ISSN 1415 - 8973

De manuscritos históricos a corpora anotados: do Documento Digital Texto (DDT) ao corpus anotado

From historical manuscripts to annotated corpora: from the Digital Document Text (DDT) to annotated corpus

Cristiane Namiuti*
Universidade Estadual do Sudoeste da Bahia, Vitória da Conquista, Bahia, Brasil

Jorge Viana Santos**
Universidade Estadual do Sudoeste da Bahia, Vitória da Conquista, Bahia, Brasil

Resumo: Este artigo apresenta a etapa de Compilação do método LAPELINC (Laboratório de Pesquisa em Linguística de Corpus), o qual envolve a integração de três objetos fundamentais na construção de corpora digitais: Documento Físico (DF); Documento Digital Imagem (DDI); Documento Digital Texto (DDT). Tal etapa é responsável por alterar a codificação do digital de visual-imagética para verbal-textual, passagem que envolve transcrição. Nos limites deste trabalho, consideramos *as limitações e as possibilidades dos suportes materiais da escrita com destaque para* as complexidades do suporte digital.

Palavras-chave: Linguística de Corpus. Escrita. Suporte Digital. Documentos Históricos. Método LAPELINC.

Abstract: This paper shows the Compilation stage of LAPELINC (in english, Corpus Linguistics Research Laboratory) method, that integrates three fundamental objects to built digital corpora: physical document (in Portuguese, DF); digital document image (in Portuguese, DDI); digital document text (in Portuguese, DDT). This stage is responsible for change the digital codification from visual-image to verbal-text, this fact uses transcription. Here we consider the limits and possibilities of digital medium for recording writting.

Keywords: Corpus Linguistics. Writting. Digital. Historical documents. LAPELINC Method.

1 INTRODUÇÃO

Como vem sendo salientado por Paixão de Sousa (2006), Namiuti, Santos e Leite (2011), Santos e Namiuti (2016a), dentre outros, a reprodução de documentos históricos para a pesquisa científica necessita da garantia de fidedignidade entre o documento físico e sua versão digital. Santos (2010a; 2010b), Santos e Brito (2014), Brito (2015), buscando soluções para o problema relativo à fidedignidade para o corpus de Documentos Oitocentistas de Vitória da Conquista e região (DOViC), vem desenvolvendo e aplicando uma forma de transposição de documentos manuscritos históricos para o meio digital através da Fotografia cientificamente controlada, hoje integrada no Método LAPELINC[1]. Neste método, por almejar a construção e disponibilização de corpora eletrônicos anotados, o pressuposto fundamental da fidedignidade

* Professor do Departamento de Estudos Linguísticos e Literários (DELL) e do Programa de Pós-Graduação em Linguística (PPGLIN), coordenador, juntamente com a Professor Dr. Jorge Viana Santos, do Laboratório de Pesquisa em Linguística de *Corpus* (LAPELINC), da Universidade Estadual do Sudoeste da Bahia (UESB), Vitória da Conquista, Bahia, Brasil. Email: cristianenamiuti@uesb.edu.br.
** Professor do Departamento de Estudos Linguísticos e Literários (DELL) e do Programa de Pós-Graduação em Linguística (PPGLIN), coordenador, juntamente com a Professora Drª Cristiane Namiuti, do Laboratório de Pesquisa em Linguística de *Corpus* (LAPELINC), da Universidade Estadual do Sudoeste da Bahia (UESB), Vitória da Conquista, Bahia, Brasil. Email: jorge.viana@pesquisador.cnpq.br.
[1] O método LAPELINC possui um fluxo de trabalho (workflow) que compreende três etapas para a construção de corpora eletrônicos anotados, cientificamente controlados: (i) transposição; (ii) transcrição paleográfica; (iii) compilação de corpora. Para uma introdução sobre o método LAPELINC, ver Namiuti e Santos (2016a).

ISSN 1415 - 8973

> [...] precisa ser integrado com as exigências impostas pelas vertentes tecnológica, computacional e linguística, tais como: o arquivo digital, a confiabilidade do código, a necessidade de quantidade e de automação no processamento de dados. (NAMIUTI; SANTOS, 2016a).

A tarefa de disponibilizar documentos para fim de investigação científica que visa lidar com grande volume de dados e com necessidade de recuperar informações com rapidez e segurança requer ferramentas que atendam a necessidade de flexibilidade e automatização na recuperação de informação e reuso de tecnologias. Tal necessidade pode ser enfrentada com sistemas de gerenciamento de informações, banco de dados e ferramentas computacionais que garantam o fluxo de trabalho completo da construção de corpora eletrônicos anotados, que, no caso do Método LAPELINC, envolve a integração de três objetos fundamentais na construção de corpora digitais: Documento Físico (DF); Documento Digital Imagem (DDI); Documento Digital Texto (DDT). Tais objetos possuem diferentes materialidades, como apresentado por Namiuti e Santos (2016b), as quais caracterizam o modo de se fazer pesquisa em humanidades utilizando-os como corpus. Neste artigo, por recorte, trazemos a seguinte questão: *Quais as limitações e as possibilidades dos suportes materiais da escrita?*

A decifração de documentos históricos para fins de pesquisa científica normalmente envolve a transcrição das fontes originais da escrita. De acordo com Namiuti e Santos (2016a) a tarefa de transcrição paleográfica das fontes pode encerrar potencial de perda de informações a depender do suporte material ou da exploração deste suporte, podendo prejudicar a fidedignidade entre o documento original e sua versão transcrita.

Para enfrentar o problema relativo à fidedignidade, no método Lapelinc, buscamos garantir que a ciência se beneficie das vantagens do suporte digital sem dispensar a autenticidade do documento original físico. Nesta perspectiva, propomo-nos, neste trabalho, apresentar as complexidades dos suportes envolvidos na transposição material para, em seguida, descrever preliminarmente a etapa de Compilação do Método LAPELINC, a qual se caracteriza por envolver algumas ferramentas computacionais tais como o eDictor (PAIXÃO DE SOUSA; KEPLER; FARIA, 2010) e o WebSinC (NAMIUTI; SANTOS; COSTA, 2015) que articulam as tarefas e informações relacionadas a transcrição e edição das fontes documentais.

2 COMPLEXIDADES DOS SUPORTES DA ESCRITA

Como mencionamos na introdução deste artigo, o suporte material da fonte documental caracteriza o modo de fazer a pesquisa nas diferentes areas das humanidades. Namiuti e Santos (2016a) postulam que, com o advento das tecnologias, surgem novos suportes para as fontes documentais, a exemplo do digital, suporte este que traz com ele novas possibilidades e limites e caracteriza uma nova forma de fazer humanidades, definida como Humanidades Digitais.

No método LAPELINC, classificamos os documentos quanto ao suporte material do seguinte modo:

> (i) Documento Físico (**DF**), cujo o suporte material é físico, no sentido de matéria sólida, densa, corpórea, a exemplo de: pedra, argila, pergaminho, papiro, papel.

> (ii) Documento Digital Imagem (**DDI**), cujo o suporte é digital codificado para exibição visual de uma imagem a qual funciona eletronicamente a partir de lógica binária que representa bidimensionalmente informação visual imagética obtida através de uma matriz binária.

ISSN 1415 - 8973

(iii) Documento Digital Texto (**DDT**), cujo o suporte é digital codificado para exibição visual de um Texto em formato de caracteres, o qual funciona eletronicamente a partir de lógica binária que representa bidimensionalmente informação textual obtida através de uma matriz binária.

Em DFs do tipo textos manuscritos ou impressos a sequência de caracteres que formam o texto, bem como diversas informações estruturais importantes (por exemplo, a paragrafação), são codificadas de modo direto e visual.

Diferentemente, em DDIs e DDTs, eletrônicos em formato de imagem ou texto, a sequência de caracteres que formam o texto, bem como diversas informações estruturais importantes (por exemplo, a paragrafação), apesar de poderem ser decodificadas de modo direto e visual, são decorrentes de sistemas de codificação computacional-matemática.

O suporte material de DF, por ser físico, tem seu acesso limitado a um tempo e a um espaço, complexidade esta que marcou (e marca) o modo de se fazer humanidades desde a antiguidade.

Por sua vez, o suporte digital (de DDIs e DDTs), por ser eletônico, tem seu acesso ampliado a tempos e espaços simultâneos e remotos, complexidade esta que marca (e marcará) o modo de se fazer Humanidades Digitais.

Na construção de corpora anotados que partem de fontes históricas manuscritas, para se chegar a um documento do tipo texto, formato necessário à anotação computacional, há de se fazer a transcrição do manuscrito.

A transcrição paleográfica tradicional, por ter sua origem bastante remota, tradicionalmente considera os limites do suporte físico, por exemplo: a solução para as segmentação de palavras que se rege pelo princípio de juntar partes de palavras que originalmente estavam separadas e separar palavras que originalmente foram escritas juntas interfere no texto, suprimindo assim pelo menos uma natureza de informação que pode ser relevante para algum estudo

A etapa de Transcrição no método LAPELINC envolve as soluções técnicas para a edição especializada de textos antigos em meio digital/eletrônico. Compreende uma etapa inicial de leitura e transcrição paleográfica dos DDIs que servirá de entrada para a compilação do DDT.

Assim, como explicitado em Namiuti e Santos (2016a),

> […] a transcrição, no método LAPELinC, segue a recomendação de Paixão de Sousa (2004) de que, na produção de textos em meio eletrônico com a finalidade específica de construção de corpora de língua, se deve fazer uso de um processamento controlado que permita a codificação de uma grande variedade de informações, de modo confiável e transportável. Conforme tal pensamento, no processamento eletrônico de textos, as estruturas precisam ser anotadas em alguma linguagem de anotação, e depois traduzidas ou lidas por uma programação que gera a apresentação final do texto. (NAMIUTI; SANTOS, 2016a).

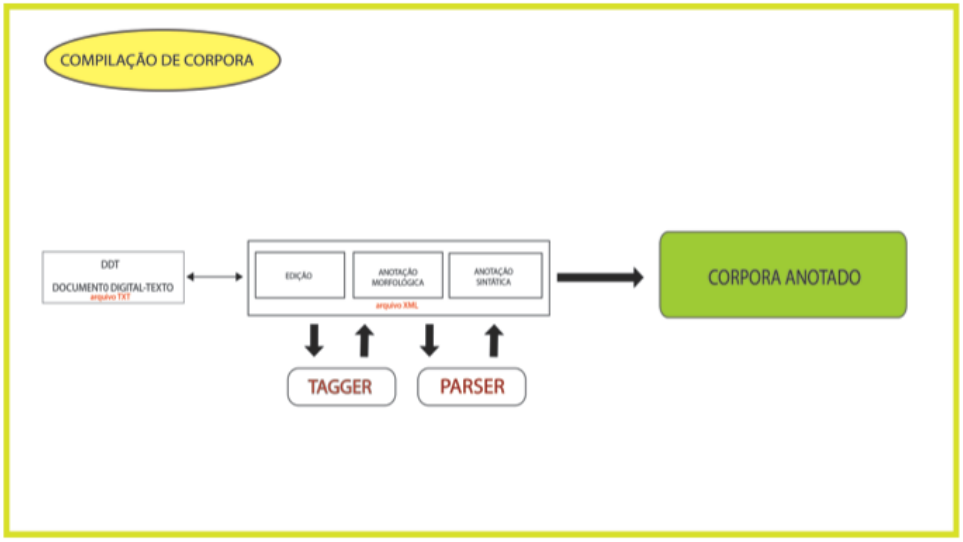
3 DO DOCUMENTO DIGITAL TEXTO (DDT) AO CORPUS ANOTADO: ETAPA DE COMPILAÇÃO DE CORPORA NO MÉTODO LAPELINC

Para a construção de corpora anotados, os textos antigos precisam ser editados, pois possuem características gráficas e grafemáticas que dificultam o processamento computacional posterior à etapa de transcrição. Não obstante, as características do texto original devem ser preservadas, devido à sua importância para estudos linguísticos e filológicos.

ISSN 1415 - 8973

As fontes documentais, no método LAPELINC, após passarem pelas etapas de transposição e transcrição ganham o formato digital de texto simples (TXT), requisito necessário para se iniciar o processo de compilação de corpora, como se vê no figura 1:

Figura 1: Representação gráfica da etapa de Compilação de corpora do fluxo de trabalho do Método LAPELINC.



Fonte: Namiuti e Santos (2016b).

Esta etapa envolve algumas ferramentas computacionais de edição, anotação linguística e gerenciamento de informações, tais como o eDictor (PAIXÃO DE SOUSA; KEPLER; FARIA, 2010) – editor de marcação extensível XML – e o WebSinC (Namiuti, Santos e Costa 2015) – aplicativo web para o trabalho de registro, armazenamento, disponibilização, visão e busca de dados em corpora cientificamente controlados, além das ferramentas de anotação linguística – *tagger* (anotador automático da morfossintaxe das palavras do texto) e *parser* (anotador automático da sintaxe das frases do texto).[2]

Os textos transcritos passam pelo processo de edição, no qual se utiliza a ferramenta eDictor. As edições dos textos são anotadas conforme o esquema de anotação proposto por Paixão de Sousa (2006) originalmente para anotação do corpus anotado do português histórico Tycho Brahe, que mantém as informações sobre a interferência realizada, o texto original e a anotação morfossintática em camadas no mesmo arquivo formato XML. Deste modo, diferentes graus de interferências de edição – das mais restritas, próprias das edições paleográficas (desdobramento de abreviaturas; decisões de leitura), às mais amplas, próprias das edições modernizadas (atualização de grafia) são contempladas nesse plano de edição. Em consequência, o DDT no método LAPELINC traz, em camadas, num único arquivo, todas as informações referentes, ao processamento do documento (metadados, transcrição, informação de edição encaixadas, anotação linguística), fato que possibilita gerar diferentes visões do texto.

> O DDT no método LAPELINC apresenta uma singularidade crucial do trabalho de edição eletrônica que explora as possibilidades próprias do suporte informático de modo a permitir a manutenção do texto original no mesmo plano em que se realizam as interferências editoriais. Assim, o documento eletrônico usado pelo editor contém todas as informações de transcrição e de edições, devidamente codificadas, de forma a garantir a integridade do texto transcrito do início ao fim do processo. Dito de outra maneira, as palavras (e todo o texto nas suas respectivas versões e graus de interferências) são mapeadas, e, por isso, podemos transitar pelas edições e recuperar as informações da palavra original no texto modernizado. É esta a característica que

[2] Para mais detalhes sobre as ferramentas de anotação, ver Costa (2015).

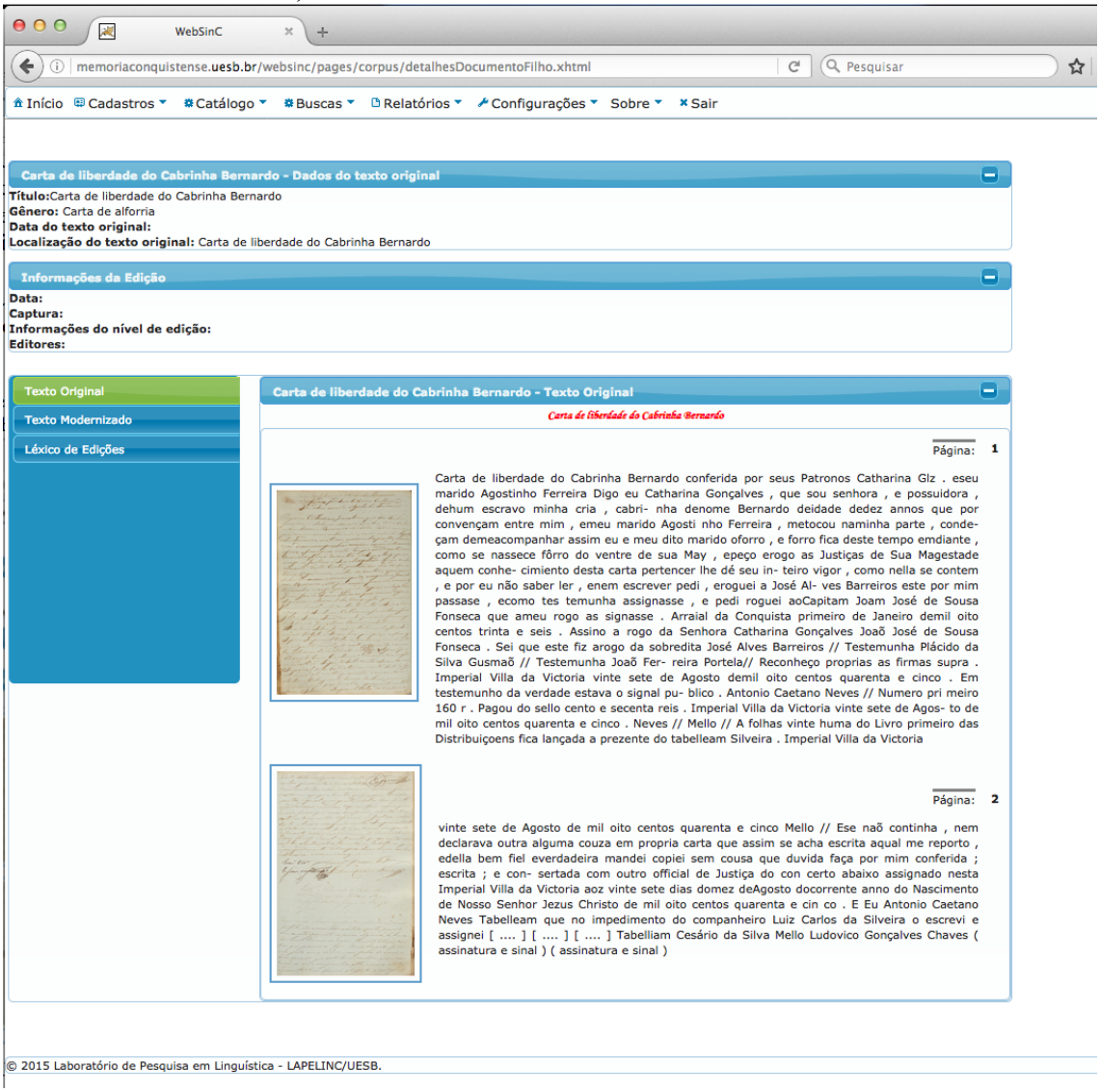
ISSN 1415 - 8973

confere controle e confiabilidade às edições eletrônicas assim desenvolvidas. (NAMIUTI; SANTOS, 2016a).

Como requisito crucial no método LAPELINC para buscar o controle e a cientificidade no processo de construção de corpora, garantindo a integração e relação entre as três fases e consequentemente entre os objetos: DF, DDI e DDT, criamos a ferramenta WebSinC com recursos específicos de sistema de gerenciamento de bancos de dados (SGDB), para gerenciar e disponibilizar os textos de corpora produzidos com esta metodologia.

A figura 2 exemplifica uma possibilidade de visão gerada pela ferramenta WebSinC, apresentando uma visualização de um DDT em sua versão editada associada ao DDI original:

Figura 2: Visão gerada pela ferramenta WebSinC a partir de DDT compilado nos moldes do Método LAPELINC, exibindo a Carta de Liberdade do Cabrinha Bernardo – 1845.



Fonte: Corpus DOViC (SANTOS; NAMIUTI, 2016b).

4 CONSIDERAÇÕES FINAIS

Para enfrentar o problema relativo as limitações e as possibilidades dos suportes materiais da escrita vimos que, com a aplicação de um método a exemplo do Método

ISSN 1415 - 8973

LAPELINC, é possível, sim, se beneficiar das vantagens do suporte digital sem dispensar a autenticidade do documento original físico.

O método LAPELINC, como demonstramos neste artigo, estabelece pontes entre as antigas fontes e o novo, explorando as complexidades do suporte digital na medida em que utiliza das vantagens da tecnologia, da computação e da linguística, como pressupõe as Humanidades Digitais, através, como vimos, do uso sistemáticos de ferramentas como o eDictor e o WebSinC.

REFERÊNCIAS

BRITO, Giovane Santos. *Do texto ao documento digital: transposição fotográfica de documentos manuscritos históricos para formação de corpora linguísticos eletrônicos*. (Dissertação) Mestrado em Linguística – Programa de Pós Graduação Linguística da Universidade Estadual do Sudoeste da Bahia, Vitória da Conquista, 2015. Orientador: Jorge Viana Santos; Co-orientadora: Cristiane Namiuti Temponi.

COSTA, Aline Silva. *WebSinC: Uma Ferramenta Web para buscas sintáticas e morfossintáticas em corpora anotados - Estudo de Caso do Corpus DOViC – Bahia*. (Dissertação) Mestrado em Linguística – Programa de Pós Graduação Linguística da Universidade Estadual do Sudoeste da Bahia, Vitória da Conquista, 2015. Orientadora: Cristiane Namiuti Temponi; Co-orientador: Jorge Viana Santos.

NAMIUTI, C.; SANTOS, J. V.; LEITE, C. M. B. Propostas e desafios dos novos meios das antigas fontes: a preservação da memória pela Linguística de Corpus. In: X Colóquio Nacional e II Colóquio Internacional do Museu Pedagógico UESB, 2011, Vitória da Conquista. *Anais do X Colóquio Nacional e II Colóquio Internacional do Museu Pedagógico UESB*. Vitória da Conquista: UESB, 2011. v. 1. p. 1-11.

NAMIUTI, Cristiane; SANTOS, Jorge Viana. Novos desafios para antigas fontes: a experiência DOViC na nova linguística histórica. In.: *E-Book do Congresso de Humanidades Digitais em Portugal: Construir pontes e quebrar barreiras na era digital – 2015*. Lisboa: Universidade Nova de Lisboa, 2016a (no prelo).

NAMIUTI, Cristiane; SANTOS, Jorge Viana. *De manuscritos históricos a corpora anotados: do Documento Digital Texto (DDT) ao corpus anotado*. Feira de Santana: UEFS, 2016b. (Conferência proferida no VIII Seminário de Estudos Filológicos, UEFS, Feira de Santana, 07 de julho de 2016. Mesa-Redonda Filologia e Linguística de Corpus).

NAMIUTI, Cristiane; SANTOS, Jorge Viana; COSTA, Aline Silva. *WebSinC*. Vitória da Conquista, UESB/LAPELINC, 2015. (Software Web desenvolvido para o trabalho de disponibilização, visão e busca de dados em corpora cientificamente controlados e anotados em diversos níveis. Disponível em: <http://memoriaconquistense.uesb.br/websinc/>

PAIXÃO DE SOUSA, Maria Clara; KEPLER, Fábio; FARIA, Pablo. E-Dictor: novas perspectivas na codificação e edição de *corpora* de textos Histórico (2010). In: SHEPHERD, Tania M.; SARDINHA, Tony B.; PINTO, Marcia (org.). *Caminhos da Linguística de Corpus*. Campinas: Mercado de Letras, 2012.

PAIXÃO DE SOUSA, Maria Clara. Memórias do Texto. *Revista Texto Digital*, n. 2., 2006. Acessado em 5 de agosto de 2014, http://www.textodigital.ufsc.br/num02/paixao.htm.

ISSN 1415 - 8973

SANTOS, J. V. *Técnicas de transporte do texto manuscrito para o meio digital*. Feira de Santana: UEFS, 2010a. (Conferência ministrada na I Oficina de Linguística de *Corpus* da Bahia (UEFS, UESB, UFBA).

SANTOS, J. V. *Apresentação de meios para o transporte do texto manuscrito para o meio digital: digitalização de documentos manuscritos e impressos*. Feira de Santana: UEFS, 2010a. (Conferência ministrada na I Oficina de Linguística de *Corpus* da Bahia (UEFS, UESB, UFBA).

SANTOS, Jorge; BRITO, Giovane Santos. Fotografia técnica de documentos para formação de corpora digitais eletrônicos: o método desenvolvido no Lapelinc. *Letras & Letras* (Online), v. 30, p. 421-430, 2014.

SANTOS, Jorge Viana; NAMIUTI, Cristiane. *De manuscritos históricos a corpora anotados: do Documento Físico (DF) ao Documento Digital Imagem (DDI)*. Feira de Santana: UEFS, 2016a. (No prelo),

SANTOS, Jorge Viana; NAMIUTI, Cristiane. *Documentos Oitocentistas de Vitória da Conquista (DOViC). Projeto Memória Conquistense*. UESB/LAPELINC, Vitória da Conquista-Bahia/Brasil, 2016b. URL: http://memoriaconquistense.uesb.br/websinc.