcfyIH>. Acesso em: 26 set. 2016.

OTHERO, G. de A.; MARTINS, R. 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.

OTHERO, G. de A.; MENUZZI, S. de M**. Linguística Computacional**: teoria & prática. São Paulo: Parábola, 2005.

PAIXÃO DE SOUSA, Mª C.; KLEPER, F. N.; FARIA, P. P. F. de. E-dictor: novas perspectivas na codificação e edição de *corpora* de textos históricos. *In*: SHEPHERD, T. M. G.; BERBER SARDINHA, T.; PINTO, M. V. (org). **Caminhos da linguística de *corpus***. Campinas, SP: Mercados das Letras, 2012. p. 191-223 (Série Espaços da Linguística de *Corpus*).

SANTOS, D. **Processamento computacional da língua portuguesa**: documento de trabalho. Lisboa: Ministério da Ciência e Tecnologia, 1999. Disponível em: <https://goo.gl/KqUzTA>. Acesso em: 26 set. 2016. Não paginado.

SARMENTO, S. **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: <https://goo.gl/0ixvf0>. Acesso em: 26 set. 2016.

SILVEIRA, F. P. da. **Integração de ferramentas para compilação e exploração de *corpora***. Dissertação (Mestrado em Ciência da Computação). Fac. de Informática, PUCRS, 2008. Disponível em: <https://goo.gl/sZM3QK>. Acesso em: 26 set. 2016.

ZYNGIER, S. 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://goo.gl/UMem7H>. Acesso em: 26 set. 2016.

ZYNGIER, S.; VIANA, V.; SILVEIRA, N. G. Discurso literário e linguística de *corpus*: uma visão empírica. **Cadernos de Letras (UFRJ)** n.28 – jul. 2011. Disponível em: <https://goo.gl/M20UH0 >. Acesso em: 26 set. 2016.