fapesb
Fundação de Amparo à Pesquisa do Estado da Bahia

SECRETARIA DE CIÊNCIA, TECNOLOGIA E INOVAÇÃO

Bahia GOVERNO
TERRA DE TODOS NÓS

RELATÓRIO TÉCNICO FINAL

EDITAL: PROJETO DE PESQUISA (X) AÇÃO REFERENCIA (X)

Nº PROTOCOLO

Coordenador do Projeto (Proponente): Zenaide de Oliveira Novais Carneiro **Vice-Coordenador:** Mariana Fagundes de Oliveira		
Edital (nº /ano): 2010	**Pedido (nº/ano):** 5566/2010	**Termo de Outorga (nº/ano):** PET0028/2010
Telefones p/ contato **TELEFONE DO PROJETO:** 3161-8522	**E-mail:** cedohs@gmail.com	
Data de Assinatura do Termo de Outorga: 09/11/2010		
Vigência: 09/11/2010 a 09/11/2012. Alterada para 16/02/2013 (Portaria 115/2012). DOE de 8/2/2012). **Projeto concluído no período de vigência indicado no T.O? SIM (X) NÃO ( )**		
Data(s) Recursos recebidos: R$ 22.150,00 (Corrente 9,300,00 e Capital 12.850,00). Bolsas 9.600,00) **Total -** R$ 31.750,00 **Recursos oriundos de aplicação financeira:** R$ 1.288,02 **Total de recursos alocados para o projeto:** R$ 33,038,02_____,___ **Total de recursos utilizados para o desenvolvimento da pesquisa: R$** R$ 33,038,02		
Período de abrangência do Relatório: 09/11/2010 a 16/02/2013. **Tal período corresponde a que meses do cronograma de pesquisa?** *Meses: **1 - 26 meses** (considerando a data de assinatura do termo de outorga do projeto).*		

1. DESCRIÇÃO DO PROJETO

Título do Projeto: CE-DOHS CORPUS ELETRÔNICO DE DOCUMENTOS HISTÓRICOS DO SERTÃO (FAPESB 5566/2010)		
Instituição de vínculo do Proponente: Universidade Estadual de Feira de Santana	**Unidade:** Departamento de Letras e Artes	**Departamento:** Letras e Artes
Área: Linguística	**Subárea: Linguística Histórica**	
Instituições Parceiras: ATENÇAO – CONFIRA **ANEXO** DETALHADO DAS PARCERIAS – INTINULADO QUEM SOMOS.		

Nome / Sigla	Município / Estado	Área e Subárea
Universidade Estadual de Campinas (UNICAMP)	Campinas/São Paulo	Linguística Linguística Histórica

Universidade Federal da Bahia (UFBA)	Salvador/Bahia	Linguística Linguística Histórica
Universidade Estadual do Sudoeste Baiano (UESB)	Vitória da Conquista/Bahia	Linguística Linguística Histórica

1.2 Mudança na equipe responsável pelo desenvolvimento do projeto Preencher apenas nos casos de mudanças na equipe aprovada. Obs. Não incluir bolsistas neste campo do relatório.

1.2.1 Equipe Principal (*Pesquisadores vinculados à instituição executora e principais responsáveis pela execução das atividades descritas neste relatório)*

Nome	**Titulação**	**Instituição**	**Período de permanência no projeto** (data de início e termino das atividades)
Zenaide de Oliveira Novais Carneiro	Pós-Doutorado	UEFS	09/11/2010 a 09/11/2012
Mariana Fagundes de Oliveira	Doutorado	UEFS	09/11/2010 a 09/11/2012
Telma Regina Garrido de Araújo	Especialização	UEFS	09/11/2010 a 09/11/2012

1.2.2 Equipe Parceira *(Pesquisadores vinculados às instituições parceiras e/ou integrados a equipe após aprovação da proposta, se ocorrer)*

Nome	**Titulação**	**Instituição**	**Período de permanência no projeto** (data de início e termino das atividades)
Charlotte Marie Chambelland Galves	Pós-Doutorado	UNICAMP	09/11/2010 a 16/02/2013
Cristiane Namiuti Temponi	Pós-Doutorado	UESB	09/11/2010 a 16/02/2013
Jorge Viana Santos	Doutorado	UESB	09/11/2010 a 16/02/2013
Klebson Oliveira	Pós-Doutorado	UFBA	09/11/2010 a 16/02/2013
Maria Clara Paixão de Sousa	Pós-Doutorado	Faculdade de Filosofia, Letras e Ciências Humanas	09/11/2010 a 16/02/2013
Tânia Conceição Freire Lobo	Doutorado	UFBA	09/11/2010 a 16/02/2013

2. DESENVOLVIMENTO DO PROJETO - *Delinear a execução do Projeto de Pesquisa, especialmente no que se refere a:*

2.1 Objetivo Geral

Transcrito do projeto original aprovado

Este projeto de pesquisa tem como objetivo principal converter um *corpus digital* de documentos históricos do sertão baiano em um *Corpus Digital Eletrônico* em linguagem XML, nos moldes do *Corpus Histórico do Português Tycho Brahe/CTB* (Unicamp) (http://www.tycho.iel.unicamp.br/~tycho), conhecido também como CTB, para uso na rede mundial de computadores, via manipulação dos dados para análise linguística com recursos das ferramentas de buscas. O *corpus* base é o do banco de textos já constituído e quase todo organizado do Projeto *Vozes do Sertão em Dados: História, Povos e Formação do Português Brasileiro* (CNPq, 2009), subprojeto de um projeto amplo de âmbito nacional, que vem sendo desenvolvido na UEFS, como parte de uma agenda do amplo *Projeto Nacional Para a História do Português (PHPB)* e do *Programa para a História Português (PROHPOR-UFBA)*, cabendo a equipe da UEFS fazer o levantamento, a compilação, a edição de documentação e os estudos linguísticos da região semiárida baiana e, no projeto atual, a edição eletrônica do banco digital, como dito[1].

Os objetivos específicos do *CE-DOHS Corpus Eletrônico de Documentos Históricos do Sertão*/BA são os seguintes:

1. Implementação do acervo digital inédito de um grande *corpus* em formato de texto computacionalmente manipulável, em linguagem XML, que permita recuperar informações gráficas de cunho filológico dos documentos originais, além disso, gerar bases anotadas (morfológica e sintática) para análise linguística em parceria com o projeto *Corpus Histórico do Português Tycho Brahe/CTB* (Unicamp) (http://www.tycho.iel.unicamp.br/~tycho. Ou seja, um *corpus* que permita usos de mecanismos que gerem versões diversas acessíveis em processamentos de buscas automáticas;

2. Agora que desenvolvemos um piloto, cujo primeiro resultado pode ser acessado em http://www.uefs.br/dohs/corpusxml. html, e cujos dados eletrônicos vão dar origem a um livro organizado por membros da UEFS/UFBA e da UNICAMP (produto do projeto *Vozes do Sertão em Dados* - CNPq), citado, já usando a busca eletrônica, um procedimento inédito que até agora somente vem sendo utilizado em teses e artigos,

[1] O Projeto *Vozes do Sertão em Dados* e o CE-DOHS relacionam-se com o *Programa para a História da Língua Portuguesa* (PROHPOR), coordenado por Rosa Virgínia Mattos e Silva, na Universidade Federal da Bahia (UFBA), especificamente em seu arco temporal da história do português brasileiro (PB), e resulta de desdobramentos de uma agenda de trabalho iniciada por Ilza Maria de Oliveira Ribeiro, na UEFS, em 1997, na qual se previa a edição de documentos diversos, no âmbito do projeto *Contribuições para a constituição de um banco de textos e de um banco de dados para o estudo da história do português do Brasil, do séc. XVII ao XX.* Integra o *Projeto Nacional Para a História do Português Brasileiro* (PHPB), coordenado por Ataliba de Castilho, da Universidade de São Paulo (USP) e da Universidade Estadual de Campinas (UNICAMP), via equipe baiana, coordenada por Tânia Conceição Freire Lobo, da UFBA.

vamos fazer um banco que será referência no estado da Bahia, pois conterá milhares de palavras a serem acessadas em segundos na rede e já com dados quantificados, exemplos separados computacionalmente. Uma demonstração com o banco do português europeu pode ser acessado em www.tycho.iel.unicamp.br, mediante senha obtida automaticamente depois de cadastro controle. Prioritariamente os dados inéditos serão objeto de exploração de nossa equipe baiana (livros, teses, artigos), que serão imediatamente disponibilizados para pesquisadores do Brasil e do mundo, dando a conhecer sobre uma região pouquíssima estudada do ponto de vista linguístico, o semiárido baiano, mas que tem uma história de contato única entre indígenas, brancos e africanos ao longo de todo o período colonial e com repercussões no nosso sistema educacional, que se baseia em uma situação de diglossia linguística. Esse projeto certamente oferecerá contraparte a outros importantes com dados orais do presente, desenvolvidos na UFBA (a exemplo do vertentes.ufba.br), UNEB e outros e dos projetos citados ligados ao PHPB nacional.

3. Oferecer catálogos do léxico eletrônicos de *corpus* serial (século XVII-XXI) de toda a região do semiárido baiano;

4. Como subprodutos, oferecer a geração computacional automática de léxicos das edições em XML, catálogos dos documentos, edições variadas (conservadoras, modernizadas, etc.), além de outros subprodutos, todos gerados pela transformação XML;

5. Disponibilizar em um portal toda a produção do banco digital nos moldes do modelo disponível em *www.tycho.iel.unicamp.br*, através de parceria firmada institucionalmente entre UEFS/Unicamp, já em implementação. Essa parceria se faz para otimização de usos e atualização de tecnologias. A Unicamp por sua vez desenvolve tecnologias para banco de dados através de parcerias com outras universidades, a exemplo da UPENN e outras, como será explicitado na metodologia.

Comentar eventuais alterações ocorridas com relação ao objetivo proposto inicialmente, lembrando que não pode ocorrer alteração do OBJETO do Termo de Outorga.

Não houve alterações.

2.2 Objetivos Específicos

Transcritos do projeto original aprovado

Os objetivos específicos estão ligados às metas e aos posteriores resultados deste projeto, a saber:

CONCLUÍDO

1. Fazer o levantamento de toda a documentação do banco para elaboração do catálogo digital. É importante salientar que já foi feito um levantamento

de quase 2.000 cartas, já está em fase final de organização;

2. Concluir a organização de documentos que estão em fase de edição digital em Word;

3. Concluir as autorizações de uso em consonância com que estabelece o código de Ética da UEFS, já em processo de execução, sobretudo para acervos do século XX e XXI;

4. Treinamento em oficinas do uso do E-Dictor para controle do texto base bruto e transformação XML, a serem agendas em parceria com a UNICAMP, aliando recursos do projeto *Corpus Histórico do Português Tycho Brahe* e o *Vozes do Sertão em Dados*, com interesses para os dois projetos através da comparação dos dados dos dois bancos, propiciando uma multiplicação de artigos e teses, etc.;

5. Testes da codificação XML;

6. Estabelecimento do formato da base de dados e do sítio-web do projeto piloto final;

7. Teste do projeto dessa primeira fase do projeto, de forma que venha ser possível a alimentação de textos até atingir milhões de palavras em projeto contínuo na UEFS, partilhado e acessado na rede mundial de computadores;

8. Complementar o banco com documentação que está em fase de prospecção, através do resgate em acervos de Recife e Campinas;

9. Digitalizar os arquivos doados por diversos arquivos privados;

10. Anotação morfológica através da equipe técnica da a ser constituída na UEFS com suporte do pessoal técnico do Corpus Histórico do Português Tycho Brahe.

Comentar eventuais alterações ocorridas com relação aos objetivos específicos propostos inicialmente

Não houve alterações.

2.3 Cronograma de atividades *(Descrever de maneira sintética e objetiva o desenvolvimento das atividades previstas e comentar eventuais alterações ocorridas no período de abrangência deste relatório)*

Atividades previstas no projeto original para o período	**Atividades realizadas no período de abrangência deste relatório / Resultados finais alcançados**

Treinamento, em Oficinas, do uso do E-dictor para controle do texto base bruto e transformação XML, aliando recursos do Projeto Tycho Brahe (UNICAMP) e do Projeto Vozes do Sertão em Dados (UEFS).	Treinamento, em Oficinas, do uso do **E-dictor** para controle do texto base bruto e transformação XML, aliando recursos do Projeto Tycho Brahe (UNICAMP) e do Projeto Vozes do Sertão em Dados (UEFS).
Testes da codificação XML.	Testes da codificação XML.
Estabelecimento do formato da base de dados e do sítio-web do projeto piloto final.	Estabelecimento do formato da base de dados e do sítio-web do projeto piloto final. Acesso: www.uefs.br/cedohs Também pode ser acessado através do *Corpus Histórico do Português Tycho Brahe* http://www.tycho.iel.unicamp.br/~tycho/corpus/index.html
Comentários sobre as diferenças entre atividades previstas e as realizadas	
Não há diferenças entre as atividades previstas e as realizadas.	

2.4 Resultados alcançados / Produtos obtidos

Os objetivos do projeto foram plenamente atingidos e os produtos correspondentes são os seguintes:

BANCO ELETRÔNICO - Acessível em www.uefs.br/cedohs na parte de apresentação e no processamento em www.tycho.iel.unicamp.br, em consonância com a parceria estabelecida via Convênio Guarda-Chuva UEFS/Unicamp e Termo Aditivo/CE-DOHS/CORPUS HISTÓRICO DO PORTUGUÊS TYCHO BRAHE, anexos.

- O banco conta com **MANUSCRITOS** (cartas pessoais do século xix e xx, intitulados cartas brasileiras, e com livros de razão inéditos), **IMPRESSOS** E O **CORPUS ORAL**.

MANUSCRITOS -- DAS CARTAS BRASILEIRAS

CF. ANEXO O SITE DO CE-DOHS PARA VISUALIZAÇÃO

Acesso: www.uefs.br/cedohs e também pode ser acessado através do *Corpus Histórico do Português Tycho Brahe* http://www.tycho.iel.unicamp.br/~tycho/corpus/index.html

Esse acervo eletrônico conta com mais de 1200 cartas pessoais, editadas em formato XML, constituindo-se no maior acervo dessa natureza no Brasil, conforme pode ser visualizado em www.uefs.br/cedohs. Para comprovar isso, acessem-se

os projetos correspondentes no Brasil, da mesma natureza, a exceção do *Corpus* Tycho Brahe, o maior, mas que se volta para o português europeu, configurando o CE-DOHS como o maior do Brasil nessa categoria de textos brasileiros.

Toda a descrição dos acervos pode vista no site www.uefs.br/cedohs, além das imagens anexas a esse relatório.

Esse acervo recebeu o título de CARTAS BRASILEIRAS para melhor correlação com o correspondente em formato impresso (versão publicada em 3 volumes[2] que compõem *Cartas brasileiras (1809-2000): coletânea de fontes para o estudo do português* (Fapesb 1493/2010/Consepe 102/2009), do banco digital do trabalho coletivo de participantes do Projeto *Vozes do Sertão em Dados*: *história, povos e formação do português,* editados de acordo com as normas da Plataforma Nacional do PHPB e transformados no CE-DOHS para o formato XML[3]. O resultado mais importante atingido é o de que esse extenso banco está agora disponível, sem prejuízo das informações filológicas, com edições confiáveis para a execução de artigos, teses, dissertações, monografias e demais trabalhos acadêmicos, o que vem sendo feito, antes mesmo do banco ficar pronto.

O banco já vem sendo amplamente consultado e já foi citado com base de trabalhos em Mesas-Redondas Nacionais, a exemplo de Mesas apresentadas no encontro recente da *Associação Brasileira de Linguística*, ocorrida na UFRN em fevereiro de 2013. A base eletrônica desenvolvida no CE-DOHS também serviu de inspiração para projetos semelhantes, desenvolvidas na Universidade de Lisboa/Portugal.

Como dito acima, o *corpus* eletrônico desse acervo foi retirado de C*artas brasileiras (1809-2000): coletânea de fontes para o estudo do português*[4]*,*

[2] O volume 3, de caráter diferenciado, produzido por mãos inábeis, está sendo objeto de escrita de um livro de análise de dados. Como não foi prevista a sua inserção no projeto original, consta apenas uma pequena amostra. O todo será disponibilizado assim que o livro de análises inéditas ficarem prontas, o que está previsto para 2014.

[3] O PHPB nacional desenvolve agendas e metodologias de pesquisa compartilhadas com diversas universidades brasileiras e tem, entre os seus objetivos, ações metodológicas voltadas para a organização de *corpora* diacrônicos com variados tipos de textos manuscritos e impressos, que vêm servindo de base para a composição de uma *Plataforma de Corpora do PHPB,* a cargo de Afrânio Barbosa, da Universidade Federal do Rio de Janeiro (UFRJ), e de Marcelo Módulo, da USP. O Projeto *Vozes do Sertão em Dados* partilha dessa agenda, por meio da prospecção, da edição de documentos e da formação de *corpora* representativos de demandas histórico-sociais da região semiárida baiana, com repercussões sobre o processo de formação histórica do PB, com amplo contato linguístico de populações de origem portuguesa, indígena e africana, bem como com projetos temáticos de análise linguística.

[4] Os volumes são: VOLUME 1 (1809-1904) Zenaide de Oliveira Novais Carneiro (Organização)/ *CD-ROM* 1. Cartas para vários destinatários (1809-1904): edição fac-similada/Zenaide de Oliveira Novais Carneiro/*CD-ROM* 2. Cartas para Severino Vieira, governador da Bahia (1901-1902): edição fac-similada/Zenaide de Oliveira Novais Carneiro / *CD-ROM* 3. Cartas para Cícero Dantas Martins, Barão de Jeremoabo (1880-1903): edição fac-similada/Zenaide de Oliveira Novais Carneiro./VOLUME 2 (1902-1993)/Zenaide de Oliveira Novais Carneiro; Mariana Fagundes de Oliveira; Norma Lucia Fernandes de Almeida (Organização)/*CD-ROM* 1. Cartas do acervo Dantas Jr. (1902-1962): edição fac-similada/ Zenaide de Oliveira Novais Carneiro; Marta Carvalho Ferreira Lisboa; Mônica Araújo Cruz; Denise Branco Cerqueira; Eliane Santos Leite/*CD-ROM* 2. Cartas baianas: o acervo de João da Costa Pinto Victoria (1911-1958): edição fac-similada/ Zenaide de Oliveira Novais Carneiro; Maria Rosane Passos; Priscila Tuy Batista; Anderléia Mascarenhas/*CD-ROM* 3. Correspondências amigas: o acervo de Valente, Bahia (1980-1993): edição fac-similada/Mariana Fagundes de Oliveira; Maiany Soares de Oliveira; Adilson Silva de

organizada em três volumes: a apresentação de cada volume e a descrição dos acervos do formato impresso foi preservada nos metadados do acervo em fichas em formato XML.

A edição XML foi organizada de maneira a permitir que os pesquisadores interessados na história do português brasileiro (PB) possam constituir *corpora* específicos de acordo com seus objetivos. Para isso, cada pesquisador poderá utilizar-se das informações sobre a documentação e sobre seus escreventes contidas na descrição dos acervos no site e nos metadados na base de dados eletrônica.

É importante salitentar que são geradas versões automáticas de edições, modernizadas e conservadoras, além do inventário lexical, sem prejuízo das informações do documento original. Do ponto de vista da documentação, é possível a sua organização por ordem cronológica, tanto por data de escrita, como é tradicional na Linguística Histórica, quanto por data de nascimento do autor (individual), no caso de autores que possuem acervos mais significativos, ou por grupos de autores, como tem sido feito na base de dados para estudos em Linguística Diacrônica, no âmbito do *Corpus Histórico do Português Tycho Brahe,* coordenado por Charlotte Galves, da UNICAMP. Essa base documental pode ajudar nos estudos sobre variação e mudança do PB, do século XIX ao século XX, se se considerar a cronologia dos textos, por data de escrita das cartas, ou a partir do século XVIII, se a cronologia for feita por data de nascimento dos seus remetentes. As informações sobre o grau de escolaridade dos escreventes, aliadas a possíveis análises do tipo de escrita, permitem também a separação das cartas representativas de variedades *standard* e não-*standard* do PB, em uma perspectiva histórica. Isso é possível agora também em versão XML na REDE MUNDIAL DE COMPUTADORES. Essas formulações podem também serem feitas com base na hipótese central de organização dos *corpora* do PHPB baiano e nacional: - Amostras de documentos produzidos por indivíduos, para quem o português tenha sido a primeira ou a segunda língua; - Amostras documentais escritas por indivíduos alvos de diferentes processos de contato com a língua escrit, de modo que se possam oferecer meios para o estudo de “normas vernáculas” e “normas cultas”, representativas do processo histórico do PB.

O *parser* e a base eletrônica não forma como estão constituídas permitem gerar automaticamente **catálogos eletrônicos do léxico de edições** de todo o corpus da base.

Com relação à anotação morfológica, essa depende exclusivamente no momento do projeto matriz *Corpus Histórico do Português Tycho Brahe*. Na fase do projeto piloto, desenvolvido no pós-doutoramento de Zenaide de Oliveira Novais Carneiro já foi feito essa anotação do corpus de cultos do século XIX, disponível em http://www.uefs.br/dohs/va004.html). Agora entrou para anotação a contraparte, o *corpus* de não-cultos, do século XIX. Paulatinamente, outros *corpora* serão submetidos a anotações morfológicas, essa é a contraparte do *Corpus*

Jesus/VOLUME 3 (1906-2000)/ Huda da Silva Santiago; Zenaide de Oliveira Novais Carneiro; Klebson Oliveira (Organização)/*CD-ROM* 1. Cartas em Sisal: Riachão do Jacuípe, Conceição do Coité e Ichu (1906-2000): edição fac-similada/Huda da Silva Santiago/No caso do volume 3, a edição das cartas vem também no formato impresso, por ser um acervo diferenciado e pouco comum, escrito exclusivamente por remetentes pouco escolarizados.

Histórico do Português Tycho Brahe para os *corpora* com datas anteriores ao século XIX, assim como a anotação sintática e o uso da ferramenta de busca, o *Corpusearch.* Essa anotação dos *corpora* é feita em blocos de textos repartidos regularmente pelo tempo por período de 50 anos.

Essa parceria com o *Corpus Histórico do Português Tycho Brahe/CTB,* que possui uma ampla base de textos portugueses, já está em processo de expansão como textos brasileiros, que segundo, a coordenadora do Projeto Charlotte Galves (2012)

> Trata-se de constituir um CTB Brasil em colaboração estreita com grupos de pesquisa brasileiros envolvidos na edição de textos não literários produzidos no Brasil ao longo da sua história (...), o resultado dessa parceria já redundou na inclusão no Corpus de textos de uma imensa relevância para a compreensão da história do português no Brasil: as *Cartas brasileiras*, já mencionadas e as *Atas da Sociedade protetora dos desvalidos*. Trata-se de ampliar e sistematizar essa parceria, de maneira que se constitua um Corpus sintaticamente anotado do português brasileiro histórico de pelo menos 500.000 palavras. Desse *Corpus* deverão também fazer parte textos literários, de maneira a permitir uma comparação sistemática das duas vertentes do português no Brasil, o 'culto' e o 'popular' (pelo menos naquilo que transparece de textos não literários). Está sendo também constituído um *corpus* de jornais, no âmbito de teses e dissertações em andamento. Enfim, serão acrescentados textos orais representativos do português afro-brasileiro. Os grupos associados a essa pesquisa são os seguintes: Projeto Brasiliana Digital, USP-São Paulo; Projeto PHPB-Rio de Janeiro, na Universidade Federal do Rio de Janeiro, coordenado pela Profa Célia Lopes; Projetos projeto Vozes do Sertão e do banco CE-DOHS e PHPB Bahia na Universidade Estadual de Feira de Santana - UEFS, coordenados pela Profa Zenaide Carneiro, Projeto *Memória Conquistense*, na Universidade Estadual do Sudoeste da Bahia –UESB, coordenado pelos Profs Cristiane Namiuti e Jorge Viana Projeto PHPB-Bahia, na Universidade Federal da Bahia coordenado pela Profa Tânia Lobo. Já existe uma intensa parceria entre os três grupos baianos e os projetos associados ao *Corpus Tycho Brahe*, concretizada em publicações já efetuadas (**cf. Lobo e Oliveira, 2009) ou em planejamento (cf. Galves, Lobo e Oliveira, em prep.; Carneiro, Galves, Lobo, em prep**.), em transferência de tecnologia e de competência, e em parcerias para disponibilização de textos na Internet. A existência de financiamento para a construção do CTB no âmbito deste projeto temático agilizará essa cooperação, uma vez que permitirá que mais bolsistas efetuem as tarefas de formatação e anotação dos textos (Trecho retirado do projeto filiado *A língua portuguesa no tempo e no espaço: contato linguístico, gramáticas em competição e mudança paramétrica/Portuguese in time and space: linguistic contact, grammars in competition and parametric change (FAPESP, 2012-2016*))

EXPANSÃO E NOVAS PERSPECTIVAS:

EXPANSÃO

Foram inseridos também o **CORPUS IMPRESSO**.

CF. ANEXO O O SITE DO CE-DOHS Acesso: www.uefs.br/cedohs e também

pode ser acessado através do *Corpus Histórico do Português Tycho Brahe* http://www.tycho.iel.unicamp.br/~tycho/corpus/index.html

Esse acervo também produto do projeto *Vozes do Sertão em Dados: história, povos e formação do português brasileiro* (CNPq) recebeu a contraparte em versão XML, isto é, a edição fac-similada em versão eletrônica de documentos impressos baianos, da cidade de Feira de Santana: cartas de leitores e redatores dos jornais *O Progresso* e *Folha do Norte*, além de anúncios do *Folha do Norte*. Data de junho de 1900 a fundação do jornal *O Progresso*, com circulação somente até 1909, inicialmente bissemanal, depois aos domingos. No mesmo ano em que este jornal saiu das bancas, foi fundado o jornal *Folha do Norte*, que circula até hoje – como diário até 1977, e como semanário, a partir daí –, já com mais de 100 anos de existência, o mais antigo da Bahia, referência importante sobre a história de Feira de Santana. *O Progresso* e o *Folha do Norte* são resultado do avanço do parque tipográfico no Brasil, partindo do Rio de Janeiro, no século XIX.

São, no total, 368 documentos reunidos neste material:

a) 36 cartas de leitores e redatores do jornal *O Progresso*, datadas de 1901 a 1908 (28 cartas de leitores e 8 cartas de redatores).
b) 121 cartas de leitores e redatores do jornal *Folha do Norte*, datadas de 1909 a 1997 (89 cartas de leitores e 32 cartas de redatores)
c) 211 anúncios do jornal *Folha do Norte*, datados de 1910 a 2006.

Além do que foi previsto no projeto original, foram acrescidos outros acervos de universidades parcerias, como a UFBA, organizados pelo PROHPOR e PHPB-ba, a exemplo dos acervos das **Cartas do Recôncavo Baiano**, da **Família Freire** e **da Família Soledade.**

CARTAS DO RECÔNCAVO BAIANO
CARTAS DA FAMÍLIA FREIRE
CARTAS DA FAMÍLIA SOLEDADE

PERSPECTIVAS

Na fase II, posterior, está previsto, e já em desenvolvimento no âmbito do projeto de pós-doutoramento, por Mariana Fagundes de Oliveira, co-coordenadora do CE-DOHS, a saber:

MANUSCRITOS RAROS DA FAZENDA BREJO SECO

Encontra-se em andamento, a edição XML de livros do acervo de Brejo Seco. Trata-se de documentação inédita, sendo editada por Mariana Fagundes, co-responsável pelo CE-DOHS, que serviu de base para o livro raro intitulado *Uma comunidade rural do Brasil antigo: aspectos da vida patriarcal no sertão da Bahia nos séculos XVIII e XIX,* recém lançado pela Fundação Pedro Calmon em parceria pela UEFS, em comemoração ao 55° aniversário da primeira e única

edição. É o livro que o autor define como “a história de uma fazenda de criação dos sertões da Bahia, uma reconstituição da fazenda do “Brejo do Campo Seco”, nos seus aspectos social, econômico e histórico”, configura-se como uma importante obra sobre os sertões baianos. Resulta de ampla prospecção nos chamados livros de “Razão” e do “Gado”, parte do que se constituiu o rico arquivo privado das famílias Almeida, Pinheiro Pinto e Pinheiro Canguçu, justamente os livros inéditos que estão em processo de edição e que fará parte futuramente do acervo CE-DOHS. A importância desses livros, o “Livro do Gado” e o “Livro de Razão”, advém do fato de se constituírem em registros raros feitos de forma sistemática por três dos seus senhores: o escrivão português Miguel Lourenço, inicialmente, como contador no “Tribunal dos ausentes” (1742-1743), cujos registros da fazenda se iniciam em 1755 e vão até 1885, o brasileiro, genro de Miguel Lourenço, Antonio Pinheiro Pinto, a partir de 1794, e o seu filho, Inocêncio Pinheiro Canguçu, neto de Miguel Lourenço, a partir de 1822. Nesses livros, constam não apenas lançamentos contábeis. Há anotações minuciosas da vida na fazenda, magistralmente interpretadas pelo autor, Lycurgo Castro Santos Filho, que contou também com parte da documentação de Exupério Pinheiro Canguçu, filho de Inocêncio Pinheiro Canguçu, constituída a partir de 1838, quando assumiu a fazenda. Desse modo, a cada capítulo dessa obra, o leitor é levado a conhecer diferentes perspectivas do sertão agropecuário baiano, seja a do homem sertanejo, através de aspectos de âmbito privado e familiar, do seu trabalho na agricultura e na pecuária e até mesmo de suas relações comercias e sociais.

Esse acervo permitirá o resgate de outros documentos dispersos do arquivo da fazenda do Brejo do Campo Seco e da edição do “Livro do Gado” e do “Livro de Razão, sob a responsabilidade do Projeto *CE-DOHS - Corpus Eletrônico de Documentos Históricos do Sertão*, generosamente disponibilizados pelo Dr. Lycurgo Castro Santos Neto, filho do autor.

CORPUS ORAL

- Também está em andamento a inserção de dados **Orais do Projeto** - A língua portuguesa no semiárido baiano (1994-2012/FAPESB). Fases 1, 2 e 3 (Feira de Santana), desenvolvido sob a coordenação de Norma Lucia Fernandes de Almeida e minha, Zenaide de Oliveira Novais Carneiro.

É um banco que vem fornecendo as bases para as dissertação da área da sociolinguística, no programa de Mestrado em Estudos Linguísticos da UEFS recém implantado, sem o qual o atual. O *corpus* oral é o seguinte, com mais de 100 amostras de fala:

Piemonte da Diamantina: Ancelino da Fonseca/Piabas

Chapada Diamantina: Rio de Contas (localidades de Barra/Bananal/Mato Grosso);

Nordeste: Jeremoabo (Casinhas, Lagoa do Inácio e Tapera);

Paraguaçu: Matinha e Feira de Santana (migrantes e não-migrantes).

Obs.: no momento em processo de edição XML. Consulte no site, versão pdf.

2.5 Fatores de facilitação ou de dificultação relativos ao desenvolvimento do Projeto **Os fatores de dificultação (espaço físico) foram dirimidos com uma pequena ampliação da sala de pesquisa.**
Acordo com Universidades com o mesmo projeto, em especial a parceria com o Projeto Tycho Brahe Corpus Histórico do Português/CTB, sediado no Instituto da Linguagem da Unicamp (www.tycho.iel.unicamp.br).

2.6 Mecanismos gerenciais de execução multi-grupo ou multi-instituição (caso existam)
Descrever e avaliar os mecanismos utilizados para gerenciamento de projetos executados através de parceria.

O projeto CE-DOHS CORPUS ELETRÔNICO DE DOCUMENTOS HISTÓRICOS DO SERTÃO filia-se ao projeto matriz o *Corpus Histórico do Português Tycho Brahe/CTB* (CNPq) disponível em (http://www.tycho.iel.unicamp.br/~tycho/ e ao projeto interdisciplinar *A língua portuguesa no tempo e no espaço: contato linguístico, gramáticas em competição e mudança paramétrica/Portuguese in time and space: linguistic contact, grammars in competition and parametric change (FAPESP, 2012-2016*), desenvolvidos no Instituto de Linguagem da Universidade Estadual de Campinas (UNICAMP), ambos sob a supervisão da Profª Drª Charlotte C. Galves. Essa parceria está formalizada através de Convênio Guarda-Chuva entre a UEFS e a Unicamp, com o propósito fundamental desenvolver novas metodologias para formação de grandes bancos de dados para o processamento de texto para fins linguísticos. É importante salientar que o CE-DOHS, através dessa parceria com o TYCHO BRAHE CORPUS também se beneficia dos acordos com o projeto pioneiro *Penn Helsinki Parsed Corpus of Middle English*, (http://www.ling.upenn.edu/hist-corpora), coordenado por Anthony Kroch na Universidade da Pensilvânia e os seus afiliados

Cabe salientar que já há uso de tecnologia integradas como o *Corpus Dialetal Sintático (CordialSin),* por Ana Maria Martins, na Universidade de Lisboa, o que é importante para um uso comparativo entre o banco de dados orais e escritos com inquéritos e textos do Português Europeu como o banco CE-DOHS. Essa parceria já destacada por Charlotte Galves no texto do projeto *A língua portuguesa no tempo e no espaço: contato linguístico, gramáticas em competição e mudança paramétrica/Portuguese in time and space: linguistic contact (UNICAMP/FAPESP)* prevê essa comparabilidade do português brasileiro com o português europeu, uma vez que grande parte do acervo brasileiro é constituido de cartas (cf.por exemplo as *Cartas brasileiras* já presentes no Corpus – va_004).

Através do projeto *Vozes do Sertão em Dados: história, povos e formação do português brasileiro (*CNPq. Número do processo: 401433/2009-9) CONSEPE - 27/08/2009 102/2009), uma vez que o CE-DOHS é a contraparte eletrônica do banco Documentos Históricos do Sertão ou DOHS, do projeto *Vozes do Sertão* citado acima, filia-se também a importantes projetos nacionais, como o *Programa para a História da Língua Portuguesa* (PROHPOR), criado pela Professora Rosa

Virgínia Mattos e Silva, na Universidade Federal da Bahia (UFBA), especificamente em seu arco temporal da história do português brasileiro (PB). Integra o *Projeto Nacional Para a História do Português Brasileiro* (PHPB), coordenado por Ataliba de Castilho, da Universidade de São Paulo (USP) e da Universidade Estadual de Campinas (UNICAMP), via equipe baiana, ou PHPB-ba, coordenada por Tânia Conceição Freire Lobo, da UFBA e da Plataforma do PHPB-ba, por Zenaide de Oliveira Novais Carneiro. O PHPB nacional desenvolve agendas e metodologias de pesquisa compartilhadas com diversas universidades brasileiras e tem, entre os seus objetivos, ações metodológicas voltadas para a organização de *corpora* diacrônicos com variados tipos de textos manuscritos e impressos, que vêm servindo de base para a composição de uma *Plataforma de Corpora do PHPB,* a cargo de Afrânio Barbosa, da Universidade Federal do Rio de Janeiro (UFRJ), e de Marcelo Módulo, da USP. O Projeto *Vozes do Sertão em Dados* e o CE-DOHS partilham dessa agenda, por meio da prospecção, da edição de documentos e da formação de *corpora* representativos de demandas histórico-sociais da região semiárida baiana, com repercussões sobre o processo de formação histórica do PB, com amplo contato linguístico de populações de origem portuguesa, indígena e africana, bem como com projetos temáticos de análise linguística.

O projeto também tem feito outras parcerias com Co-cooperação com projetos regionais. Isso tem sido feito através da parceria com o projeto *Corpora digitais para a história do português brasileiro: Documentos históricos da região do Sudoeste da Bahia, aliança PHPB-TYCHO BRAHE* (FAPESB 6171/ 2010), sob a coordenação de Jorge Viana Santos, da UESB.

E recentemente iniciou processo de acordo com a Brasiliana Digital/Biblioteca Mindlin, coordenada por Maria Clara Paixão de Souza, uma das consultora do Corpus CE-DOHS.

É importante consultar o **ANEXO QUEM SOMOS** para visualização dessas parcerias.

2.6.1 Mecanismos de devolução dos resultados da pesquisa à sociedade

TEXTOS MANIPULÁVEIS E BUSCAS AUTOMÁTICAS

O CE-DOHS tem todo o seu acervo em formato de texto computacionalmente manipulável, em linguagem XML, que permita recuperar informações gráficas de cunho filológico dos documentos originais, além disso, gerar bases anotadas (morfológica e sintática) para análise linguística. Ou seja, um *corpus* que permita usos de mecanismos que gerem versões diversas acessíveis em processamentos de buscas automáticas, conforme visualizações anexas e uso efetivo no site do projeto em www.uefs.br/cedohs. Essa geração automática de versões de edições, sem perder as informações filológicas de fundamental importância para os estudos em linguística histórica, é feita via utilização de linguagem XML, como dito e convertida através da ferramenta desenvolvida especialmente para esse fim, o e-dictor. O *e-dictor* (http://oncoto.dyndns.org:44880/projects/edictor), desenvolvido como parte de desdobramentos do *Corpus Histórico do Português Tycho Brahe/CTB*, por

Kepler e Paixão e Souza (2004) e implementada por Fábio Kepler (USP) e Pablo Faria (UNICAMP), a partir de 2009. Essa ferramenta permite não só a edição em XML de forma fácil e rápida como também a aplicação de outras ferramentas desenvolvidas pelo projeto, tais como: correção de etiquetação automática e correção do *parser* automático, entre outras.

A geração da edição técnica permitirá também usos mais avançados de busca automática de dados através do corpus de buscas, através do convênio com a Unicamp, Convênio Guarda-Chuva entre a UEFS e a Unicamp e o Termo Aditivo de Transferência mútua de tecnologia, através do Corpus Histórico do Português Tycho Brahe/CTB (www.tycho.iel.unicamp.br), já citado e também com parcerias como projetos regionais. Isso tem sido feito através da parceria com o projeto *Corpora digitais para a história do português brasileiro: Documentos históricos da região do Sudoeste da Bahia, aliança PHPB-TYCHO BRAHE* (FAPESB 6171/ 2010).

O *corpus* permite a geração de catálogos eletrônicos de *corpus* serial (século XVII-XXI) de toda a região do semiárido baiano, além de da geração computacional automática de léxicos das edições em XML, catálogos dos documentos, edições variadas (conservadoras, modernizadas, etc.), além de outros subprodutos, todos gerados pela transformação XML.

DISPONIBILIZAÇÃO NA REDE

O CE-DOHS está disponível na rede mundial em www.uefs.br/cedohs e para geração de edições em www.tycho.iel.unicamp.br, também indicado nessa página ao lado de outros corpora brasileiros (Rio de Janeiro/UFRJ), Estados Unidos/Pensilvânia/UPENN, Finlândia e Portugal.

Cabe salientar que esse projeto foi possível graças ao piloto desenvolvido no âmbito do meu pós-doutoramento na Unicamp, acessado em http://www.uefs.br/dohs/corpusxml. html, e cujos dados eletrônicos vão dar origem a um livro organizado por membros da UEFS/UFBA e da UNICAMP (produto do projeto Vozes CNPq), já usando a busca eletrônica, um procedimento inédito que até agora somente vem sendo utilizado em teses e artigos, um banco que será referência no estado da Bahia, pois conterá milhões de palavras a serem acessadas em segundos na rede e já com dados quantificados exemplos separados computacionalmente.

Essa parceria já foi implementada com a FAPESP, via o projeto sob a coordenação da Professora Charlotte Galves da Unicamp, **A língua portuguesa no tempo e no espaço: contato linguístico, gramáticas em competição e mudança paramétrica/***Portuguese in time and space: linguistic contact, grammars in competition and parametric change* (FAPESP, 2012-2016), e coordenadora do Corpus Histórico do Português Tycho Brahe/CTB (cf. o banco do português europeu pode ser acessado em www.tycho.iel.unicamp.br, mediante senha obtida automaticamente depois de cadastro controle).

Cabe salientar que, conforme previsto, priorizamos os textos inéditos da região semiárida que estão sendo objeto de exploração de nossa equipe baiana (livros, teses, artigos), e também de diversos pesquisadores no Brasil e em outras partes do mundo. Conforme dito, isso é importante para contribuir para o

conhecimento de dados de uma região pouquíssima estudada do ponto de vista linguístico, o semiárido baiano, mas que tem uma história de contato única entre indígenas, brancos e africanos ao longo de todo o período colonial e com repercussões no nosso sistema educacional que se baseia em uma situação de diglossia linguística.

2.7 Bolsas – Esse item só deverá ser preenchido no caso de projetos com Bolsas FAPESB vinculadas.

2.7.1 Panorama de bolsas vinculadas ao projeto *(Informar o número total de bolsas previstas no projeto e a relação das já implementadas no período de que trata o presente relatório,, contendo: o nome dos bolsistas, a modalidade da bolsa e o período da vigência das bolsas implantadas).*

Nome do bolsista	Modalidade da bolsa	Vigência	Entrega do relatório FINAL de bolsa		Status do termo de outorga (rescindido,finalizado, substituído)
			Sim	*Não*	
Amanda Lopes de Souza Martins	I.C. Jr./ Referência/ UEFS	01/09/2011 -01/09/ 2012	X		FINALIZADO Obs.: A aluna continuou trabalhando como voluntária e atualmente presta vestibular pela UEFS
Igor Leal Souza	I.C. Computaçã o/Referênci a/UEFS	010 01/09/2011 - 01/09/2012	X		FINALIZADO Obs.: atualmente o bolsista permanece no Projeto como Pibic/CNPq
Maiany Soares de Oliveira	I.C. Referência/ UEFS	0 01/09/2011 - 01/09/2012	X		FINALIZADO Obs.: A aluna encontra-se matriculada como aluna especial no Programa de Estudos Linguísticos da UFBa/ Mestrado Nota 5

Justificar diferenças no projeto original, caso existam.
Não existem.

2.7.2 Análise geral da participação dos bolsistas

Relacionamento da coordenação e equipe com os bolsistas
Excelente [X] Bom [ ] Regular [ ] Insuficiente [ ] Não se aplica [ ]

Analisar a participação dos bolsistas em termos de sua formação e desenvolvimento do projeto durante o período abrangido pelo relatório.

Os bolsistas Amanda Lopes de Souza Martins, Igor Leal Souza e Maiany Soares de Oliveira desenvolveram um trabalho excepcional. Amanda Martins, I.C. Jr. teve uma participação assídua e contribuiu de forma excelente com o projeto, superando as expectativas. A bolsista Mayane Soares também teve excelente participação e continua contribuindo com o projeto, agora na condição de aluna especial do Mestrado em Linguística da UFBa e o bolsista Igor Leal Souza, da área da computação vem desenvolvendo um trabalho

fundamental para o projeto, agora como bolsista PIBIC.
Cabe salientar que participaram do projeto 12 bolsistas (Probic, Fapesb e Pibic, Pevic e voluntários) todos com excepcional contribuição ao projeto e a execução do acervo em formato digital, participando ativamente de atividades do projeto, nas oficinas e eventos nacionais e internacionais, conforme consulta aos anexos.

2.8 Coordenação e Gerenciamento

Comentar as atividades de coordenação de gerenciamento do projeto, incluindo observações sobre as parcerias estabelecidas.

ATENÇÃO: Consultar ANEXO do QUEM SOMOS referentes às parcerias.

A equipe da instituição executora reúne-se quinzenalmente, planejando e distribuindo entre os seus membros as atividades do projeto, discutindo a metodologia de trabalho e fazendo o relatório das atividades já realizadas. Mantem com as equipes parceiras da UNICAMP, da UFBA e da UESB contato permanente, discutindo sobre o desenvolvimento do projeto, avaliando os resultados obtidos e programando e organizando novos encontros de formação.

3. PRODUÇÃO GERADA ATRAVÉS DO DESENVOLVIMENTO DA PESQUISA *(Trabalhos da equipe executora aprovada pela FAPESB, individuais ou em cooperação, submetidos e/ou publicados, relativos à pesquisa apoiada, no período de abrangência deste relatório).*

Quantificar a produção de:
Relatórios/notas técnicas [3] Anais [em publicação [5]
Livros e Capítulos de livros [11]
Trabalhos apresentados em eventos científicos [6]
Cf. anexos

4. PARTICIPAÇÃO EM EVENTOS RELACIONADA AO DESENVOLVIMENTO DA PESQUISA *(Participação dos membros da equipe executora da pesquisa aprovada pela FAPESB, individual ou em grupo em eventos de formação e/ou de natureza científica, tecnológica e/ou de inovação, ou ainda em atividades de extensão no período de abrangência deste relatório).*

Quantificar a participação em:
Eventos científicos e/ou tecnológicos [9]
Cursos, workshops ou outras atividades de formação [3]
Outros (especificar) - ***I Oficina de Corpora Bahia*** **(UEFS, UFBA, UESB**), **II Oficina de Linguística de Corpus da Bahia**
Conferência (UEFS/UFBA/UNICAMP) – Internet

5. PARECER DO COORDENADOR DO PROJETO

Classificação de desempenho da equipe executora
Excelente [X] Bom [ ] Regular [ ] Insuficiente [ ]

Apreciar o desempenho da equipe executora.

A equipe executora do Projeto realizou as atividades previstas no plano de

trabalho, tendo resultados bastante significativos.
I Oficina de Corpora Bahia (UEFS, UFBA, UESB) com a participação de representantes dessas três universidades (cf. Primeiro relatório parcial)
Coordenou *a II Oficina de Linguística de Corpus da Bahia*, no período de 3 a 7 de outubro de 2011, com a participação do professor Pablo Faria (UNICAMP), que ministrou para o grupo do CE-DOHS – professores e bolsistas – o minicurso intitulado *Construção de corpora anotados usando o E-dictor* e prestou serviço de consultoria relativa ao sítio-web do projeto piloto final. Treinados pelo professor Pablo Faria e sob a orientação da equipe executora, havendo reuniões semanais no projeto, os bolsistas vêm desenvolvendo, a contento, as atividades de edição de textos no formato XML, usando o E-dictor.
- Participou em co-coordenação do encontro com Charlotte Galves, coordenadora do *Corpus Histórico do Português Tycho Brahe/CTB,* projeto que subsidia o Projeto CE-DOHS entre os dias 14 e 15 na UFBA, contando com a participação de todos os bolsistas.
Foram significativas as participações dos alunos em eventos regionais, nacionais e internacionais.

VER ANEXOS.

Avalie (considerando o período tratado por este relatório)

Infra-estrutura da Instituição onde está sendo desenvolvido o projeto Excelente [ ] Boa [X] Regular [ ] Insuficiente [ ]
Relacionamento com a equipe executora da pesquisa Excelente [X] Bom [ ] Regular [ ] Insuficiente [ ]
Quantidade e qualidade do trabalho desenvolvido Excelente [X] Boa [ ] Regular [ ] Insuficiente [ ]

Descrição e avaliação do apoio institucional recebido no período
A instituição, dentro de suas possibilidades, tem apoiado a pesquisa desenvolvida no âmbito do CE-DOHS. Foram ao todo 3 oficinas, a primeira descrita no relatório parcial, ocorrida com a participação da UFBa e Uneb, em dezembro de 2010, na Feira do Semiárido na UEFs, a *II Oficina de Linguística de Corpus da Bahia*, promovida pelo CE-DOHS no período de 3 a 7 de outubro de 2011, em que a equipe executora do projeto contou com a colaboração da Direção do Departamento de Letras e Artes (DLA), que nos disponibilizou o Laboratório de Informática, onde aconteceu o minicurso intitulado *Construção de corpora anotados usando o E-dictor*, ministrado pelo professor Pablo Faria (UNICAMP). Durante a Oficina, o projeto contou também com a colaboração do Setor de Informática, que auxiliou na instalação de programas, bem como com o Setor de Transportes, que colocou um carro e motorista à disposição do CE-DOHS e do professor Pablo Faria. A UEFS financiou também parte da alimentação do professor Pablo Faria, no período da Oficina. E a III com a participação da Professora Charlotte Galves, da Unicamp, na UFBa, com o apoio do PROHPOR, via Professora Tânia Lobo, da UFBa, contando com a participação de todos os bolsistas teve apoio da UEFS no transporte dos mesmos.

Avalie de maneira geral

Infra-estrutura da Instituição onde está sendo desenvolvida o projeto
Excelente [ ] Boa [X] Regular [ ] Insuficiente [ ]
Relacionamento com a equipe executora da pesquisa
Excelente [X] Bom [ ] Regular [ ] Insuficiente [ ]
Quantidade e qualidade do trabalho desenvolvido
Excelente [X] Boa [ ] Regular [ ] Insuficiente [ ]

Local /Data _____	Coordenador: _____ (Nome do Coordenador do Projeto)

Anexo a este relatório devem constar os seguintes documentos:

1 - Ofício de encaminhamento do Relatório Final à FAPESB. OK

2 - Relação de documentos entregues.OK ANEXO

3 - Cópia dos certificados de apresentação de membros da equipe executora do projeto em eventos científicos e/ou tecnológicos (desde que relacionados à pesquisa apoiada) . OK ANEXO

4 - Cópia dos certificados de participação de membros da equipe executora em atividades de extensão e outras, desde que relacionadas ao desenvolvimento desta pesquisa. OK ANEXO

5 - Lista dos trabalhos preparados ou submetidos (e ainda não aceitos) para publicação, acompanhada de cópias deste trabalho. OK ANEXO

6 - Cópia das primeiras páginas dos trabalhos científicos publicados individualmente ou por membros da equipe executora, ainda não enviados anteriormente, desde que relacionados ao projeto apoiado.

OBS.: Para encaminhamento de artigos elaborados pela equipe executora do projeto aprovado pela FAPESB, a tabela que compõe o Anexo I deste formulário deverá ser preenchida. Para encaminhamento dos certificados de participação em eventos dos membros da equipe executora do projeto aprovada pela FAPESB, a tabela que compõe o Anexo II deste formulário deverá ser preenchida.

7 - Fotos das ações desenvolvidas, com legenda, quando couber.

8- Sumário executivo do Projeto, com o objetivo de demonstrar como se deu o desenvolvimento e subsequentes resultados da pesquisa. O Sumário Executivo poderá ser disponibilizado para consulta pública, no Portal Fapesb.

RELATÓRIO FINAL RESUMIDO

CE-DOHS CORPUS ELETRÔNICO DE DOCUMENTOS HISTÓRICOS DO SERTÃO (FAPESB 5566/2010) EDITAL REFERÊNCIA 2010

PROPONENTE
Zenaide de Oliveira Novais Carneiro Bolsista

ABRANGÊNCIA

09/11/2010 a 09/11/2012
Alterada para 16/02/2013 (Portaria 115/2012/DOE de 8/2/2012)

SUMÁRIO

RESUMO

O projeto CE-DOHS CORPUS ELETRÔNICO DE DOCUMENTOS HISTÓRICOS DO SERTÃO (FAPESB 5566/2010), EDITAL REFERÊNCIA 2010, acessível em www.uefs.br/cedohs e em www.tycho.iel.unicamp.br, define-se, como o próprio nome sugere, como um Portal Eletrônico, constituído por um banco elaborado em linguagem XML, formado por textos escritos em português no Brasil, em especial, na região do sertão baiano, entre os séculos XIX e XX (fase atual, ou fase I) e séculos XVII e XVIII (fase posterior, ou fase II). O CE-DOHS filia-se ao Projeto Matriz, o *Corpus Histórico do Português Tycho Brahe/CTB* (CNPq), disponível em (http://www.tycho.iel.unicamp.br/~tycho/ e com o projeto interdisciplinar *A língua portuguesa no tempo e no espaço: contato linguístico, gramáticas em competição e mudança paramétrica/Portuguese in time and space: linguistic contact, grammars in competition and parametric change (FAPESP, 2012-2016*), desenvolvidos no Instituto de Linguagem da Universidade Estadual de Campinas (UNICAMP), ambos sob a supervisão da Profª Drª Charlotte C. Galves. Essa parceria está formalizada através de Convênio Guarda-Chuva entre a UEFS e a Unicamp, com o propósito fundamental de desenvolver novas metodologias para formação de grandes bancos de dados para o processamento de texto para fins linguísticos. É importante salientar que o CE-DOHS, através dessa parceria com o TYCHO BRAHE CORPUS também se beneficia dos acordos com o projeto pioneiro *Penn Helsinki Parsed Corpus of Middle English*, (http://www.ling.upenn.edu/hist-corpora), coordenado por Anthony Kroch na Universidade da Pensilvânia e os seus afiliados. Através do projeto *Vozes do Sertão em Dados: história, povos e formação do português brasileiro (*CNPq. Número do processo: 401433/2009-9) CONSEPE - 27/08/2009 102/2009), uma vez que o CE-DOHS é a contraparte eletrônica do banco Documentos Históricos do Sertão ou DOHS, do projeto *Vozes do Sertão* citado acima, filia-se também a importantes projetos nacionais, como o *Programa para a História da Língua Portuguesa* (PROHPOR), criado pela Professora Rosa Virgínia Mattos e Silva, na Universidade Federal da Bahia (UFBA), especificamente em seu arco temporal da história do português brasileiro (PB). Integra o *Projeto Nacional Para a História do Português Brasileiro* (PHPB), coordenado por Ataliba de Castilho, da Universidade de São Paulo (USP) e da Universidade Estadual de Campinas (UNICAMP), via equipe baiana, ou PHPB-ba, coordenada por Tânia Conceição Freire Lobo, da UFBA e da Plataforma do PHPB-ba, por Zenaide de Oliveira Novais Carneiro. O PHPB nacional desenvolve agendas e metodologias de pesquisa compartilhadas com diversas universidades brasileiras e tem, entre os seus objetivos, ações metodológicas voltadas para a organização de *corpora* diacrônicos com variados tipos de textos manuscritos e impressos, que vêm servindo de base para a composição de uma *Plataforma de Corpora do PHPB,* a cargo de Afrânio Barbosa, da Universidade Federal do Rio de Janeiro (UFRJ), e de Marcelo Módulo, da USP. O Projeto *Vozes do Sertão em Dados* e o CE-DOHS partilham dessa agenda, por meio da prospecção, da edição de documentos e da formação de *corpora* representativos de demandas histórico-sociais da região semiárida baiana, com repercussões sobre o processo de formação histórica do PB, com amplo contato linguístico de populações de origem portuguesa, indígena e africana, bem como com projetos temáticos de análise linguística.

O projeto também tem feito outras parcerias com Co-cooperação com projetos regionais. Isso tem sido feito com o projeto *Corpora digitais para a história do português brasileiro: Documentos históricos da região do Sudoeste da Bahia, aliança PHPB-TYCHO BRAHE* (FAPESB 6171/ 2010), sob a coordenação de Jorge Viana Santos, da UESB. E, recentemente iniciou processo de acordo, ainda informal, com o Projeto da Brasiliana Digital/ Projeto Brasiliana Digital, USP-São Paulo coordenado por Maria Clara Paixão de Souza, uma das consultora do Corpus CE-DOHS, extraído da Biblioteca Mindlen, coordenado por Pedro Puntoni, USP, em processo de tramitação formal, firmado por membros recentes do Projeto CE-DOHS para digitalizações de acervos inéditos do semiárido baiano.

É importante consultar o **ANEXO QUEM SOMOS** para visualização das parcerias.

O objetivo central do banco é contribuir, de forma específica, mas não exclusivamente, para o estudo da história do português brasileiro, a partir de questões levantadas por pesquisadores da história do português, quais sejam:

1. Como se dá a formação do português brasileiro, do século XVII ao XX, desde o seu período inicial, na fase colonial, no contexto de geral de multilinguismo/multidialetalismo que caracteriza o Brasil nesse período: *o português europeu, as línguas gerais indígenas e o português geral brasileiro.* E, posteriormente, na fase de multilinguismo localizado que caracteriza os séculos subsequentes, com a generalização do português brasileiro, em suas vertentes cultas e não-cultas.
2. Explicar, após a conclusão das fases I e II, qual é a trajetória temporal do português brasileiro.
3. Identificar como e quando se dá a emergência do português brasileiro nos textos escritos no Brasil.

O CE-DOHS tem todo o seu acervo em formato de texto computacionalmente manipulável, em linguagem XML, que permiea recuperar informações gráficas de cunho filológico dos documentos originais, além disso, gerar bases anotadas (morfológica e sintática) para análise linguística. Ou seja, um *corpus* que permita usos de mecanismos que geram versões diversas acessíveis em processamentos de buscas automáticas, conforme visualizações anexas e uso efetivo no site do projeto em www.uefs.br/cedohs. Essa geração automática de versões de edições, sem perder as informações filológicas, de fundamental importância para os estudos em linguística histórica, é feita via utilização de linguagem XML, como dito e convertida através da ferramenta desenvolvida especialmente para esse fim, o e-dictor. O *e-dictor* (http://oncoto.dyndns.org:44880/projects/edictor) foi desenvolvido como parte de desdobramentos do *Corpus Histórico do Português Tycho Brahe/CTB*, por Kepler e Paixão e Souza (2004) e implementada por Fábio Kepler (USP) e Pablo Faria (UNICAMP), a partir de 2009. Essa ferramenta permite não só a edição em XML de forma fácil e rápida como também a aplicação de outras ferramentas desenvolvidas pelo projeto, tais como: correção de etiquetação automática e correção do *parser* automático, entre outras.

A geração da edição técnica permitirá também usos mais avançados de busca automática de dados através do corpus de buscas, através do convênio com a Unicamp, Convênio Guarda-Chuva entre a UEFS e a Unicamp e o Termo Aditivo de Transferência mútua de tecnologia, através do Corpus Histórico do Português Tycho Brahe (www.tycho.iel.unicamp.br), já citado e também com parcerias como projetos regionais. Isso tem sido feito através da parceria com o projeto *Corpora digitais para a história do português brasileiro: Documentos históricos da região do Sudoeste da Bahia, aliança PHPB-TYCHO BRAHE* (FAPESB 6171/ 2010).

Esse estágio buscou também consolidar uma parceria para troca de experiências, através da associação entre edição tradicional e edição computacional na construção desse piloto de *Corpus digital-eletrônico.* Também foi proposto a aplicação e uso da ferramenta integrada de anotação de corpus denominada *e-dictor* (http://oncoto.dyndns.org:44880/projects/edictor), desenvolvida como parte de desdobramentos do *Corpus Histórico do Português Tycho Brahe*, por Kepler e Paixão e Souza (2004) e implementada por Fábio Kepler (USP) e Pablo Faria (UNICAMP), a partir de 2009. Essa ferramenta permite não só a edição em XML de forma fácil e rápida como também a aplicação de outras ferramentas desenvolvidas pelo projeto, tais como: correção de etiquetação automática e correção do *parser* automático, entre outras.

BREVE FUNDAMENTAÇÃO TEÓRICA E METODOLÓGICA

A apresentação da base teórica está dividida em tópicos, como abaixo descritos, com o objetivo de melhor explicitar o que se constitui um banco eletrônico de dados linguísticos, seu uso e aplicações nas análises linguísticas. Essa divisão foi feita da seguinte forma:

1. Discussão sobre a questão da edição de documentos nos moldes filológicos tradicionais;
2. Apresentação do conceito de mudança linguística;
3. Discussão sobre a análise de fenômenos linguísticos e

1. Da edição de documentos nos moldes filológicos tradicionais

Em resumo, essa parte foi centrada no projeto *Para a História do Português/Prohpor*, na UFBA (www.prohpor.ufba.br) e no *Projeto Nacional Para a História do Português do Brasil/PHPB* do qual fazemos partes. A base empírica para a construção do nosso banco de texto é discutida e construída nas reuniões com esses projetos. São discussões que fomentam as bases teórico-metodológicas na construção de grandes *corpora,* de 1997 até os dias atuais[5]. Cabe salientar que participamos de todas as reuniões desse projeto e participamos ativamente na construção do corpus, especificamente na grande área do semi-árido. Portanto, são projetos que configuram as suas balisas na área da Linguística Histórica em

[5] Na última reunião realizada entre 31 de maio a 4 de junho de 2010, no âmbito do "Projeto Para a História do Português Brasileiro"/PHPB, na Paráiba, foi discutida também já a viabilidade de um corpus eletrônico.

três campos de atuação: construção material/*corpus, base histórica que fundamenta a composição do material e análise da mudança linguística*. As especificidades da Equipe da UEFS, que se volta ao estudo da língua portuguesa no semi-árido, bem como os produtos das fases de atuação, estão descritas no histórico do DOHS (www.uefs.br/dohs).

No caso dos estudos históricos do PB, alguns procedimentos têm orientado a formação dos *corpora*, a saber:

a) Separação de documentos produzidos por indivíduos que têm o português como primeira língua (L1) ou como segunda (L2) e, nesse caso, tanto em situações "regulares" quanto naquelas que resultante de transmissão linguística irregular;
b) Separação de documentos escritos por indivíduos sem contato prolongado com a escola (populares) ou com relativo contato (semi-popular e semi-cultos) daqueles com muito contato (cultos).

2 - Sobre as questões teóricas do conceito de mudança linguística e sobre a análise de fenômenos linguísticos

A orientação que sedimenta o olhar sobre o *corpus* é baseada na concepção de que a mudança ocorre durante o processo de aquisição de linguagem, nos moldes do é proposto no modelo de Princípios e Parâmetros (Chomsky, 1981). Para isso, assumimos a distinção entre língua-I e língua-E, proposta pelo autor. Desse modo, embora os estudos sejam feitos com dados de língua-E, o interesse é observar fenômenos que caracterizam a língua-I. Dentro dessa visão, as mudanças são tratadas como alterações paramétricas (Lightfoot, 1999, Kroch, 2001, entre outros). Essas alterações no PB, com consequências importantes para a mudança paramétrica no português como apontado nos trabalhos de Galves (1987, 2001) na linha dos trabalhos de autores como Tarallo (1989) e dos trabalhos organizados por Roberts e Kato (1993), entre outros. No âmbito da linguística histórica tem sido considerado fundamental o uso de dados quantificados que apontem as alterações nas frequências e sejam interpretados como indicativos de mudança. Desse modo, fazer estudos de Linguística Diacrônica nessa perspectiva requer o manuseio de um número bastante grande de dados.

3. Da formação de bancos digitais eletrônicos

Como dito no projeto, as análises quantitativas demandam o manuseio de dados extensos, lugar em que as contribuições de tecnologias computacionais são decisivas e fundamentais, como a desenvolvida pelo *Projeto Corpus Histórico do Português Tycho Brahe/CTB.* Esse tipo de banco de dados vem se mostrando uma tendência mundial. Elencamos os projetos que já trabalham com essa técnica: o pioneiro *Penn Helsinki Parsed Corpus of Middle English*, (http://www.ling.upenn.edu/hist-corpora), coordenado por Anthony Kroch na Universidade da Pensilvânia e os seus afiliados, o *York Helsinki Parsed Corpus of Old English Poetry*, por Susan Pintzuk e Leendert Plug, o *York Toronto Helsinki*

Parsed Corpus of Old English Prose, por Ann Taylor, Anthony Warner, Susan Pintzuk, Frank Beths, ambos na Universidade de York, o *Penn Helsinki Parsed Corpus of Early Middle English,* por Anthony Kroch e Beatriz Santorini na Universidade da Pensilvânia, e enfim o *Parsed Corpus of Early English Correspondence,* por Ann Taylor, Anthony Warner, Susan Pintzuk na Universidade de York, e por Terttu Nevalainen e Arja Nurmi na Universidade de Helsinki, além de outros como o projeto de *Corpus annoté syntaxiquement de textes de français (9è au 17è siècle),* por F. Martineau e Paul Hirschbuhler na Universidade de Ottawa, e o projeto *Corpus Dialetal Sintático (CordialSin),* por Ana Maria Martins, na Universidade de Lisboa. Defendeu-se, o uso da linguagem XML pelas vantagens que apresenta.

METODOLOGIA

A metodologia está dividida em dois planos: da análise dos dados e da relação como os bancos eletrônicos

Para descrição da dinâmica da mudança que tem como objetivo responder questões levantadas por pesquisadores da história do português quais sejam: - Como se dá a formação do português brasileiro, do século XVII ao XX, desde o seu período inicial, na fase colonial, no contexto de geral de multilinguismo/multidialetalismo que caracteriza o Brasil nesse período: *o português europeu, as línguas gerais indígenas e o português geral brasileiro.* E, posteriormente, na fase de multilinguismo localizado que caracteriza os séculos subsequentes, com a generalização do português brasileiro, em suas vertentes cultas e não-cultas; - Explicar, após a conclusão das fases I e II, qual é a trajetória temporal do português brasileiro e - Identificar como e quando se dá a emergência do português brasileiro nos textos escritos no Brasil, as análises serão feitas a partir das descrições quantificadas da evolução dos fenômenos no tempo, a partir dos dados extraídos da base de dados do CE-DOHS. O objetivo é descrever possíveis correlações entre propriedades morfossintáticas distintas da gramáticas ou gramáticas do português brasileiro e sua evolução no tempo, usando-se a noção de taxa constante de Kroch (1989) que se baseia em um modelo teórico de mudança dentro da teoria de princípios e parâmetros, segundo a qual a mudança está intimamente ligada à aquisição. Os parâmetros são definidos como a contraparte de uma capacidade biológica constante, a faculdade da linguagem (cf. Chomsky, 1981, 1988, Chomsky & Lasnik, 1991, Chomsky, 1995, entre outros). A língua-E, definida como o produto cotidianamente exteriorizado em situações de uso, ativa essa capacidade inata. A distinção entre língua-I e língua-E, fundamental para o programa de investigação da gramática gerativa, que tem como objeto de estudo a língua-I, ganha especial relevância nos estudos diacrônicos, uma vez que as mudanças definidas a partir dessa concepção de gramática são tratadas como alterações paramétricas. Essas alterações seriam decorrentes de falhas de transmissão lingüística durante o processo de aquisição da linguagem por crianças (língua materna ou L1), ou por adultos em situação de contato lingüístico (segunda língua ou L2). A concepção dessa teoria de gramática, construída segundo o pressuposto de que há princípios universais e princípios parametrizáveis responsáveis pela variação que

se observa de língua para língua, permitiu que a mudança adquirisse um novo enfoque dentro dessa teoria. A conclusão imediata dessa formulação é que a mudança se daria durante o processo de aquisição da linguagem. A relação entre mudança paramétrica e aquisição da linguagem motivou diversas formulações que viessem a dar uma resposta adequada a um dos problemas cruciais dessa teoria: o que leva uma criança a marcar diferentemente dos seus pais, ou da geração anterior, os parâmetros da língua que lhe serviram de *input* (Lightfoot (1979, 1991, 1993, 1999), Kroch (1989, 1994, 2001) e Roberts (1993a e 1993b, Roberts & Holmberg, 2010; entre outros).

Com relação aos bancos eletrônicos, suas bases são fundamentais para análises de identificação temporal para determinar quando se inicia uma mudança, através do acesso integral aos textos históricos através de análises quantificadas fundamentais para o estudo de mudança dentro da teoria linguística assumida na análise do banco de dados.

A metodologia do Banco CE-DOHS baseia-se fundamentalmente na metodologia do Corpus Histórico do Português Tycho Brahe/CTB, composto por um corpus eletrônico anotado de textos em português escritos por autores nascidos entre 1435 e 1845, desenvolvido desde 1998, em cujas bases é possível fazer buscas automáticas de construções sintáticas. Na fase atual, o CTB pretende ampliar os textos sintaticamente anotados para 1.500.000 palavras em textos portugueses, 500.000 palavras em textos brasileiros, via entre outros projeto, via parceria com CE-DOHS, além de 300.000 palavras em documentos africanos, e da elaboração de analisador sintático automático (parser) para o português, que ampliará a base anotada sintaticamente e no qual o CE-DOHS se beneficiará através de buscas dos textos do projeto na base do CTB. A evolução da maturidade metodológica desse projeto pode ser vista no site: http://www.tycho.iel.unicamp.br/~tycho/corpus/, onde estão definidos as ferramentas e modelos seguidos.

O CE-DOHS vem ajudando na exploração dessa experiência pioneira realizada na conversão de edição em Word em linguagem XML, através da ferramenta e-dictor (cf http://oncoto.dyndns.org:44880/projects/edictor), desenvolvido por Faria, Kepler e Paixão e Souza (2004-2009). As diretrizes na anotação dos documentos são as seguintes, de acordo Namiuti e Santos (2009): a catalogação dos textos; a transcrição dos textos; a codificação da interferência editorial sobre os textos; (iv) a apresentação dos textos. Por fim, Galves (2004) resume bem as vantagens desse sistema para um “Corpus Digital Eletrônico”: a) um *melhor gerenciamento* de arquivos do *corpus*; b) a *otimização dos processos* que levam às anotações morfológica e sintática; c) a *ampliação da finalidade* do corpus, explorando as potencialidades dos textos ortograficamente transcritos e d) 4. a *padronizacão* do *corpus* de modo a poder inseri-lo em catálogos internacionais. Para facilitar a transformação em XML está sendo usada a Ferramenta Integrada de anotação de corpus, e-dictor, como já dito. Em que consiste essa ferramenta? O *e-dictor* destina-se a transcrição e codificação de textos em formato XML para análises linguísticas (morfológica, sintática, entre outras), acessível em http://oncoto.dyndns.org:44880/projects/edictor, na linguagem de programação Python, com código-fonte aberto a ser disponibilizado à comunidade. O objetivo dos autores, ao adotar o padrão XML, deveu-se a necessidade de se abarcar informações de edição, de etiquetagem (morfológica),

além de parte1 do *layout* do texto original (títulos, quebras de linha, página, etc.). Essa ferramenta permite entrada de texto puro a ser convertido para uma estrutura XML; manipulação das principais informações XML: páginas, parágrafos, sentenças e palavras; manipulação das principais edições: opções <ed>, forma fonológica e morfológica; navegação por páginas; layout básico (para exibição das informações na tela); atalhos de teclado para tornar a edição mais eficiente; inserção de comentários de edição; manipulação de metadados; acesso aos demais elementos do texto (títulos, cabeçalho/rodapé, etc.); mais poder de formatação (quebra de seções, títulos e subtítulos, etc.); melhoras gerais no layout da tela da ferramenta; identificação visual dos limites dos elementos, como parágrafos, seções, etc. mais opções na barra de ferramentas, melhor acesso a informação contextual, durante a edição (propriedades de elementos, propriedades do documento, etc.) e melhorar a apresentação HTML do texto.

Na preparação dos textos - Tipos de Texto-Fonte e Tipos de Edição. Parte-se do texto-Fonte com grafia preservada (originais impressos e transcrições diplomáticas): e*dição Completa* que implica modernização controlada do texto fonte. Texto-Fonte com grafia modernizada (edições intermediárias, usadas na Fase I): *Edição Técnica* devido as dificuldades de processamento; e - 2. Versões Disponíveis - Versão transcrição do texto-fonte onde mostra a transcrição fidedigna em relação ao texto tomado como fonte (seguindo, portanto sua grafia: a grafia original dos originais impressos; ou a grafia modernizada pelo editor anterior). Versão Texto Editado em que mostra o texto com as interferências realizadas pela equipe do corpus (modernizações completas ou modificações técnicas, conforme o caso). Dois tipos de arquivos estão disponíveis nesses casos: arquivos .html para leitura e arquivos .txt sem formatação, para uso das ferramentas automáticas e busca de dados e, por fim, a versão glossário de edições que mostra uma lista das intervenções realizadas pela equipe do corpus, seja no caso de edições completas ou técnicas. Para correção dos textos há ferramentas disponíveis criada na Universidade da Pensilvânia por Beth Randall, também autora da ferramenta de busca *CorpusSearch*. *CorpusDraw* e *CorpusSearch* são distribuídos no mesmo pacote no endereço corpussearch.sourceforge.net. *CorpusDraw* (cf. relatório, 2007 http://www.tycho.iel.unicamp.br/~tycho/prfpml/fase2/relatorios/2007/inicio.pdf). O manual para uso do sistema de anotação morfologia e sintática pode ser cessada em http://www.tycho.iel.unicamp.br/~tycho/corpus/manual/index.html.

Atualmente, o CTB tem investido na elaboração de novas ferramentas e melhoria das já existentes como define Charlotte Galves (2012) e cita os principais tipos de ferramentas necessárias ao desenvolvimento do Corpus e seu uso para pesquisa linguística:

- ***ferramentas de edição eletrônica***. Um importante produto tecnológico do projeto temático *Padrões rítmicos, fixação paramétrica e Mudança Linguistica Fase II* foi a ferramenta *E-dictor*, planejada por Maria Clara Paixão de Sousa durante seu pós-doutorado no âmbito do projeto, e implementada por Fábio Kepler, então doutorando no Departamento de Computação da USP e Pablo Faria, bolsista TT4 do projeto. Essa ferramenta está sendo usada na construção do Corpus Tycho Brahe mas também por vários grupos envolvidos na construção de corpora históricos (Vozes do Sertão em Dados, banco CE-DOHS na UEFS, Memória Conquistense, na UESB, PHPB-Rio de Janeiro, UFRJ). A ferramenta permite que a

anotação morfológica seja efetuada no mesmo arquivo. Ela deve ser melhorada no sentido de permitir mais funcionalidades, como a inserção de imagens e a anotação sintática dentro do arquivo. Esse ultimo ponto representará um avanço importante na integração dos aspectos filológicos e sintáticos do Corpus, uma vez que, em estando reunidos num mesmo arquivo todos os passos da anotação e codificação dos textos, todas as combinações se tornam possíveis, e a anotação sintática pode ser disponibilizada com todos os níveis de edição, inclusive os mais próximos dos textos originais.

- **ferramentas de anotação morfológica e sintática**. O etiquetador morfológico foi elaborado no decorrer do primeiro projeto temático, graças à participação decisiva de Marcelo Finger, do Departamento de Computação da USP, na elaboração inicial do Corpus Tycho Brahe. Temos atualmente uma segunda versão, de Fábio Kepler. Para a anotação sintática, utilisamos o analisador sintático automático ("parser") universal elaborado por Dan Bickel na Universidade da Pensilvânia, e treinado com nossos dados. É parte importante deste projeto a elaboração de um novo "parser", específico para a língua portuguesa e mais eficiente do que o que usamos atualmente, graças à colaboração com Marcelo Finger da do Departamento de Computação da USP, Fábio Kepler agora professor na Universidade Federal do Pampa, e Jesus Garcia e Veronica Gonzalez Lopez, do Departamento de Estatística da Unicamp. A originalidade do projeto consiste na articulação de métodos computacionais novos com algoritmos probabilísticos de análise da cadeia sintática.
- **ferramentas de buscas**. O trabalho com o Corpus conta com duas ferramentas essenciais, Corpus Draw e Corpus Search, (cf. http://corpussearch.sourceforge.net/) elaboradas por Beth Randall no âmbito do projeto do *Penn-Helsinki Parsed Corpus of Middle English*, coordenado por Anthony Kroch na Universidade da Pensilvânia (cf. http://www.ling.upenn.edu/hist-corpora/annotation/index.htm). Corpus Draw permite agilizar a fase de revisão da saída do analisador, e Corpus Search é a ferramenta de busca com a qual se fazem as buscas de estruturas sintáticas nos textos anotados. Uma versão para buscas em textos morfologicamente anotados já está disponível *on line* no site do *Corpus Tycho Brahe*. Uma interface deverá ser elaborada no decorrer do projeto para possibilitar buscas on line em textos sintaticamente anotados. Tanto para o desenvolvimento do analisador novo quanto para a procura de interfaces mais eficientes e amigáveis no uso das ferramentas, ou ainda para o estabelecimento de anotações sintáticas mais adequadas, a parceria com Anthony Kroch e sua equipe, em particular Beatrice Santorini, continuará fundamental neste projeto como foi nos projetos anteriores. (Trecho retirado do projeto filiado *A língua portuguesa no tempo e no espaço: contato linguístico, gramáticas em competição e mudança paramétrica/Portuguese in time and space: linguistic contact, grammars in competition and parametric change (FAPESP, 2012-2016*))

A nova implementação do termo (2012-2016) entre o CE-DOHS e o CTB, em execução se beneficiará do desenvolvimentos dessas ferramentas, primordiais para o acesso aos textos que subsidiarão as análises linguísticas, incluindo as previstas pelos projetos individuais e coletivos.

CONCLUSÃO

O projeto CE-DOHS CORPUS ELETRÔNICO DE DOCUMENTOS HISTÓRICOS DO SERTÃO foi bem sucedido na medida em que atingiu as principais metas propostas, como explicitado a seguir:

- Execução do Banco Eletrônico, um instrumento de pesquisa essencial para o estudo do português brasileiro, sobretudo em sua vertente baiana, do século XIX e XX e a perspectiva de consolidação de textos anteriores, dos séculos XVII e XVIII. Esse banco está acessível na rede mundial de computadores, disponibilizando textos na sua integralidade. Acessível em www.uefs.br/cedohs na parte de apresentação e no processamento em www.tycho.iel.unicamp.br, atendendo não somente a base prevista de cartas brasileiras, quanto de impressos, acervos da UFBa e com a perspectiva da inserção de textos raros e inéditos do século XVIII e XIX do semiárido baiano, além dos textos orais da coleção de Amostras da Língua Falada, do Projeto "Amostras da Língua Falada do Semiárido Baiano), www.uefs.br/nelp, com publicação financiada pela FAPESB, um projeto em que atuei como co-coordenadora, desde a sua fundação em 1993;

- O CE-DOHS foi pioneiro na Bahia em linguagem manipulável e com geração automática de edições para uso linguístico, fornecendo aos estudiosos do português brasileiro uma base sólida para a sua descrição temporal a partir de uma base escrita;

- Estabelecimento de parcerias com projetos **regionais**, **nacionais** e **i*nternacionais*** de referência, como o projeto matriz o *Corpus Histórico do Português Tycho Brahe/CTB* (CNPq) disponível em (http://www.tycho.iel.unicamp.br/~tycho/ e ao projeto interdisciplinar *A língua portuguesa no tempo e no espaço: contato linguístico, gramáticas em competição e mudança paramétrica/Portuguese in time and space: linguistic contact, grammars in competition and parametric change (FAPESP, 2012-2016*), desenvolvidos no Instituto de Linguagem da Universidade Estadual de Campinas (UNICAMP), ambos sob a supervisão da Profª Drª Charlotte C. Galves. Essa parceria está formalizada através de Convênio Guarda-Chuva entre a UEFS e a Unicamp, com o propósito fundamental desenvolver novas metodologias para formação de grandes bancos de dados para o processamento de texto para fins linguísticos. É importante salientar que o CE-DOHS, através dessa parceria com o TYCHO BRAHE CORPUS também se beneficia dos acordos com o projeto pioneiro *Penn Helsinki Parsed Corpus of Middle English*, (http://www.ling.upenn.edu/hist-corpora), coordenado por Anthony Kroch na Universidade da Pensilvânia e os seus afiliados. Cabe salientar que já há uso de tecnologia integradas como o *Corpus Dialetal Sintático (CordialSin),* por Ana Maria Martins, na Universidade de Lisboa, o que é importante para um uso comparativo entre o banco de dados orais e escritos com inquéritos e textos do Português Europeu como o banco CE-DOHS. Essa parceria já destacada por Charlotte Galves no texto do projeto *A língua portuguesa no tempo e no espaço: contato linguístico, gramáticas em competição e mudança paramétrica/Portuguese in time and space: linguistic contact (UNICAMP/FAPESP)* prevê essa comparabilidade do português brasileiro com o português europeu, uma vez que grande parte do acervo brasileiro é constituido de cartas (cf.por exemplo as *Cartas brasileiras* já presentes no Corpus – va_004). Através do projeto *Vozes do Sertão em Dados: história, povos e formação do português*

*brasileiro (*CNPq. Número do processo: 401433/2009-9) CONSEPE - 27/08/2009 102/2009), uma vez que o CE-DOHS é a contraparte eletrônica do banco Documentos Históricos do Sertão ou DOHS, do projeto *Vozes do Sertão* citado acima, filia-se também a importantes projetos nacionais, como o *Programa para a História da Língua Portuguesa* (PROHPOR), criado pela Professora Rosa Virgínia Mattos e Silva, na Universidade Federal da Bahia (UFBA), especificamente em seu arco temporal da história do português brasileiro (PB). Integra o *Projeto Nacional Para a História do Português Brasileiro* (PHPB), coordenado por Ataliba de Castilho, da Universidade de São Paulo (USP) e da Universidade Estadual de Campinas (UNICAMP), via equipe baiana, ou PHPB-ba, coordenada por Tânia Conceição Freire Lobo, da UFBA e da Plataforma do PHPB-ba, por Zenaide de Oliveira Novais Carneiro. O PHPB nacional desenvolve agendas e metodologias de pesquisa compartilhadas com diversas universidades brasileiras e tem, entre os seus objetivos, ações metodológicas voltadas para a organização de *corpora* diacrônicos com variados tipos de textos manuscritos e impressos, que vêm servindo de base para a composição de uma *Plataforma de Corpora do PHPB,* a cargo de Afrânio Barbosa, da Universidade Federal do Rio de Janeiro (UFRJ), e de Marcelo Módulo, da USP. O Projeto *Vozes do Sertão em Dados* e o CE-DOHS partilham dessa agenda, por meio da prospecção, da edição de documentos e da formação de *corpora* representativos de demandas histórico-sociais da região semiárida baiana, com repercussões sobre o processo de formação histórica do PB, com amplo contato linguístico de populações de origem portuguesa, indígena e africana, bem como com projetos temáticos de análise linguística.O projeto também tem feito outras parcerias com Co-cooperação com projetos regionais. Isso tem sido feito através da parceria com o projeto *Corpora digitais para a história do português brasileiro: Documentos históricos da região do Sudoeste da Bahia, aliança PHPB-TYCHO BRAHE* (FAPESB 6171/ 2010), sob a coordenação de Jorge Viana Santos, da UESB.E recentemente iniciou processo de acordo com a Brasiliana Digital/Biblioteca Mindlin, coordenada por Maria Clara Paixão de Souza, uma das consultora do Corpus CE-DOHS;

- Colaboração com a discussão para o melhoramento e a produção de ferramentas com o projeto matriz, o CTB;

- Formação de diversos bolsistas, projetos de dissertações, teses de doutorado, pós-doutorado e de projeto individuais, agregando diversos pesquisadores do Departamento de Letras e Artes.
- Além do reforço na internacionalização do Banco CE-DOHS, permitiu parcerias e convites para elaboração de diversos trabalhos acadêmicos.

- Perspectivas de Extensão do Banco do CE-DOHS e novos acordos entre projetos;

- A publicação de livro sobre o português brasileiro em parceria com diversas universidades;

- Colaboração na elaboração no novo projeto temático da Professora Charlotte Galves voltado para uma história comparada do português europeu e brasileiro intitulado *A língua portuguesa no tempo e no espaço: contato linguístico, gramáticas em competição e mudança paramétrica/Portuguese in time and space: linguistic contact, grammars in competition and parametric change (FAPESP, 2012-2016). Um* projeto que é a continuação natural dos anteriores, e redireciona o Corpus Tycho Brahe em duas direções: Corpus do Tycho Brahe Brasil);

- Colaboração na organização do Plataforma do PHPB-ba, cuja plataforma nacional do PHPB está a cargo de Afrânio Barbosa (UFRJ) e de Marcelo Módulo (USP), a ser sitiado no Museu de Língua Portuguesa, além das parcerias com a Brasiliana Digital em acordo informal, e colaboração na elaboração do BIT-Prohpor na UFBA;

- Participação dos bolsistas e membros do Projeto em eventos regionais, nacionais e internacionais;

- Convite para a Professora Zenaide de Oliveira Novais Carneiro representar a Área de Linguística Histórica juntamente com Tânia Lobo, no Congresso da Associação Brasileira de Linguística/ABRALIN 2013, em Natal, na UFRN.

CORPUS

Nomes das Coletâneas Publicação	**Número de volumes/ Tomos/CD-ROM**	**Discriminação**
COLETÂNEA MANUSCRITOS [6] **Cartas brasileiras (1809-2000):** coletânea de	3 volumes 7 tomos/CD-ROM **Obs.:** Orelhas de: **Volume 1** - Rosa	**VOLUME 1 (1809-1904)** - Zenaide de Oliveira Novais Carneiro (Organizadora) **CD-ROM 1.** *Cartas para vários destinatários (1809-1904):* edição fac-similada/Zenaide de Oliveira Novais Carneiro **CD-ROM 2.** *Cartas para Severino Vieira, governador da Bahia (1901-1902):* edição fac-

[6] A coletânea foi organizada de maneira a permitir que pesquisadores interessados na história do PB possam constituir *corpora* específicos de acordo com seus objetivos. Para isso, cada pesquisador poderá utilizar-se das informações sobre a documentação e sobre seus escreventes contidas na Coletânea. Do ponto de vista da documentação, é possível a sua organização por ordem cronológica, tanto por data de escrita, como é tradicional na Linguística Histórica, quanto por data de nascimento do autor (individual), no caso de autores que possuem acervos mais significativos, ou por grupos de autores, como tem sido feito na base de dados para estudos em Linguística Diacrônica, no âmbito do *Corpus Histórico do Português Tycho Brahe,* coordenado por Charlotte Galves, da UNICAMP. Desse modo, essa base documental pode ajudar nos estudos sobre variação e mudança do PB, do século 19 ao século 20, se se considerar a cronologia dos textos, por data de escrita das cartas, ou a partir do século 18, se a cronologia for feita por data de nascimento dos seus remetentes. As informações sobre o grau de escolaridade dos escreventes, aliadas a possíveis análises do tipo de escrita, permitem também a separação das cartas representativas de variedades *standard* e não-*standard* do PB, em uma perspectiva histórica.

fontes para o estudo do português (*1023 cartas manuscritas*) (*Edital Publicação Fapesb 1493/2010)* **UEFS Editora (http://uefseditora.com.br/)**	Virgínia Mattos e Silva **Volume 2** - Ataliba Castilho **Volume 3** - Afrânio Barbosa	similada/Zenaide de Oliveira Novais Carneiro **CD-ROM 3**. *Cartas para Cícero Dantas Martins, Barão de Jeremoabo (1880-1903):* edição fac-similada/Zenaide de Oliveira Novais Carneiro **VOLUME 2 (1902-1993) -** Zenaide de Oliveira Novais Carneiro; Mariana Fagundes de Oliveira; Norma Almeida (Organização) **CD-ROM 1.** *Cartas do acervo Dantas Jr. (1902-1962):* edição fac-similada/ Zenaide de Oliveira Novais Carneiro; Marta Carvalho Ferreira Lisboa; Mônica Araújo Cruz; Denise Branco Cerqueira; Eliane Santos Leite ***CD-ROM* 2**. *Cartas baianas: o acervo de João da Costa Pinto Victoria (1911-1958):* edição fac-similada/ Zenaide de Oliveira Novais Carneiro; Maria Rosane Passos; Priscila Tuy Batista; Anderléia Mascarenhas **CD-ROM 3**. *Correspondências amigas: o acervo de Valente, Bahia (1980-1993):* edição fac-similada/Mariana Fagundes de Oliveira; Maiany Soares de Oliveira; Adilson Silva de Jesus **VOLUME 3 (1906-2000)** Huda da Silva Santiago; Zenaide de Oliveira Novais Carneiro; Klebson Oliveira (Organização) **CD-ROM 1.** Cartas em Sisal: Riachão do Jacuípe, Conceição do Coité e Ichu (1906-2000): edição fac-similada/Huda da Silva Santiago
PUBLICA-SE EM FEIRA DE SANTANA (1901-2006)[7] *Das cartas de leitores e*		**Orgaizadoras:** Zenaide de Oliveira Novais Carneiro e Mariana Fagundes de Oliveira **CD-ROM 1.** Cartas de Leitores e Redatores em o Progresso e na Folha do Norte (1901-1997) **CD-ROM 2.** Anúncios na Folha do Norte (1910-

[7] Esta Coletânea é um produto do projeto *Vozes do Sertão em Dados: história, povos e formação do português brasileiro*, financiado pelo CNPq. Concretizando um dos objetivos específicos do *Vozes* – que tem por um dos seus objetivos gerais contribuir para a edição de documentos históricos do Semiárido baiano escritos em língua portuguesa, possibilitando a execução de pesquisas em estudos linguísticos temáticos sobre o português brasileiro (PB) –, apresentamos nessa Coletânea a edição fac-similada de documentos impressos baianos, da cidade de Feira de Santana: cartas de leitores e redatores dos jornais *O Progresso* e *Folha do Norte*, além de anúncios do *Folha do Norte*, publicados entre 1901 a 1997.

retadores e dos anúncios e o progresso e na Folha do Norte *(Edital Publicação Regular da UEFS Editora, 2012)* **UEFS Editora (http://uefseditora.com.br/)**		2006) **Editores:** Zenaide de Oliveira Novais Carneiro Mariana Fagundes de Oliveira Priscila Tuy Batista Huda da Silva Santiago Gilvânia da Silva Almeida Oliveira Tárcia Priscila Lima Dória Denise Branco Cerqueira Amanda Lopes de Souza Martins Janaina de Oliveira Costa Marinalda Silva Freitas Adilson Silva de Jesus Liliane de Jesus e Jesus Lorena Rosa Santos Luana Manuela Lima Silva Maiany Soares de Oliveira Mônica Araújo Cruz Sueli Meireles Conceição Tárcia Priscila Lima Dória

REFERÊNCIAS

BRITTO, Helena & FINGER, Marcelo (1999): *Constructing a parsed corpus of historical portuguese.* ACHALLC' 99 International Humanities Computing Conference, University of Virginia, Charlottesville, 9-14 de junho de 1999.

CARNEIRO, Z. O. N. (2008). Vozes do sertão em dados: história, povos e formação do português brasileiro. In: *VI Feira do Semi-Árido: desertificação, perspectivas de autonomia produtiva frente aos desafios socioambientais,* 2008, Feira de Santana. VI Feira do Semi-Árido: desertificação, perspectivas de autonomia produtiva frente aos desafios socioambientais. Feira de Santana, v. 1.

CARNEIRO, Zenaide & Charlotte GALVES. 2010. "Variação e Gramática: colocação de clíticos na história do português brasileiro"*. Revista de Estudos da Linguagem* 18, 1.

CARNEIRO, Zenaide de Oliveira Novais (2011). *Resultados sobre medição de letramento nos "Sertões de Dentro" a partir de livros cartoriais (1853-1908*). Apresentação no XVI Congresso Internacional da Associação de Linguística e Filologia da América Latina, Alcalá de Henares, Madrid.

CARNEIRO, Zenaide de Oliveira Novais (2012). *Lei de terras e ocupação privada Elementos para avaliação dos níveis de alfabetização no interior da Bahia oitocentista*. 2012. Apresentação em Mesa-Redonda no Castilho - II Congresso Internacional de Linguística Histórica. Estudos de História Social.

CARNEIRO, Zenaide Novais (2005). *Cartas brasileiras (1809-1907): um estudo filológico-linguístico.* Campinas: UNICAMP. Tese de doutorado inédita.

CARNEIRO, Z. O. N. (2008). Estudo de escolarização de aldeados no Brasil do século XVII: um caminho para a compilação de possíveis fontes escrita em português por. In: XV Congresso Internacional de La Asociación de Linguística y Filología de América Latina, 2008, Montevideo. Libro de resúmenes de XV Congresso Internacional de La Asociación de Linguística y Filología de América Latina/ALFAL,p. 263-263.

CARNEIRO, Zenaide Novais (2005). *Cartas brasileiras (1809-1907): um estudo filológico-linguístico.* Campinas: UNICAMP. Tese de doutorado inédita.

CES, 1996: Corpus Encoding Standard, Document CES 1 Version 1.2 . Em http://www.cs.vassar.edu/CES/CES1-1.html

CHOMSKY, Noan. (1986a). *Knowledge of language:* Its nature, origin and use. New York: Praeger.

CHOMSKY, Noan. (1988). *Language and problems of knowledge*. The Managua Lectures. Cambridge: MIT Press.
CHOMSKY, Noan. (1995). *A minimalist program*. Press. Massachusetts: Cambridge, MIT PRESS.

FINGER, Marcelo, 1998: "Tagging a morphologically rich language", *Proceedings of the first workshop on text, speech and dialogue* (TSD98), Brno, Tchecoslováquia.

FINGER, Marcelo, 2000: *Técnicas de otimização da precisão empregadas no etiquetador Tycho Brahe*, Propor 2000.

GALVES, C. (1987). A sintaxe do português brasileiro. *Ensaios de linguística,* p. 13:31-50.

GALVES, C. (2002). *Ensaios sobre gramáticas do português.* Campinas: Editora da UNICAMP.

GALVES, Charlotte, 2004: *Projeto Padrões Rítmicos, Fixação de Parâmetros e Mudança Gramatical, II* –FAPESP. GALVES, Charlotte (http://www.tycho.iel.unicamp.br/~tycho/prfpml/fase2/index.html) relatórios anuais.

GUEDES, Marymarcia & BERLINCK, Rosane Andrade (Orgs. 2000). E os preços eram commodos…Anúncios de jornais brasileiros do século XIX. São Paulo: Humanitas [Série Diachronica, vol. 2].

IDE, Nancy e Laurent Romary, 2003: *Outline of the international standard linguistic Annotation framework.* Proceedings of ACL'03 Workshop on Linguistic

Annotation: Getting the Model Right, Sapporo, 1-5. Em http://www.cs.vassar.edu/faculty/ide/pubs.html

ROBERTS, I; KATO, M. (Orgs). *Português brasileiro*: uma viagem diacrônica. Campinas: Editora da UNICAMP,

KROCH, A. (1989b). Reflexes of grammar in patterns of language change. *Language variation and change,* 1, p. 199-244.

KROCH, A. (1994). Morphosyntactic variation. In: BEALS, K. et al. (Eds.). *Papers from the 30th regional meeting of the Chicago linguistics society:* parasession on variation and linguistic theory, v. 2, p. 180-201.

KROCH, A. (2001). Syntactic change. In: BALTIN, M.; COLLINS, C. (eds.). *The handbook of contemporary syntactic theory*, Oxford: Blackwell Publishers Inc., p. 699-729.

LABOV, W. (1982). Building on empirical foundation. In: LEHAMANN, W. P.; MAKIEL, Y (eds.). *Perspectives on historical linguistics*. Philadelphia: John Benjamins Publishing Company.

LAPESA, Rafael (1993). Sobre los orígenes y evolución del leísmo, laísmo y loísmo. In: FERNÁNDEZ, Soriano (Ed.). (1993). *Los pronombres átonos*. Madrid, Taurus, Alfagurara, S.A.

LIGHTFOOT, David. (1979). *Principles of diachronic syntax*. Cambridge: The MIT Press.

LIGHTFOOT, David. (1991). *How to set parameters. Arguments from language change.* Cambridge: MIT Press.

LIGHTFOOT, David. (1999). *The development of language:* Acquisition, change, and evolution. Maryland lectures in language and cognition. Malden.Blackwell.
LABOV, W. (1982). Building on empirical foudation. In: LEHAMANN, W. P.;
MAKIEL, Y (eds.). *Perspectives on historical linguistics*. Philadelphia: John Benjamins Publishing Company.
LEITE, S.J., Serafim (1938-1950). *História da Companhia de Jesus no Brasil*, 10 vols. Lisboa e Rio de Janeiro: Portugália/Civilização Brasileira.
LIGHTFOOT, David. (1999). *The development of language:* Acquisition, change, and evolution. Maryland lectures in language and cognition. Malden.Blackwell.
LOBO, Tânia / RIBEIRO, Ilza Ribeiro / CARNEIRO, Zenaide / ALMEIDA, Norma (Orgs. 2006). Para a história do português brasileiro: novos dados, novas análises. Salvador: Editora da Universidade Federal da Bahia, vol. VI, 2 tomos.
LOBO, Tânia; MATTOS E SILVA, Rosa Virgínia; VENÂNCIO, Américo L. M. Filho (2006). Indícios de uma língua geral no sul da Bahia na segunda metade do século XVIII. In: LOBO, Tânia; RIBEIRO, Ilza; CARNEIRO, Zenaide de O. N.; ALMEIDA, Norma Lucia F. de. *Para a história do português brasileiro*: *novos dados, novas análises*. Salvador: Edufba, vol. 6, 1-2, p. 609-630.

LOBO, Tânia (2003). A questão da periodização da história do Brasil. In: CASTRO, Ivo; DUARTE, Inês (orgs.). *Razão e emoção: miscelânea de estudos em homenagem a Maria Helena Mateus*. Lisboa: Imprensa Nacional- Casa da Moeda, p.395 et.passim.

MARQUILHAS, Rita, (2001). *A faculdade das letras*, Imprensa Nacional Casa da Moeda, Lisboa.

MATTOS E SILVA, Rosa Virgínia (2008). *Caminhos da Linguística histórica: ouvir o inaudível,* São Paulo: Parábola Editorial.

MATTOS E SILVA, Rosa Virgínia (2001). De fontes sócio-históricas para a história social linguística do Brasil: em busca de indícios. In: MATTOS E SILVA, Rosa Virgínia (org.). *Para a história do português brasileiro: primeiros estudos*. São Paulo: Humanitas/FFCHL/USP:FAPESP, v.2, t. 2, p. 275-302.

MATTOS E SILVA, Rosa Virgínia; OLIVEIRA, Klebson; LOBO, Tânia (2007). Panorama preliminar do letramento de negros na Bahia. In: RAMOS, J.; ALKMIM, Mônica A. *Para a história do português brasileiro*: *estudos sobre mudança linguística e história social*. Belo Horizonte: Faculdade de Letras da UFMG, v.5. p. 373-442.

MATTOS E SILVA, Rosa Virgínia (2002). Para a história do português culto e popular brasileiro: sugestões para uma pauta de pesquisa. In: ALKMIM, Tânia M. *Para a história do português brasileiro*: *novos estudos*. São Paulo: Humanitas/FFCHL/USP:FAPESP, v. 2, p. 443-464.

MATTOS E SILVA, Rosa Virgínia (Org.2001). Para a História do Português Brasileiro, vol. II, Primeiros Estudos, 2 tomos. São Paulo: Humanitas / Fapesp.

MONTEIRO, John M. (2001). *Tupis, tapuias e historiadores*: *estudos de história indígena e do indigenismo*. Campinas: IFCH/UNICAMP. Tese de livre docência.

Kroch, Anthony. 1994. “Morpho-syntactic variation”, in Kenneth Beals et al. (eds.), *Papers from the 30th Regional Meeting of the Chicago Linguistics Society: Parasession on Variation e Linguistic Theory,* vol. 2, pp. 180-201.

Kroch, Anthony. 2002. “Syntactic change”, in Mark Baltin & Cris Collins (eds.) *The handbook of contemporary syntactic theory*. Oxford: Blackwell, 699-729.

Lightfoot, David. 1999. *The development of language: evolution, change and acquisition*. Oxford: Blackwell.

Lightfoot, David. 2006. *How new languages emerge*. Cambridge: Cambridge University Press.

Lobo, Tânia & Klebson Oliveira. 2009. *África à vista*, Salvador: EDUFBa.

Longobardi, Giuseppe & Cristina Guardiano. 2009. “Evidence for syntax as a signal of historical relatedness”. *Lingua*, 229, spec. issue *The Forests behind the trees*, ed. By John Nerbonne, 1679-1706.

Lucchesi, Dante. 2003. “O conceito de transmissão linguística irregular e o processo de formação do português do Brasil”, in Roncarati, Claudia; Abraçado, Jussara. (orgs) *Português brasileiro, contato linguístico, heterogeneidade e história*, Rio de Janeiro: Viveiros de Castro Editora, pp. 272-282.

ROBERTS, Ian. (1992a). *Object movement and verb movement in early modern english*. University of Wales, Bangor.

ROBERTS, Ian. (1993a). *Verbs and diachronic syntax.* Dordrecht: Kluwer.

Roberts, Ian. a sair "Macroparameters and Minimalism: A Programme for Comparative Research" in Galves et al., a sair.
Roberts, Ian & Anders Holmberg. 2010. "Introduction: parameters in minimalist theory", in Biberauer et al., 1-57.

PAIXÃO DE SOUSA, M.C., CAVALCANTE, S.R.O., NAMIUTI, C.. *Linguística de Corpus e História da Língua Portuguesa: Propostas, Resultados e Desafios.* Mesa Redonda. V Congresso Internacional da ABRALIN. 2007.*História da Língua Portuguesa: Propostas, Resultados e Desafios.* Mesa Redonda. V Congresso Internacional da ABRALIN. 2007.

PAIXÃO DE SOUSA, M.C. *Memórias do Texto*. Revista Texto Digital. n. 2. Universidade Fedral de Santa Catarina. 2006.
PAIXÃO DE SOUSA, Maria Clara. *Projeto Memórias do Texto*. FAPESP-UNICAMP, 2004.
PAIXÃO DE SOUSA, Maria Clara e Thorsten Trippel (2004): *Single source processing of historic corpora for diverse uses* ALLC/ACH 2004, proceeding.
PROJETO *CE-DOHS: Corpus eletrônico de documentos históricos do sertão*. Coordenação: Zenaide de Oliveira Novais Carneiro; Mariana Fagundes de Oliveira Disponível em: www.uefs.br/cedohs. 2011.
PROJETO *Vozes do sertão em dados: história, povos e formação do português brasileiro*. FASE 1- Coordenação: Zenaide de Oliveira Novais Carneiro. Disponível: www.uefs.br/nelp. 2011.·
RAMOS, Jânia / ALCKMIN, Mônica A. (Orgs. 2007). Para a História do Português Brasileiro, vol. V: Estudos sobre mudança linguística e história social. Belo Horizonte: Faculdade de Letras da Universidade Federal de Minas Gerais.
SAXON. <http://saxon.sourceforge.net/>
TYCHO BRAHE. <http://www.tycho.iel.unicamp.br/~tycho/corpus/index.html>
TRIPPEL, Thorsten and PAIXÃO DE SOUSA, Maria Clara (2006). "Metadata and XML standards at work: a corpus repository of Historical Portuguese texts". *Papers from the V International Conference on Language Resources and Evaluation (LREC 2006).*
W3C (1997). "Extensible Markup Language". http://www.w3.org/XML

OFÍCIO

Feira de Santana-Ba, 30 de janeiro de 2013.

À Diretoria Geral
Fundação de Amparo à Pesquisa do Estado da Bahia
Rua Aristides Novis, 203, Federação
40.210-720 Salvador Bahia

Ref.: RELATÓRIO FINAL - TERMO DE OUTORGA Nº PET0028/2010 - EDITAL 019/2010 – PEDIDO 5566/2010
OBS.: ACOMPANHA ANEXOS E RELATÓRIO FINANCEIRO

Prezados Senhores,

Encaminhamos, para os devidos fins, ENVIO DE RELATÓRIO FINAL do projeto apoiado por esta Fundação, como detalhado a seguir.

Outorgado/Convenente: ZENAIDE DE OLIVEIRA NOVAIS CARNEIRO

Coordenadora: ZENAIDE DE OLIVEIRA NOVAIS CARNEIRO **Vice-Coordenadora: MARIANA FAGUNDES DE OLIVEIRA**

Título do Projeto: CE-DOHS CORPUS ELETRONICOS DE DOC.HIST.	**Pedido Nº: 5566/2010**	**Telefone:** (75)3224-9023

Termo de Outorga: PET0028/2010	**Convênio:**
Data de Assinatura do Termo de Outorga: 09/11/2010	
Vigência: 09/11/2010 a 09/11/2012. Alterada para 16/02/2013 (Portaria 115/2012). DOE de 8/2/2012). **Projeto concluído no período de vigência indicado no T.O? SIM (X) NÃO ( )**	

Período de vigência do instrumento legal: 08/11/2010 ATÉ 08/11/2012

Atenciosamente,

Zenaide de Oliveira Novais Carneiro
Professor Titular UEFS/DLA/Área de Linguística/Linguística Histórica
NELP - Núcleo de Estudos de Língua Portuguesa (Coord.)
Projeto CE-DOHS – CORPUS ELETRÔNICO DE DOCUMENTOS HISTÓRICOS DO SERTÃO (FAPESB5566/2010) CONSEPE: 17/11/2010 202/2010)/ **Coord.)DOHS**