FAPESB
Fundação de Amparo à Pesquisa do Estado da Bahia

EDITAL FAPESB AÇÃO REFERÊNCIA Pedido Nº 5566/2010	Nº PROTOCOLO

Dados do Proponente

Coordenador do Projeto: Zenaide de Oliveira Novais Carneiro

Titulação Máxima	Ano de Conclusão	Telefone	Celular	Email
Doutorado	2005			

Instituição do Proponente

Instituição/ Unidade/ Departamento

Universidade Estadual de Feira de Santana

Departamento de Letras e Artes

Natureza: UNIVERSIDADE ESTADUAL (UE)
CNPJ: 14.045.546/0001-73
Endereço: AVENIDA UNIVERSITÁRIA,S/N KM 03 DA BR 116
Complemento: CAMPUS UNIVERSITÁRIO
Bairro:
Cidade: Feira de Santana
Estado: BA
Telefone:
Fax:
Representante Legal: José Carlos Barreto de Santana
Cargo: Reitor

Identificação do Projeto

Titulo do Projeto: CE-DOHS - CORPUS ELETRÔNICO DE DOCUMENTOS HISTÓRICOS DO SERTÃO

Área do Conhecimento: Linguística, Letras e Artes

Sub-área do Conhecimento: Linguística

Obra-Referência: Portal

Palavras Chaves

Banco Eletrônico / Português Brasileiro / Sertão Baiano / Ling. Histórica

Resumo do Projeto

Esta pesquisa insere-se no contexto maior dos projetos Vozes do Sertão em Dados: história, povos e formação do português brasileiro (CNPq 401433/2009-9 (www.uefs.br/nelp). E nasce a partir desse projeto com o meu projeto de pós -doutoramento intitulado Proposta de um sistema de edições eletrônicas e seus aspectos tecnológicos: a construção de um Piloto de Corpus Eletrônico a ser implantado na UEFS para o estudo da língua portuguesa no semi-árido baiano (séculos XVII-XXI) (Fapesb 1648/2009). O projeto Referência, intitulado CE DOHS Corpus Eletrônico de Documentos Históricos do Sertão/BA, que estamos propondo, pretende, portanto, dá um formato eletrônico do nosso extenso banco de dados, resultado de uma longa prospecção de fontes inéditas do Sertão (www.uefs.br/dohs) e também complementar a parte que está faltando, aliando os dois formatos: o digital e o eletrônico, esse último para fins de pesquisa lingüística (que dará um retorno incomensurável na qualidade e rapidez de trabalhos acadêmicos (tanto do nosso

programa quanto de outros que desejem acessar o banco, mediante senha de controle) e o outro disponível ao grande público. O formato digital é o que está sendo feito no Vozes do Sertão, citado acima, e cujo resultado parcial será já publicado agora em 2011 (Edital Fapesb, publicação 2010). Ou seja, 1.000 cartas da Bahia editadas em formato tradicional e fac-similada segundo os critérios do Projeto Nacional Para a História do Português Brasileiro (cf. www.uefs.br/dohs/phpb.html) que agrega várias universidades brasileiras (UFBA, UEFS, UFAL, UFPB, UFRN, UFPE, UFMG, UFRJ, USP, UNICAMP, UFSC, UFRS, UEL, entre outras) e somos a chamada equipe baiana, na qual a UEFS está responsável pela quase totalidade área do semi-árido baiano (cf. http://www.uefs.br/dohs/historico.html), um projeto iniciado por Ilza Ribeiro (UEFS). Hoje estamos dividindo a pesquisa com a UESC (cf. http://www.uefs.br/dohs/equipe.html), que ficará responsável pela região Oeste e parte da Chapada Diamantina. O projeto Vozes (fase atual), por sua vez é filiado ao Prohpor Programa Para a História do Português (www.prohpor.ufba.br), coordenado por Rosa Virgínia Mattos e Silva, desde 1996. Então, por que estamos propondo o formato eletrônico? Por que apesar da importância do banco digital para o grande público, por tornar acessível a documentação rara e escassa do sertão baiano, contribuindo para a sua preservação em formato digital (fotos digitais e edição diplomática de todos os documentos), precisamos tornar esse mesmo material manipulável para fins de pesquisa. Nesse caso, o mesmo banco ganhará um novo formato, o eletrônico. E o mais importante, o transporte se dará de forma bastante rápida, porque já dispomos de programas de conversão. No nosso caso, o edictor (http://purl.org/edictor/), desenvolvido na USP e na UNICAMP. O importante é que acompanhamos a elaboração desse programa e fizemos intervenções ao longo da nossa participação no projeto Tycho Brahe desde 2000. Fizemos o pós-doc para fazer um piloto a ser aplicado ao banco baiano. Esse formato vai elevar o alcance do conteúdo do acervo constituído ao longo de 14 anos de pesquisa, tornando-o acessível para fins acadêmicos bem como da pesquisa em geral. A constituição do acervo digital se dará nos moldes do Projeto Corpus Histórico do Português Tycho Brahe (www.tycho.iel.unicamp.br), a partir da proposta do projeto de pós-doutorado citado acima. Toda a viabilização da transferência de tecnologia mútua está sendo feita através de convênios institucionais (UEFS/Unicamp), através da AERI, e já está em fase adiantada de tramitação, como um dos nossos produtos do projeto de pós-doc. A parte específica no que tange a este projeto consta de um termo aditivo já aprovado no Instituto da Linguagem na Unicamp (cf. convênio e termo aditivo em anexo). Estamos agora buscando o financiamento no edital referência para implementação do projeto que na verdade já se iniciou apoiado pela instituição, mas que precisa de mais recursos sobretudo em equipamentos (basicamente computadores) e de bolsistas.

Projeto

Introdução e justificativa

Continuando o que está especificado, a nossa justificativa centra-se no desenvolvimento de novas metodologias para formação de grandes bancos de dados, segundo os mais recentes avanços do campo do processamento de linguagem natural, em tecnologias próprias da Inteligência Artificial com repercussões em diferentes áreas de conhecimento. Vamos implementar o que foi efetivamente feito como proposta do projeto piloto de pós-doutorado, a ser realizado entre 2009-2010, junto ao Instituto de Linguagem da Unicamp, sob a supervisão da Profa Charlotte C. Galves, com tecnologia desenvolvida no âmbito do Corpus Histórico do Português Tycho Brahe e do projeto interdisciplinar Padrões Rítmicos, Fixação de Parâmetros e Mudança Lingüística - Fase II (http://www.tycho.iel.unicamp.br/~tycho/). O principal objetivo deste estágio foi atingido, já que desenvolvemos a parceria, tanto científica quanto formal através de convênios institucionais para a consolidação, troca e aperfeiçoamento de novas tecnologias de processamento de texto para construção de um Corpus Eletrônico, o CE DOHS Corpus Eletrônico de Documentos Históricos do Sertão/BA, sediado na UEFS. Como dissemos, trata-se de dar um novo formato a um amplo conjunto de dados inéditos de toda grande área do semi-árido baiano, desenvolvido sob minha supervisão ao longo de mais de dez anos de pesquisa na UEFS,tratados sob rigor filológico, através de parceria com a UFBA/IL, no Projeto Para a História do Português (PROHPOR/ http:www.prohpor.ufba.br), mas com acesso ainda limitado a meios impressos. Essa parceria com a Unicamp visou a otimização para exploração computacional do material do projeto Vozes do Sertão em Dados (www.uefs.br/nelpP) , através de aquisição de novas tecnologias, com o uso da linguagem XML e da ferramenta integrada de anotação de corpus, E-Dictor (cf http://oncoto.dyndns.org:44880/projects/edictor), desenvolvido por Faria, Kepler e Paixão e Souza (2004-2009).

Objetivo geral

Este projeto de pesquisa tem como objetivo principal transformar um corpus digital em um Corpus Digital Eletrônico nos moldes do Corpus Histórico do Português Tycho Brahe (Unicamp) (http://www.tycho.iel.unicamp.br/~tycho), a partir de um corpus já constituído e já quase todo organizado no âmbito do projeto Vozes do sertão em dados: história, povos e formação do português brasileiro, subprojeto de um plano central de trabalho que vem sendo desenvolvido na UEFS e como parte do Programa para a história

do português (PROHPOR), sediado na UFBA, e de acordo com a agenda do Projeto Nacional Para a História do Português (PHPB), cabendo a equipe da UEFS, fazer o levantamento, a compilação, a edição de documentação e estudos lingüísticos da região semi-árida baiana. Os objetivos 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 lingüística. 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/corpus-xml.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, vamos fazer 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. 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 lingüístico, o semi-árido 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 lingüística. 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.
3. Oferecer catálogos eletrônicos de corpus serial (século XVII-XXI) de toda a região do semi-árido 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.

Objetivos específicos

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

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. Complementar o banco com documentação que está em fase de prospecção, através do resgate em acervos de Recife e Campinas;
4. Digitalizar os arquivos doados por diversos arquivos privados;
5. 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;
6. 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 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.
7. Testes da codificação XML;
8. Anotação morfológica através da equipe técnica da a ser constituída na UEFS com suporte do pessoal técnico do Corpus Histórico do Português Tycho Brahe;
9. Estabelecimento do formato da base de dados e do sítio-web do projeto piloto final;
10. 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.

Objeto do levantamento e registro de dados

O que pretendemos é oferecer um Corpus Eletrônico de documentos produzidos no sertão desde o século XVII até a atualidade nos formatos especificados ao longe deste projeto. O que pretendemos no âmbito desse projeto é disponibilizar materiais para análise e preservar a cultura sertaneja, resgatada através da língua que usam e usaram ao longo da sua história. A nossa motivação na separação para compor o banco seguirá as perspectivas que consideramos fundamentais para os estudos sócio-históricos do PB, quais sejam:

a) separar os materiais 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

lingüística irregular, no sentido defendido por Lucchesi e Baxter (2009);
b) separar materiais escritos por indivíduos sem contato prolongado com a escola (populares) ou com relativo contato (semi-cultos) daqueles com muito contato (cultos).

Desse modo, o nosso produto em versão computacional oferecerá um corpora sócio-históricos organizados e seriados, para usos diversos (edições de documentos inéditos e raros) e versão manipulável para uso lingüístico. Como dito tem como de partida o uso de uma base documental parcialmente inédita, que está em processo de digitalização no banco, intitulada Documentos históricos do Sertão/DOHS (cf. www.uefs.br/dohs). Um banco que resulta de uma ampla prospecção de fontes feita ao longo de mais de dez anos de pesquisa por uma equipe de pesquisadores da UEFS, filiada ao Projeto Para a História do Português/PROHPOR (http:www.prohpor.ufba.br), na UFBA, coordenado por Rosa Virgínia Mattos e Silva.
Uma primeira proposta de organização de um subconjunto de textos está em processo de finalização. Trata-se da Coleção Diacrônica Série Cartas Particulares, composta por seis volumes, a saber: Volume 1:cartas de brasileiros cultos, nascidos entre 1724 e 1880; Volume 2: cartas de populares da Bahia do século XIX; Volume 3: cartas de ilustres desconhecidos do século XIX ; Volume 4: cartas de brasileiros cultos do século XX; Volume 5: cartas de populares do século XX e Volume 6: cartas de mulheres (século XIX e século XX). A proposta de organização da coleção, nos termos colocados em a e b, foi pensada a partir de dois artigos de Rosa Virgínia Mattos e Silva, Idéias para a história do português brasileiro: fragmentos para uma composição posterior e De fontes sócio-históricas para a história social lingüística do Brasil: em busca de indícios apresentados, respectivamente, nos dois primeiros encontros do Projeto Para a História do Português Brasileiro (PHPB), em 1997 e 1998, publicados posteriormente nas Atas desse projeto.
Essa coleção já organizada e editada em formato digital será transformada em formato eletrônico. As cartas do produto final atingirá a meta de 2000 cartas.
Pensamos que é necessário, como sugeriu Mattos e Silva (1998:49) um corpus documental seriado que represente tanto as normas vernáculas como as normas cultas, comparar o português brasileiro com o português europeu. As nossas reflexões, mediante os resultados que os estudos em lingüística histórica vêm mostrando, sobretudo no âmbito do Projeto Nacional Para a História do Português Brasileiro/PHPB, é que somente conjuntos extensivos e seriados de documentos do PB podem oferecer um painel mais claro dos modelos de escrita e de que forma podem oferecer também uma amostra da língua do período, em um continuum: documentos que expressam mais claramente a fala, em que há praticamente uma transposição da fala para a escrita (os mais populares) até o extremo do continuum, em que um modelo de escrita bloqueia a língua falada (os documentos mais formais produzidos pelos cultos).
Enfim, este texto tem como propósito iniciar um plano de trabalho para o estudo do PB. Um passo importante e fundamental é tentar fazer um levantamento exaustivo das mudanças lingüísticas captadas nos textos produzidos no Brasil desde o século XVI até os nossos dias, como disse Mattos e Silva (1998) para que possa ter uma visão mais clara da competição de gramáticas no PB e dos fenômenos que ainda estão instáveis daqueles mais estáveis e buscar as razões de tal comportamento, se motivado por fatos lingüísticos ou se bloqueados pela opacidade dos textos diacrônicos.
E para finalizar e defender a importância dos textos populares, há que se registrar fenômenos lingüísticos somente encontrados na correspondência de populares.

Metodologia

Como se trata de um plano de conversão de um corpus já parcialmente construído faz-se necessário demonstrar a metodologia já aplicada no banco bruto e depois demonstrar o plano propriamente dito em um Corpus Eletrônico, pensado inicialmente a partir de um Piloto executado, como foi posposto no estágio de pós-doutoramento na UNICAMP. Dessa forma, parto inicialmente para a apresentação do estágio atual do banco de dados que coordeno na UEFS com o objetivo de mostrar em que medida é necessária a transformação para um Corpus Eletrônico e como isso será executado passo a passo. Nesse ponto, será necessário fazer um balanço breve do material das pesquisas feitas na UEFS em parceria com a UFBA, através do Projeto PROHPOR/Programa Para a História do Português e, também, com uso das metodologias desenvolvidas nos últimos dez anos do Projeto Nacional, integrado com várias universidades, o PHPB (Para a História do Português Brasileiro), a saber: UFPE, Bahia UFBa/UEFS, UFPb, UFAL, UFCE, UFMG, UFRJ (http://www.letras.ufrj.br/phpb-rj/,) UEL, UFSC, USP e UNICAMP.
1. DA PARTE DO CORPUS COMPILADO PELA UEFS (1998-2008) e de outras etapas previstas:

O semi-árido é uma região ainda pouco estudada do ponto de vista lingüístico e histórico. Neste item, apresentamos o resultado de dez anos de pesquisa, referente a estudos sócio-históricos e compilação de

documentação na região de forma a permitir estudos na área da lingüística histórica. É um banco que assume uma feição multidisciplinar, envolvendo pesquisas históricas e lingüísticas, além, é claro, da interface com a filologia, que fornece todo o aparato para o tratamento da documentação. Da parte histórica, é fundamental o que chamamos de controle de produção do material, ou seja, todas as classificações que, necessariamente, precisam ser feitas com relação ao tipo de material a ser analisado. De modo geral, para além das contribuições nas áreas já citadas, oferece uma contrapartida às regiões estudadas, por colaborar na preservação da cultura regional e oferecer subsídios para uma história lingüística local com implicações sócio-educacionais.

O objetivo é geral desse banco é compor uma amostragem de documentos históricos (século XVII-XXI), editados para fins lingüísticos de várias regiões do semi-árido baiano, a saber: Nordeste, Serra Geral, Chapada Diamantina e Paraguaçu, etc., para entender como se deu o avanço e a consolidação da língua portuguesa nessa região. São regiões que exemplificam o processo de ocupação européia falante do português na região. Através da extensão e consolidação de uma fase já concluída que trata de formação de banco de dados e análise lingüística de natureza sócio-histórica.

O segundo objetivo é usar o Corpus para estudar a história do português a partir de questões levantadas por projetos em Lingüística Histórica no que concerne a história do português europeu, como o PROHPOR. As questões para o português brasileiro estão sendo elaboradas com base no conhecimento acumulado, em dez anos de pesquisa, pelo PHPB, do qual fazemos parte. Essas questões são as seguintes: 1. quais são as características da gramática do português europeu no Brasil? 2. qual é a trajetória no tempo dessa gramática? 3. como se dá a emergência do português brasileiro? Essas questões já podem ser formuladas, uma vez que foi composto um vasto material de pesquisa inédito e fidedigno com análises já feitas na documentação do século XIX. Dessa forma, o problema da mudança nos textos escritos por brasileiros, em diferentes períodos do português, torna-se mais apreensível, uma vez que o projeto já detém um banco de dados sólido. O terceiro objetivo é propiciar um amplo conhecimento sócio-cultural e histórico de povos que viveram na região do semi-árido baiano, que embora seja conhecida por seu aspecto homogêneo climático, abriga uma rica diversidade geo-ambiental com impactos importantes no processo de ocupação e domínio de população de origem portuguesa e outras em detrimento de população autóctone durante todo o período colonial. Esse banco poderá fornecer meios para o resgate, não apenas documental, como o descrito, a seguir, mas também da diversidade lingüística da região. O quarto objetivo é o de resgatar documentação dispersa no Brasil e no exterior, gerada no semi-árido, uma vez que pouco ou quase nada se encontra em comunidades locais, fato que em muito contribuirá para a preservação da história regional, pouco documentada.

A metodologia seguiu as seguintes fases:

(i) seleção e identificação da autenticidade dos documentos; localização espacial e temporal cuja importância, do ponto de vista documental, é fundamental porque busca a preservação e disponibilidade de material para análise em diversas áreas de conhecimento, além da lingüística; (ii) identificação de dados relevantes sobre os escreventes e destinatários; (iii) contextualização da amostra com base na história externa do português brasileiro e (iv) edição em versão conservadora para preservar marcas lingüísticas relevantes para o estudo histórico do português. O produto final apresenta-se a seguir, separado por fases: concluída, em andamento e previstas, concluídos, em andamento e previstos (cf. www.uefs.br/dohs).

2 DA PARTE DO CORPUS ELETRÔNICO

A metodologia do Projeto de Corpus Eletrônico baseia-se fundamentalmente na metodologia do Corpus Histórico do Português Tycho Brahe, composto por um corpus eletrônico anotado de textos em português escritos por autores nascidos entre 1435 e 1845, desenvolvido desde 1998. 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 e que apresentarei de forma resumida. O resultado da aplicação da metodologia esulta em um produto atual de 52 textos (2.357.626 palavras), disponíveis para pesquisa livre, com um sistema de anotação lingüística em duas etapas: anotação morfológica (aplicada em 25 textos); e anotação sintática (aplicada em 5 textos). Atualmente, o projeto vem aplicando a sua metodologia em textos produzidos por brasileiros. A experiência pioneira realizada na minha tese de doutoramento com textos identificados por data de nascimento de autores serviu como um teste para alimentação desse banco com textos escritos por brasileiros, além de outros, como pode ser visualizada na página citada acima, item 48. A demonstração da aplicação sobre conversão de edição em Word em linguagem XML, que pode ser vista através da ferramenta E-Dictor (cf http://oncoto.dyndns.org:44880/projects/edictor), desenvolvido por Faria, Kepler e Paixão e Souza (2004-2009), demonstrada na I Oficina do Projeto na UFBA, em 2006, foi o que motivou este plano de projeto piloto na Unicamp a ser implantado posteriormente na UEFS. A demonstração inicial feita sob a coordenação de Paixão de Sousa pode ser visualizada no seguinte endereço (http://www.ime.usp.br/~tycho/participants/psousa), onde a autora apresenta um primeiro manual, usando exemplos de documentos editados em minha tese de doutoramento.

Então, em que consiste o processamento eletrônico de textos? Em seu projeto de pós-doutoramento na

Unicamp, Paixão de Souza apresenta um plano para elaboração de corpus em linguagem XML. Nesse projeto, que intitula de memórias de textos (disponível em http://www.ime.usp.br/~tycho/participants/psousa/memorias/critical_hyper/ece_Frameset.html), a autora mostra as diferenças nos tipos de edição, opondo o texto manuscrito versus texto editado manualmente em Word, como se configura o corpus editado no projeto apresentado acima no item (1) e do
texto digital/eletrônico versus texto produzido eletronicamente. A autora defende, apropriadamente, que isso pode ser fundamental no trabalho de natureza filológica - no qual o texto é o objeto central do estudo. Ultimamente, a transcrição e edição de textos antigos têm sido realizadas em meio eletrônico - ao menos, no plano do arquivamento da documentação. Neste trabalho de transcrição e edição, codificam-se as informações relevantes sobre os textos. A autora mostra como a linguagem XML apresenta todas as vantagens da linguagem HTML, fundamentalmente por inteiramente flexível, permitindo o controle e a geração de várias camadas de textos a ser produzida de acordo com a necessidade do pesquisador. Essa anotação na Linguagem de Marcação Extensível (extensible markup language, cf. W3C, 1997 permite gerar estruturas de dados na forma de árvore, cujos nós são os elementos XML nos quais se pode preencher valores nos campos adequados da interface de janelas da ferramenta. Para ver uma demonstração desse processo é possível acessar http://www.tycho.iel.unicamp.br/~tycho/corpus/manual/prep/index.html), devido a impossibilidade de inseri-lo aqui. As diretrizes na anotação dos documentos explicitadas por Namiuti e Santos (2009) são as seguintes: a catalogação dos textos; a transcrição dos textos; a codificação da interferência editorial sobre os textos e (iv) a apresentação dos textos. Por fim, para tornar mais claro esse processo, cito os argumentos de Galves (2004), no qual resume bem as vantagens desse sistema para um Corpus Digital Eletrônico, a saber: a) um melhor gerenciamento de arquivos do corpus, prevendo sua ampliação no futuro próximo; 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) a padronizacão do corpus de modo a poder inserí-lo em catálogos internacionais.
Nesse projeto, pretendemos usar recursos já desenvolvidos no Corpus Histórico do Português Tycho Brahe, ou seja, a Ferramenta Integrada de Anotação de Corpus denominada de E-Dictor, que foi pensada por Paixão de Souza e Fábio Kepler, em 2007. Em que consiste essa ferramenta? O E-Dictor destina-se a transcrição e a codificação de textos em formato XML para análises lingüísticas (morfológica, sintática, entre outras). Essa ferramenta, atualmente, está sendo implementada por Fábio Kepler (USP) e Pablo Faria (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 parte do layout do texto original (títulos, quebras de linha, página, etc.). Essa ferramenta permite a 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.
1. DA PREPARAÇÃO DOS TEXTOS Na preparação dos textos, serão usadas as seguintes etapas (Cf. Paixão de Souza, 2007), a saber: 1) PREPARAÇÃO DE EDIÇÔES Parte se um texto-fonte com grafia preservada (originais impressos e transcrições diplomáticas) para uma edição completa que implica modernização controlada do texto fonte para uma edição técnica para sanar dificuldades de processamento; 2) FORMAS DE SAÍDA DAS EDIÇÕES: transcrição do texto-fonte em edição fidedigna em relação ao texto tomado como fonte (grafia original e modernizada dos originais impressos); texto editado onde se mostra as interferências realizadas pela equipe do corpus (modernizações completas ou modificações técnicas, conforme o caso). Ambos em formato html para leitura e arquivos .txt, sem formatação, para uso das ferramentas automáticas e busca de dados e 3) GLOSSÁRIO - onde se mostra uma lista das intervenções realizadas pela equipe do corpus, seja no caso de edições completas ou técnicas. Nesse processo são utilizadas ferramentas disponíveis criadas na Universidade da Pensilvânia por Beth Randall, como a de correção dos textos. E para acesso dos usuários, a ferramenta de busca também autora denominadas de CorpusSearch. CorpusDraw e CorpusSearch, 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). Por fim, o leitor contará com o manual do Corpus Histórico do Português Tycho Brahe para uso do sistema de anotação morfologia e sintática que pode ser acessado em http://www.tycho.iel.unicamp.br/~tycho/corpus/manual/index.html.
2. DA ANOTAÇÃO SINTÁTICA E MORFOLÓGICA
Essa fase será feita em parceria com o Corpus Histórico do Português Tycho Brahe, por necessitar de constantes atualizações, como demonstra o histórico desse processo e que se encontra ainda em

desenvolvimento. Na página desse projeto é possível ver todo o histórico, desde a primeira fase em que Marcelo Finger (IME-USP) desenvolveu um etiquetador automático para o português baseado em parte no etiquetador de Eric Brill para o inglês (cf. Finger 1998, 2000) com precisão de 97% para os textos modernizados, como a elaboração de um analisador sintático que segue o sistema de anotação desenvolvido para o inglês no projeto de Anthony Kroch na Universidade da Pensilvânia (cf. http://www.ling.upenn.edu/mideng). Essa fase que pode ser vista no manual de anotação sintática disponível em http://www.ime.usp.br/~tycho/corpus/manual foi elaborada inicialmente por Helena Brito, responsável pela fase de treinamento do analisador sintático (parser) multilíngüe, desenvolvido por Dan Bikel na Universidade da Pensilvânia, usado pelo Corpus Histórico do Português Tycho Brahe, cedido por Anthony Kroch. Como descreve Galves em http://www.tycho.iel.unicamp.br/~tycho/prfpml/fase2/projeto_completo.html, a seqüência treinamento/correção é própria da metodologia dos analisadores probabilísticos, que segundo ainda a autora parece mais adequados que os analisadores baseados em regras (como o de Eckard Bick), por serem mais flexíveis e evolutivos.

Infra-estrutura disponível

DA UEFS

Atualmente, contamos com um núcleo chamado NELP- Núcleo de Estudos da Língua Portuguesa, no qual sou uma das coordenadoras (cf. www.uefs.nelp). Nesse núcleo, localizado no MT, módulo 2, contamos com sala, ar-condicionado, mesas, cadeiras, computadores de um projeto financiado pela Fapesb (A língua portuguesa no semi-árido, fase 3, no qual, sou integrante, e criadora, e do Vozes do Sertão em Dados (cf. www.uefs.br/nelp), onde tem links para esses projetos. Temos uma câmera semi-profissional para digitalização, uma impressora, um computador com uma tela de 22 para visualização dos documentos, um lap top e um HD externo de 1.000 bytes para armazenamento dos dados. Contamos com um armário grande onde guardamos toda a documentação, composto por vários partes por século (séc. XVII ao XXI), impressos e manuscritos.

DO TYCHO BRAHE QUE DARÁ O SUPORTE TECNOLÓGICO

O Projeto Corpus do Português Histórico Tycho Brahe têm como linhas centrais o seguintes pontos: análise e a modelamento dos sons das línguas naturais; a história da língua portuguesa, lingüística de corpus; a modelagem estocástica dos padrões rítmicos das línguas naturais e a sintaxe gerativa das línguas naturais. Define-se, conforme consta na plataforma Lattes, pela elaboração do corpus anotado do português histórico com mais de 2 milhões de palavras de autores portugueses e brasileiros nascidos entre os século XV-XIX, disponível em formato eletrônico no site do projeto (http://www.tycho.iel.unicamp.br/~tycho). Apresenta uma equipe formada por 14 pesquisadores, entre os quais me incluo, além de 24 estudantes de Doutorado, Mestrado e Iniciação Científica e do programador de computação e da técnica Cíntia Pano. Em 2003, por exemplo, através de um convênio CAPES/DAAD, com o Departamento de Lingüística da Universidade de Bielefeld, dois especialistas alemães visitaram o projeto Corpus do Português Histórico Tycho Brahe, Jan-Torsten Milde, e Thorsten Trippel. Na segunda visita, Maria Clara Paixão de Sousa trabalhou nas questões levantadas pela formatação do corpus nessa linguagem. Para a parte de suporte tecnológico computacional conta com a rede IME-USP. Fora do Brasil, filia-se ao pioneiro Penn-Helsinki Parsed Corpus of Middle English, coordenado por Anthony Kroch. Apresenta uma completa infra-estrutura com uma sala específica e com todo o mobiliário no IEL/UNICAMP. A estrutura computacional atual do projeto tem por finalidade atender a dois requisitos principais: suportar a hospedagem do sítio virtual e do corpus de textos, em si e hospedar aplicações que permitam aos bolsistas manterem o corpus, bem como manipulá-lo para consultas, etc. e apresenta a seguinte configuração:

- 1 servidor Dell, com capacidade para ficar 24hrs no ar, hospedando o sítio virtual do projeto e do corpus, os arquivos do corpus e as ferramentas ligadas a ele; Possui como sistema operacional o Ubuntu (Linux) e é onde os bolsistas do projeto são cadastrados. 2 estações (desktops) de trabalho extras, para bolsistas, com uma delas (a mais moderna) servindo também como espelho do servidor (uma salvaguarda para situações de falha no servidor). 4 laptops (1 permanente, na sala, e 3 para participantes do projeto); Tantos as estações, quanto os laptops, possuem sistema operacional Ubuntu, com todas as ferramentas necessárias instaladas e constantemente atualizadas. O servidor e as estações estão na rede do IEL. Para laptops, há um roteador na sala (ligado à rede do IEL) que fornece acesso Wi-Fi. O no-break específico para o servidor, evitando problemas graves quando há falta de energia.

Referências teóricas e metodológicas

Essa parte teórica será organizada em quatro planos para facilitar a apresentação da questão tratada, nos seguintes termos:

1. Da edição de documentos nos moldes filológicos tradicionais; 2. Das questões teóricas do conceito de mudança lingüística; 3. Da análise de fenômenos lingüísticos e 4. Da formação de bancos digitais eletrônicos.

1. Da edição de documentos nos moldes filológicos tradicionais

A questão mais complexa envolvida na construção de grandes corpora foi redimensionada em 1997 com a agenda do Projeto Para a História do Português Brasileiro/PHPB/BA/Feira de Santana, consolidada por ocasião do VII Seminário Nacional do PHPB (Londrina, maio de 2007), tendo sido publicados vários volumes sobre os trabalhos apresentados que culminarão em um conjunto de livros conclusivos sobre intitulados História do Português Brasileiro (planejado para 2011-2012), no qual escreveremos dois capítulos. O

planejamento e a redação desse livro refletem as agendas das equipes regionais, como por exemplo, a nossa, a do PHPB/BA. Os alvos são três: corpus, história social e análise de mudança lingüística. A formação do corpus pressupõe edição conservadora de documentos fidedignos, localização temporal, geográfica e o perfil biográfico dos autores dos documentos, se brasileiros, portugueses, data de nascimento, uma vez que se acredita que a mudança se dá no processo de aquisição. Para esse estudo faz-se fundamental a contribuição da micro-história, como a desenvolvida pelo historiador Erivaldo Fagundes Neves, pesquisador da UEFS e colaborador do projeto da Equipe UEFS.

2. Da formação e consolidação da mudança lingüística

A concepção da gramática gerativa, 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. Essa formulação teve grande aceitação e logo se tornou consensual. O problema passou a ser o de dimensionar a forma como a experiência lingüística atua nesse processo. Através do modelo de Princípios e Parâmetros (Chomsky, 1981), construído com base da comparação das línguas, é possível interpretar os fatos lingüísticos encontrados em textos diacrônicos e lhes atribuir uma gramática subjacente. A mudança sintática é interpretada como a mudança na fixação paramétrica.

Assim sendo, a relação entre mudança paramétrica e aquisição da linguagem motivou diversas tentativas de elaboração de uma teoria de mudança que viesse 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. Por outro lado, o fato de a lingüística histórica não ter acesso a dados introspectivos, mas apenas dados de língua-E foi alvo de discussões importantes na linha de interpretação da gramática gerativa. Para os pesquisadores dessa área, a mudança paramétrica é por definição abrupta, diferindo de outras interpretações tradicionais que defendem a sua natureza lenta e gradual, expressa através de processos de coexistência e concorrência das formas variantes pelo período que a antecede.

Para resolver esse impasse, estudiosos na área da gerativa defendem que a mudança é gradual nos textos porque a mudança paramétrica é o fim de um processo, como Roberts (1993) que reinterpreta as fases da lingüística histórica tradicional e diz que é possível ver isso através de alterações de freqüência, baixa na freqüência, reanálise e desaparecimento de um dado fenômeno. Ou seja, em seus termos, a mudança paramétrica é captada nessa última fase. Desse seu ponto de vista, assim como o de Lightfoot (1999), é, nesse momento, que a mudança se torna abrupta, daí ser vista como o fim de um processo. Por outro lado, Kroch (1994, 2001), ao associar a relação entre língua-I e alterações de freqüência, defende que o que se vê nos textos é a tensão entre a gramática nova e a gramática antiga, captada através de alterações na freqüência que atingem um conjunto de propriedades associadas a um determinado parâmetro, resultando no que denomina de Hipótese da Taxa Constante. Segundo essa assunção o que aparece nas amostras de língua-E é apenas o efeito da mudança.

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).

3. Da análise de fenômenos lingüísticos: estudos diacrônicos sobre o português brasileiro

Além dessas questões colocadas acima, há um aspecto importante que marca os estudos diacrônicos no Brasil no âmbito da gramática gerativa, o de interpretar dados quantificados em amostras de língua-E com base nas hipóteses sobre gramáticas abstratas apreensíveis nos textos. Os resultados desses estudos trouxeram um avanço teórico significativo para a compreensão do português brasileiro. Uma das mais importantes contribuições deve-se à hipótese sobre mudanças no sistema pronominal brasileiro. Essas mudanças estariam relacionadas à reestruturação do sistema de caso e ao surgimento de uma gramática distinta do português entre o século XVII e o século XIX.

Esses projetos lingüísticos pautam-se em importantes linhas de investigação da lingüística atual que buscam compreender, com dito acima, o que provoca a mudança lingüística, e como essa mudança se dá ao longo do tempo. Como contraparte social final, ajudar os educadores a entender a língua com a qual trabalham metalingüisticamente nas salas de aula, quando o resultado desses estudos chega aos livros didáticos através de projetos pedagógicos.

Os estudos em lingüística histórica, numa perspectiva gerativa, utilizam os dados reais de língua-E para extrair uma gramática abstrata, a língua-I. A metodologia impõe o como analisar as construções de língua-E. Por exemplo, se ao filólogo, no cumprimento de suas funções, cabe a preocupação com a autenticidade do texto, localização espacial e temporal, etc., ao lingüista cabe saber se se trata de documentos escritos por pessoas que têm essa língua como materna (L1) ou segunda língua (L2), uma vez que se defende que a gramática é construída durante a aquisição da linguagem. Por outro lado, o conhecimento prévio do que

determina uma mudança paramétrica é também crucial para que as construções que a representam, normalmente pouco representativas nos textos, não sejam ignoradas, principalmente em se tratando de amostras de língua escrita, por natureza conservadora e que podem estar refletindo a norma ou a tradição escrita de um dado período. A separação entre usos provocados por influências artificiais como, por exemplo, as influências de cunho estilístico ou discursivo, é importante para a identificação da mudança relevante em análises a partir de textos escritos, como é o caso desta pesquisa.
Atualmente no Brasil, há já muitos resultados com de estudos de diversas linhas teóricas e mostrando as propriedades específicas do português brasileiro, atestadas por diversos pesquisadores em resultados quantitativos e baseados em corpora diacrônicos da segunda metade do século XVII-XX (cf. Galves, 1987; Tarallo, 1989 e Roberts e Kato; 1993). Daí a vantagem do corpus digital eletrônico por permitir que se trabalhe com grande volume de dados para quantificação.

4. Da formação de bancos eletrônicos digitais

Para atingir uma análise qualitativa através da análise quantitativa é que tem se buscado aliar o recurso a grandes volumes de dados, através da contribuição de tecnologias computacionais, como a desenvolvida pelo Projeto Corpus Histórico do Português Tycho Brahe e do projeto interdisciplinar Rítmicos, Fixação de Parâmetros e Mudança Lingüística - Fase II, a partir de investigações teóricas solidamente fundadas, reunindo e fomentando a contribuição entre pesquisadores de formação variada - sintaticistas, fonólogos, estatísticos, probabilistas, teóricos da computação e lingüistas computacionais - contribuindo, por fim, na renovação das perspectivas para o campo, como especificado nessa fase do projeto no seguinte endereço http://www.tycho.iel.unicamp.br/~tycho/prfpml/fase2/projeto_completo.html). Esse tipo de banco de dados vem se mostrando uma tendência mundial, como podemos ver a partir do número de projetos elencados a seguir: 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. Diante desse quadro vê-se que os avanços tecnológicos são iminentes. De modo geral, tem se usado a linguagem XML pelas vantagens que apresenta (cf. apresentado metodologia)
Para finalizar, é importante salientar que esse tipo de edição eletrônica em linguagem XML e seus produtos abrem uma grande janela de oportunidades para edições filológicas com a aplicação eletrônica de textos, por possibilitar diferentes camadas de informação em um texto, ao invés do uso não manipulável da digitação em Word (Paixão de Sousa, 2006). O uso dessa tecnologia poderá dar novo impulso a Lingüística Histórica.

Essa parte teórica será organizada em quatro planos para facilitar a apresentação da questão tratada, nos seguintes termos:
1. Da edição de documentos nos moldes filológicos tradicionais; 2. Das questões teóricas do conceito de mudança lingüística; 3. Da análise de fenômenos lingüísticos e 4. Da formação de bancos digitais eletrônicos.
1. Da edição de documentos nos moldes filológicos tradicionais
A questão mais complexa envolvida na construção de grandes corpora foi redimensionada em 1997 com a agenda do Projeto Para a História do Português Brasileiro/PHPB/BA/Feira de Santana, consolidada por ocasião do VII Seminário Nacional do PHPB (Londrina, maio de 2007), tendo sido publicados vários volumes sobre os trabalhos apresentados que culminarão em um conjunto de livros conclusivos sobre intitulados História do Português Brasileiro (planejado para 2011-2012), no qual escreveremos dois capítulos. O planejamento e a redação desse livro refletem as agendas das equipes regionais, como por exemplo, a nossa, a do PHPB/BA. Os alvos são três: corpus, história social e análise de mudança lingüística. A formação do corpus pressupõe edição conservadora de documentos fidedignos, localização temporal, geográfica e o perfil biográfico dos autores dos documentos, se brasileiros, portugueses, data de nascimento, uma vez que se acredita que a mudança se dá no processo de aquisição. Para esse estudo faz-se fundamental a contribuição da micro-história, como a desenvolvida pelo historiador Erivaldo Fagundes Neves, pesquisador da UEFS e colaborador do projeto da Equipe UEFS.
2. Da formação e consolidação da mudança lingüística
A concepção da gramática gerativa, 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. Essa formulação teve grande aceitação e logo se tornou consensual. O problema passou a ser o de dimensionar a forma como a experiência lingüística atua nesse processo. Através do modelo de Princípios e Parâmetros (Chomsky, 1981), construído com base da comparação das línguas, é possível interpretar os fatos lingüísticos encontrados em textos diacrônicos e lhes atribuir uma gramática subjacente. A mudança sintática é interpretada como a mudança na fixação paramétrica.
Assim sendo, a relação entre mudança paramétrica e aquisição da linguagem motivou diversas tentativas de elaboração de uma teoria de mudança que viesse 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. Por outro lado, o fato de a lingüística histórica não ter acesso a dados introspectivos, mas apenas dados de língua-E foi alvo de discussões importantes na linha de interpretação da gramática gerativa. Para os pesquisadores dessa área, a mudança paramétrica é por definição abrupta, diferindo de outras interpretações tradicionais que defendem a sua natureza lenta e gradual, expressa através de processos de coexistência e concorrência das formas variantes pelo período que a antecede.
Para resolver esse impasse, estudiosos na área da gerativa defendem que a mudança é gradual nos textos porque a mudança paramétrica é o fim de um processo, como Roberts (1993) que reinterpreta as fases da lingüística histórica tradicional e diz que é possível ver isso através de alterações de freqüência, baixa na freqüência, reanálise e desaparecimento de um dado fenômeno. Ou seja, em seus termos, a mudança paramétrica é captada nessa última fase. Desse seu ponto de vista, assim como o de Lightfoot (1999), é, nesse momento, que a mudança se torna abrupta, daí ser vista como o fim de um processo. Por outro lado, Kroch (1994, 2001), ao associar a relação entre língua-I e alterações de freqüência, defende que o que se vê nos textos é a tensão entre a gramática nova e a gramática antiga, captada através de alterações na freqüência que atingem um conjunto de propriedades associadas a um determinado parâmetro, resultando no que denomina de Hipótese da Taxa Constante. Segundo essa assunção o que aparece nas amostras de língua-E é apenas o efeito da mudança.
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).
3. Da análise de fenômenos lingüísticos: estudos diacrônicos sobre o português brasileiro
Além dessas questões colocadas acima, há um aspecto importante que marca os estudos diacrônicos no Brasil no âmbito da gramática gerativa, o de interpretar dados quantificados em amostras de língua-E com base nas hipóteses sobre gramáticas abstratas apreensíveis nos textos. Os resultados desses estudos trouxeram um avanço teórico significativo para a compreensão do português brasileiro. Uma das mais importantes contribuições deve-se à hipótese sobre mudanças no sistema pronominal brasileiro. Essas mudanças estariam relacionadas à reestruturação do sistema de caso e ao surgimento de uma gramática distinta do português entre o século XVII e o século XIX.
Esses projetos lingüísticos pautam-se em importantes linhas de investigação da lingüística atual que buscam compreender, com dito acima, o que provoca a mudança lingüística, e como essa mudança se dá ao longo do tempo. Como contraparte social final, ajudar os educadores a entender a língua com a qual trabalham metalingüisticamente nas salas de aula, quando o resultado desses estudos chega aos livros didáticos através de projetos pedagógicos.

Os estudos em lingüística histórica, numa perspectiva gerativa, utilizam os dados reais de língua-E para extrair uma gramática abstrata, a língua-I. A metodologia impõe o como analisar as construções de língua-E. Por exemplo, se ao filólogo, no cumprimento de suas funções, cabe a preocupação com a autenticidade do texto, localização espacial e temporal, etc., ao lingüista cabe saber se se trata de documentos escritos por pessoas que têm essa língua como materna (L1) ou segunda língua (L2), uma vez que se defende que a gramática é construída durante a aquisição da linguagem. Por outro lado, o conhecimento prévio do que determina uma mudança paramétrica é também crucial para que as construções que a representam, normalmente pouco representativas nos textos, não sejam ignoradas, principalmente em se tratando de amostras de língua escrita, por natureza conservadora e que podem estar refletindo a norma ou a tradição escrita de um dado período. A separação entre usos provocados por influências artificiais como, por exemplo, as influências de cunho estilístico ou discursivo, é importante para a identificação da mudança relevante em análises a partir de textos escritos, como é o caso desta pesquisa.

Descrição do(s) produto(s) Final(is)

O Corpus Eletrônico tem como principal resultado a aplicação de tecnologias que possibilitem a convergência de edições tradicionais e limitadas a usos impressos ou mesmos digitais estáticos, em camadas de textos manipuláveis via uso da linguagem XML e da ferramenta integrada de anotação de corpus, E-Dictor (cf http://oncoto.dyndns.org:44880/projects/edictor), desenvolvido por Faria, Kepler e

Paixão e Souza (2004-2009) sobre o produto armazenado em diversos anos de pesquisa na UEFS, através da metodologia subsidiada, como enfatizado neste projeto, pela técnica desenvolvida pelo projeto Corpus Histórico do Português Tycho Brahe com o sistema de edições eletrônicas. De forma resumida, os resultados serão os seguintes:

1. Iniciar um plano de Corpus Digital Eletrônico permanente na UEFS, a partir da metodologia de edições eletrônicas em XML do Corpus Corpus Histórico do Português Tycho Brahe / CNPq - Edital Universal 2007 (Dez/2007 - Dez/2009).
2. Colaborar na metodologia de aplicação do E-Dictor (cf http://oncoto.dyndns.org:44880/projects/edictor) para troca mútua de experiências
3. Disponibilizar a transferência das metodologias para outros bancos regionais;
4. Subsidiar o Corpus Digital Eletrônico de forma que acompanhe os avanços científicos, através de avaliações com outros grupos;
5. Disponibilizar formas de acesso a produção automática propiciadas pela transformação XML, tais como léxico das edições, catálogos, etc.
6. Produzir materiais através do uso direto do Corpus Digital Eletrônico na rede mundial de computadores para levantamento extensivo de dados de forma confiável e rápida para uso acadêmico.
7. Elevação do número de artigos dos membros participantes;
8. Subsidiar o programa de pós-graduação stricto sensu em Lingüística em fase de implantação na UEFS.

Mecanismos de Disseminação e Transferência de Resultados

O acesso será gratuito através da rede em dois formatos.

1. A edição em XML no mesmo formato do Cards - (http://www.clul.ul.pt/sectores/filologia/projecto_cards.php) e no acesso a edições de todo o banco em linguagem compatível com todos os tipos de processamento (mecânico e humano), filologicamente preparadas para leitores com diferentes interesses e contextualizadas mediante recurso à documentação que as rodeia e anotadas por historiadores e linguistas.
2. Na parte de acesso a dados codificados será no molde do Corpus Histórico do Português Tycho Brahe, um parâmetro para o nosso modelo, isso é feito através de catálogos de listas ordenadas; acesso ao diretório de textos e consulta aos textos com CorpusSearch, uma ferramenta citada acima. Há, ainda, a possibilidade de se baixar o compactador em ZIP, disponibilizando-se os seguintes produtos: anotação sintática; anotação morfológica e corpus completo sem anotação. Essa edição em XML será a base para a edição manipulável a ser desenvolvida em parceria com a Unicamp. Trata-se da anotação morfológica e sintática feita no paser do projeto Tycho Brahe e disponibilizada para correção com estudantes da pós-graduação. Por isso, foi feito o convênio com a Unicamp. Essa é parte onde nos beneficiaremos com parceiras do Projeto Tycho Brahe. Justamente por ser uma das mais importantes para análise lingüística, uma vez que o seu produto é extraído de forma rápida em sistemas de buscas, permitindo que se trabalhe de forma segura com um grande volume de dados, o que dificilmente se conseguiria manualmente, diminuindo incomensuravelmente o tempo de análise nos moldes tradicionais. Contudo essa fase de anotação sintática é a parte mais complexa do nosso plano, é que necessário a parceria com a Unicamp, como dito.

Pensamos em fazer um protótipo do portal, mas no formulário não cabem diagramas. Mas o formato pode ser visto nos dois sites. Da edição no http://alfclul.clul.ul.pt/cards-fly/index.php?page=showLetter, onde o leitor poderá ver como fazer as buscas e da parte manipulário no site do Tycho Brahe (www.tycho.iel.unicamp.br) e também ver o modelo que desenvolvir para ilustração no pós-doc em www.uefs.br/dohs.

Impactos previstos

O estudo da mudança lingüística, com fortes repercussões na área da linguagem, constitui-se a gênese da própria lingüística como ciência, iniciada em fins do século XIX, e avança com suas metodologias de acordo com o desenvolvimento da própria ciência em geral, tanto do ponto de vista social, biológico cerebral quanto no plano tecnológico na área da inteligência artificial, beneficiados com a automação tecnológica. Como se trata de um campo que cresce com a troca de tecnologias e essas demandam grandes investimentos financeiros e acadêmicos, diversas universidades brasileiras vêm fazendo parcerias em todas as áreas de conhecimento e principalmente na construção de grandes bancos de dados compartilhados mundialmente na rede de computadores, como um recurso imprescindível para fazer lingüística na atualidade. No campo da Lingüística Histórica, no Brasil, um dos pioneiros nesse ramo é o projeto Corpus Histórico do Português Tycho Brahe No caso desse projeto, em particular, vem fazendo parcerias importantes com diversas instituições internacionais. A colaboração científica nessa parceria que nos propomos é importante porque esse tipo de banco de dados encontra problemas de adequação entre materiais editados na tradição filológica e sua conversão para linguagens computacionais, como a XML.

O avanço científico de modo geral depende crucialmente da manipulação segura e rápida de um grande volume de dados com entrecruzamento com redes espalhadas por todo mundo, como uma necessidade global. Diversos projetos na área de lingüística para formação de grandes bancos de dados vêm avançando rapidamente. São exemplos desse o próprio projeto Corpus Histórico do Português Tycho Brahe que é filiado ao Penn-Helsinki Parsed Corpus of Middle English, coordenado por Anthony Kroch, que vem se unindo em uma rede de colaboração mútua via conhecimento de experiências em outros grupos de pesquisa, no Brasil e no exterior, e o Projeto Lácio-Web do Núcleo Interdisciplinar de Lingüística Computacional (NILC), além de contar com o suporte computacional com a rede IME-USP. Sendo assim, o

projeto de Corpus Eletrônico, de tecnologia limpa, com tudo que isso implica, e como estamos propondo, tem como impacto a sua inserção em uma rede internacional de construção de grandes corpora históricos anotados. A parceria que proponho, através da troca de experiência com o maior centro desse tipo no Brasil, é uma necessidade para a Bahia, a fim de que possa contribuir, através da conversão de seus dados armazenados em projetos locais para linguagens mais adequadas. Os avanços se darão em várias frentes. No plano da pós-graduação (mestrado) na UEFS, em fase de implantação, que poderá já sair com grau de excelência na área de grandes bancos de dados, além da iniciação precoce tanto na própria universidade quanto no ensino médio.

Equipe Executora do Projeto - Com cadastro na Fapesb

Membro	Maior Titulação	Função no Projeto	Carga Horária	Instituição
Zenaide de Oliveira Novais Carneiro	34935614587 Pós-Doutorado/Unicamp Doutorado - Concluído em: 2005 Doutora/UFBA	Coordenador do Projeto Coordenadora	40 15	Departamento de Letras e Artes/UEFS Departamento de Letras e Artes/UEFS
Mariana Fagundes de Oliveira		Vice-Coordenadora	12	Departamento de Letras e Artes/UEFS
Patrício Nunes Barreiros	Doutor/UFBA	Pesquisador UEFS	8	Departamento de Letras e Artes/UEFS
Liliane L. S. Barreiros	Doutoranda-UFBA Doutora/UFBA	Pesquisador UEFS	8	Departamento de Letras e Artes/UEFS
Suani de Almeida Vasconcelos			8	Departamento de Letras e Artes/UEFS
Pedro Daniel dos Santos Souza	Doutorando-UFBA	Pesquisador UNEB	8	Departamento de Letras e Artes/UNEB
Huda da Silva Santiago	Doutoranda-UFBA	Pesquisador UEFS	8	Departamento de Letras e Artes/UEFS
Charlotte Marie Chambelland Galves	Pós-Doutorado - Concluído em: 1995	Colaborador Externo	5	Departamento de Lingüística - UNICAMP
Cristiane Namiuti Temponi	Pós-Doutorado - Concluído em: 2010	Colaborador Externo	5	Departamento de Estudos Linguisticos e Literarios - UESB
Jorge Viana Santos	Doutorado - Concluído em: 2008	Colaborador Externo	5	Departamento de Estudos Linguisticos e Literarios - UESB
Tânia Conceição Freire Lobo	Doutorado - Concluído em: 2001	Integrante	5	USP Departamento de Letras Vernaculas - UFBA
Maria Clara Paixão de Sousa	Pós-Doutorado - Concluído em: 2008	Colaborador Externo	5	Faculdade de Filosofia, Letras e Ciências Humanas -

Justificativa para a Formação da Equipe Executora

A equipe é formada por profissionais de produção de profundo conhecimento na área em que atuam. São profissionais com ampla experiência na formação de corpora como pode ser visto nos seus currículos disponibilizados na plataforma lattes. Sobre o consultor Pablo Faria e Maria Clara P. de Sousa, que são dois dos três criadores do Edictor, é importante contar com os mesmos para acompanhamento nas oficinas. O Colaborador Pablo Faria dará adicionalmente o suporte para a transferência de tecnologia e conhecimento na criação e manutenção de um corpus de texto eletrônico, em formato XML. Suas atividades incluem: Projeto e desenvolvimento de um site para acesso ao corpus eletrônico, que conterá todas as informações pertinentes sobre o corpus, o projeto e a equipe. Os arquivos do corpus, em si, serão hospedados no site do Corpus Tycho Brahe, simplificando o processo de manutenção, disponibilização e segurança dos arquivos digitais respectivos; Instalação da ferramenta de edição de textos, E-Dictor, e formação (através de workshop e/ou minicurso) de um primeiro grupo de editores, que posteriormente poderão treinar outros, conforme a necessidade; - Acompanhamento e consultoria, durante o período de edição e disponibilização online dos primeiros textos, até que a equipe local se mostre capaz de prosseguir de modo independente.
Quanto aos pesquisadores da UFBA, Tânia Lobo e Klebson Oliveira, os mesmos já são parceiros do projeto há mais de 10 anos, desde a sua fundação. A Profa. Charlotte Galves, criadora do Tcyho Brahe dará todo o suporte para transferência de tecnologia, responsável oficial na Unicamp pelo Convênio Guarda Chuva e Termo Aditivo.
Os professores Cristiani Namiutti terá participação na elaboração e execução do projeto, pois também tem doutorado na Área na Unicamp e Jorge Viana dará o suporte na organização dos bancos.
A presente proposta conta com uma equipe principal alocada na UEFS, constituída de dois pesquisadores estudantes de graduação. Além de uma equipe externa de pesquisadores colaboradores.

Os pesquisadores que integram a equipe interna somam experiências nas áreas necessárias para o desenvolvimento do projeto: contamos com experiência no campo de processamento computacional de textos e metodologias automáticas de buscas de dados da proponente, e com experiência em fotografia de documentos e paleografia do segundo coordenador.

O essencial será a contratação dos bolsistas que executarão sob orientação a transformação do banco em linguagem XML que será corrigido por toda a equipe em suas áreas de atuação.

Cronograma de Atividades

	2015											
1) Elaboração do plano central do Corpus Digital Eletrônico.	1	2	3	4	5	6	7	8	9	10	11	12
-Transcrição e edição em XML com a ferramenta E-dictor de textos históricos diversos e orais	X	X	X	X	X	X	X					
Codificação em XML dos textos da estrutura bruta em Word Piloto de Corpus Digital Eletrônico.	1	2	3	4	5	6	7	8	9	10	11	12
Anotação da estrutura bruta e parcial dos textos												X
	2016											

Revisão final da estrutura para teste de ~~diretórios do Piloto de Corpus Digital Eletrônico.~~	1	2	3	4	5	6	7	8	9	10	11	12
							X	X	X			
Treinar os pesquisadores no uso ~~EDICTOR.~~	1	2	3	4	5	6	7	8	9	10	11	12
Treinamento e aprofundamento da pesquisa de técnicas de anotação XML; estudo detalhado das diretrizes do Piloto de Corpus Digital Eletrônico; esboço das diretrizes de anotação.			X	X	X							
1) Elaboração do plano central do ~~Corpus Digital Eletrônico.~~	1	2	3	4	5	6	7	8	9	10	11	12
Apresentando o esboço para as diretrizes de preparação e a anotação XML dos primeiros 50 textos						X	X	X	X	X	X	X
Codificação em XML dos textos da ~~estrutura bruta em Word Piloto de Corpus Digital Eletrônico.~~	1	2	3	4	5	6	7	8	9	10	11	12
Anotação da estrutura bruta e parcial dos textos						X	X	X				
Revisão final da estrutura para teste de ~~diretórios do Piloto de Corpus Digital Eletrônico.~~	13	14	15	16	17	18	19	20	21	22	23	24
Preparação de uma primeira versão do Piloto do Corpus Digital Eletrônico e extensão da anotação para todos os textos	X	X	X	X								
Re-estruturação da arquitetura do ~~corpus eletrônico (a partir da página web) do Piloto do Corpus Digital Eletrônico~~	13	14	15	16	17	18	19	20	21	22	23	24
Inserção/Base Corpus Histórico do Português Tycho www.tycho.iel.unicamp.br					X	X						
1) Elaboração do plano central do ~~Corpus Digital Eletrônico.~~	13	14	15	16	17	18	19	20	21	22	23	24
Reelaboração do portal www.uefs.br/cedohs							X	X				

Resultado final do portal	13	14	15	16	17	18	19	20	21	22	23	24
Testa o portal: formas de acesso									X	X		
Redigir relatórios finais e apresentar os resultados ao público	13	14	15	16	17	18	19	20	21	22	23	24
Trabalho final de revisão, relatórios e produto final											X	X

Apoios Submetidos à FAPESB Pelo Solititante

Exercício: 2006 **Nº do pedido:** 663

Título do projeto:

DIGS9 THE9 DIACHRONIC GENERATIVE SINTAX CONFERENCE,HTTP://WWW.UNITS.IT/~DIGS9/

Modalidade: Participação em Evento Científico/Tecnológico Nacional

Situação do pedido: NEGADO

Exercício: 2006 **Nº do pedido:** 1640

Título do projeto:

CONTRIBUIÇÃO PARA A CONSTITUIÇÃO DE UM BANCO DE DADOS E DE UM BANCO DE TEXTOS PARA ESTUDOS LINGÜÍSTICOS HISTÓRICOS DO PORTUGUÊS BRASILEIRO, DO SÉCULO XVII AO SÉCULO XX: FASE

Modalidade: Projeto de Pesquisa Fluxo Contínuo

Situação do pedido: NEGADO

Exercício: 2008 **Nº do pedido:** 2444

Título do projeto:

XV CONGRESSO INTERNACIONAL DA ASSOC. DE LING. E FIL. DA AMÉRICA LATINA

Modalidade: Participação em Evento Científico/Tecnológico Internacional

Situação do pedido: NEGADO

Exercício: 2009 **Nº do pedido:** 1658

Título do projeto:

PROPOSTA DE UM SISTEMA DE EDIÇÕES ELETRÔNICAS E SEUS ASPECTOS TECNOLÓGICOS: A CONSTRUÇÃO DE UM PILOTO DE CORPUS ELETRÔNICO A SER IMPLANTADO NA UEFS PARA O ESTUDO DO PORTUGUÊS NO SEMI-ÁRIDO BAIANO

Modalidade: Pós Doutorado 2 - Fluxo Contínuo

Situação do pedido: APROVADO

Exercício: 2009 **Nº do pedido:** 5493

Título do projeto:

PARA A HISTÓRIA DO PORTUGUÊS BRASILEIRO: CARTAS DE BRASILEIROS CULTOS NASCIDOS ENTRE 1724-1880

Modalidade: Publicação Científica/Tecnológica - Edital

Situação do pedido: NEGADO

Local _____ Data _____/_____/ 20___

Zenaide de Oliveira Novais Carneiro Proponente do Projeto	José Carlos Barreto de Santana Respresentante legal da instituição