Projeto Banco Informatizado de Textos
do PROGRAMA PARA A HISTÓRIA DA LÍNGUA PORTUGUESA

Projeto BIT-PROHPOR

COORDENADOR RESPONSÁVEL: ROSA VIRGÍNIA MATTOS E SILVA

Supervisão do Português Brasileiro: Tânia Conceição Freire Lobo
Supervisão do Português Arcaico: Américo Venâncio Lopes Machado Filho

1 INTRODUÇÃO

O **Programa para a História da Língua Portuguesa – PROHPOR,** desde 1992 constituído como Grupo de Pesquisa com Auxílio Integrado junto ao CNPq (cf. Programa Geral – CNPq Proc. 500038/92-9), tem desenvolvido diversos trabalhos coletivos e individuais, voltados à constituição histórica da língua portuguesa, estabelecendo como arco temporal suas origens ao final do período arcaico, e a partir desse período, infletindo para a história do português brasileiro.

Os avanços tecnológicos empreendidos pelas ciências da computação têm, cada vez mais, disponibilizado instrumentos de levantamento e aferição para tratamento de dados empíricos, fazendo com que a demanda por *corpora* informatizados tenha, nos últimos anos, nomeadamente na Lingüística moderna, aumentado consideravelmente.

Diversos são os pesquisadores integrantes do Grupo PROHPOR, e não só, que já utilizam como suporte de análise, programas dessa natureza para quantificação e tratamento de dados, a exemplo do pacote de programas VARBRUL (Variable Rules Analyses).

Um dos grandes entraves, com que se deparam, contudo, esses pesquisadores, relaciona-se com a exigüidade de *corpora* representativos dos diversos momentos históricos da língua

portuguesa, devidamente digitalizados, para a consecução dos seus trabalhos de pesquisa.

Um projeto voltado à elaboração de um banco de textos informatizado, já havia sido proposta do Programa para a História da Língua Portuguesa no ano de 1992 de autoria de Dante Lucchesi, da Universidade Estadual de Feira de Santana e Rosa Virgínia Mattos e Silva, da Universidade Federal da Bahia, intitulado Banco de Dados para a História da Língua Portuguesa, que serve de base para a presente proposta.

As limitações técnicas advindas da precariedade dos equipamentos na época alocados ao projeto, assim como pelas inadequadas formas de armazenamento de dados pouco desenvolvidas naquela altura, que se restringiam a disquetes de 3,5 polegadas, como se sabe, de capacidade bastante reduzida e pouca durabilidade, fizeram com que o referido projeto não tivesse obtido os resultados inicialmente esperados.

Ademais, muitos dos colaboradores desse projeto por razões de diversa ordem, mas principalmente para cumprimento de cursos de Mestrado e Doutorado fora do âmbito da universidade, afastaram-se temporariamente das tarefas relacionadas ao banco de textos, o que contribuiu para seu postergamento.

Considerando que novos recursos passaram a ser oferecidos pela cibernética atual, nomeadamente no que concerne a digitalização e arquivamento de dados, o significativo aumento do quadro do grupo de membros do PROHPOR, inclusive de jovens voluntários, parece justificar-se a retomada do projeto original para a constituição de um banco de textos que venha a atender a real demanda de dados para trabalhos empíricos dessa natureza.

O Banco de Textos que se pretende, pois, implementar será composto por textos escritos nas diversas fases da história da Língua Portuguesa, que devem ser selecionados e armazenados com critérios e metodologia previamente definidos, e na medida do possível uniformes.

Esses critérios e metodologia deverão garantir: (i) a utilização dos seus materiais por meios informáticos; (ii) a representatividade desses materiais relativamente ao período histórico abrangido pelo PROHPOR; (iii) o rigor filológico na seleção e armazenamento dos materiais, conferindo a quem os utilize a segurança de que eles refletem os usos (o *modus scribendi*) empregados em sua feitura.

2 OBJETIVOS

Os objetivos do Banco de Textos são basicamente os seguintes: (i) constituir um *corpus* comum sobre o qual incidirá a análise dos estudos realizados no PROHPOR; e (ii) colocar à disposição desses estudos um instrumental de análise informatizado, constituído pelos programas desenvolvidos para a pesquisa lingüística.

A constituição de um *corpus* comum, organizado por critérios previamente definidos, e na medida do possível uniformes, proporcionaria um maior intercâmbio entre os estudos realizados sobre os diversos tópicos da língua. A base de observação comum facilitaria o cotejo dos resultados obtidos, possibilitando algumas generalizações como, por exemplo: (i) delimitação de fases da história da língua; (ii) asserções sobre tendências gerais em curso na língua; (iii) estabelecimento de uma cronologia para as mudanças observadas; e (iv) cadeias implicacionais entre as mudanças identificadas.

3 JUSTIFICATIVA

A tradição dos estudos de lingüística histórica é marcada pela natureza atomística das análises feitas nesse campo de estudo da língua.

Esse caráter atomístico das análises dos fatos lingüísticos, que inicialmente refletia concepções igualmente atomizadas (pré-saussurianas) do fenômeno lingüístico, manteve-se na Lingüística Histórica, mesmo quando essas concepções que o fundamentavam já estavam superadas, em boa medida devido à dificuldade de se proceder a uma observação sistemática, e na medida do possível exaustiva, dos materiais disponíveis.

Portanto, a proposta de constituição do Banco de Textos visa exatamente a romper com essa tendência ainda presente nos estudos de história da língua, possibilitando, com a facilidade de um amplo acesso aos materiais, a aplicação das novas teorias que propugnam uma apreensão globalizante do objeto através de sua estrutura interna (lingüística), e daquelas que, ainda mais globalizantes, propõem a apreensão dos fatos através da interação

do sistema de relações lingüísticas com as disposições e relações nas quais esse sistema se atualiza (i. é. as relações sociolingüísticas).

Esta iniciativa tem em mente, por outro lado, a dificuldade, já destacada por W. Labov (1972), em relação aos dados para o estudo da língua no *tempo real*. Um obstáculo irrefutável, diante do qual só resta à ciência buscar contorná-lo através da maximização dos recursos existentes, i. é. dos textos remanescentes escritos em fases pretéritas da língua.

Constituído inicialmente para servir ao PROHPOR, o Banco de Textos estará, entretanto, à disposição dos pesquisadores individuais e instituições que dele necessitarem para o desenvolvimento de suas pesquisas. Por outro lado, não se pensa obviamente que essa iniciativa irá sanar as dificuldades referidas acima, e sim que essa iniciativa é apenas uma contribuição dentre as que serão necessárias para que o problema seja satisfatoriamente solucionado. Em vista disso, os organizadores do Banco de Textos buscarão sempre o contato com promotores de iniciativas análogas, visando ao intercâmbio de informações, instrumental e materiais coletados; e intentando, nesse intercâmbio, a possibilidade de uniformização dos critérios e equipamentos utilizados na recolha e armazenamento dos materiais, para uma efetiva integração dos Bancos. Com isso, poder-se-ia proceder a uma distribuição complementar dos campos de recolha, com a conseqüente otimização de recursos.

4.0 DESENVOLVIMENTO

4.1 Os materiais

4.1.1 O objeto da observação

O Banco de Textos será constituído por textos, escritos em português, ao longo do período histórico da língua (i. é. o período do qual se conhecem registros escritos). Dentro desse período, que se estende do início do século XIII à atualidade, proceder-se-á a um novo recorte, na medida em que, nomeadamente a partir do século de XVII, a observação incidirá sobre os textos que documentem o processo de constituição do português do Brasil.

Portanto, o objeto de observação do PROHPOR, que define os limites do campo de coleta dos textos, consiste no Português Arcaico (séc. XIII a XVI) e no Português do Brasil (séc. XVII a XXI).

4.1.2 Os tipos de textos

Dentro dos limites definidos no item anterior, será feita uma seleção de textos cronologicamente seriados e de natureza e registros vários, buscando-se iluminar ao máximo a pluralidade e heterogeneidade que caracterizam o modo concreto de existir da língua.

Nesse sentido, serão selecionados textos de várias naturezas: textos notariais/foros; textos em prosa literária (traduções e textos escritos originalmente na língua vernácula); poesia; e a prosa epistolar.

Esta seleção de textos de natureza variada fundamenta-se, entre outras coisas, na idéia de que, através da diversidade de registros, se pode entrever a variação sócio-cultural da língua (cf. ROMAINE, 1982)

4.2 Metodologia

4.2.1 Critérios para a seleção dos textos

Na seleção dos textos, será observada, em primeiro lugar, a qualidade das edições disponíveis, considerando-se para isso o instrumental metodológico desenvolvido no âmbito da Crítica Textual. Tal critério visa a garantir que os materiais selecionados conservem o máximo de elementos presentes na feitura dos textos, os quais oferecem importantes indícios à análise lingüística.

A seleção dos textos obedecerá também a um seriamento cronológico, de forma a que todas as fases do período considerado sejam, na medida do possível, satisfatoriamente representadas.

Também, será dedicada uma atenção às informações extralingüísticas (datação, localização, informações sobre o escriba e sobre o *scriptorium*, etc.). Nesse aspecto, destacam-se os *foros* e *textos notariais*, que podem fornecer importantes indícios para a análise da variação diatópica da língua (cf., por

exemplo, MAIA, 1986).

4.2.2 Extensão dos textos

Deverão ser definidos também parâmetros de tratamento dos textos selecionados, segundo os critérios definidos acima, consoante a sua extensão. O problema se coloca particularmente para os textos de grande extensão. Nestes casos, dever-se-á considerar a proposta de estabelecimento de cotas, desenvolvida por ROMAINE 1982; buscando-se garantir, contudo, a representatividade dos materiais, tendo em vista os diversos tópicos da língua que serão analisados.

4.2.3 Armazenamento dos textos

4.2.3.1 Equipamento

Os textos deverão ser armazenados no disco rígido do aparelho do PROHPOR destinado exclusivamente a essa finalidade, com cópias produzidas em CD ROM, que possibilitem sua utilização por pesquisadores do Grupo ou outros de instituições nacionais ou estrangeiras que manifestem o interesse por esse material.

4.2.3.2 Critérios editoriais

Todos os textos selecionados deverão ser digitalizados a partir de equipamento *scanner*, com o suporte de programas de tratamento de texto, a exemplo do *OCR*, que se constitui, certamente, no processo mais adequado e seguro para migração de dados, evitando-se com isso erros de cópia inerentes à intervenção humana.

Cada texto deve ser armazenado em uma primeira versão, que deve ser uma cópia fidedigna da edição escolhida, acompanhada do seu *aparato crítico*. Deve-se considerar também a possibilidade de novas versões do textos, feitas através de *programas de reescrita* (dentro de princípios análogos aos enunciados em PARKINSON 1983), que visem à constituição de um *corpus* mais uniformizado e/ou de um *corpus* constituído de versões simplificadas, para as análises que possam prescindir de uma edição tão fidedigna.

4.2.3.3 Possibilidades de aproveitamento

Para o aproveitamento dos materiais por meios informáticos será utilizado o pacote de programas VARBRUL, de Sankoff & Cedergren, que permite a análise quantitativa dentro do modelo das regras variáveis (cf., por exemplo, SANKOFF, 1988).

Obviamente que os materiais do Banco de Textos, poderão ser usados tanto para análises que utilizem outros programas, quanto para análises que eventualmente não lancem mão de programas informáticos.

5.0 Recursos envolvidos no projeto

5.1 Humanos

A equipe responsável pela consecução do Projeto Banco Informatizado de Textos do PROHPOR (Projeto BIT-Prohpor) compõe-se dos seguintes membros:

Rosa Virgínia Mattos e Silva (Coord.)	Doutor c/ est. pós doc
Américo Venâncio Lopes Machado Filho	Doutor
Eliéte Oliveira Santos	Mestranda
Klebson Oliveira	Mestrando
Mariana Fagundes de Oliveira	Mestranda
Tânia Conceição Freire Lobo	Doutor
Sílvia Santos da Silva Gonçalves	Mestre

5.2 Materiais

Com computador multimídia e *scanner* já alocados exclusivamente ao Projeto, está prevista a aquisição de um gravador de CD, para o armazenamento dos dados e da alocação de uma impressora matricial, para racionalizar o grande volume de impressão de provas durante os processos de revisão e

posteriomente para impressão dos resultados das análises lingüísticas.

6 Orçamento

Produto	Característica	Cotação
01 Gravador de CD	CD-RW externo USB Philips 210109	R$1.056,00
01 Impressora matricial	Epson FX2180	R$1.090,00
20 *Compact discs*	CD formatado para gravação	R$100,00
Total do orçamento		**R$2.246,00**

6 Cronograma

O cronograma de execução do projeto do Banco de Textos compreende duas fases: a primeira destinada à seleção, digitalização e armazenamento dos textos relativos ao Português Arcaico (séc. XIII-XVI); e a segunda destinada aos textos do Português do Brasil (séc. XVII-XX). Propõe-se, assim, o seguinte o esquema de cronograma para execução do projeto Banco de Textos:

1ª FASE

1° semestre de 2002

- seleção dos textos do Português Arcaico
- definição dos "critérios editorias" para digitalização e armazenamento dos textos
- treinamento do pessoal de apoio (bolsistas).

2° semestre de 2002

- digitalização do textos do período arcaico da língua
- armazenamento dos textos selecionados
- início da seleção dos textos do português do Brasil

2ª FASE

1° semestre de 2003 a 1° semestre de 2005

- seleção dos textos do português do Brasil (conclusão)

- digitalização dos textos selecionados
- armazenamento dos textos selecionados

2° semestre de 2005

- disponibilização do material produzido
- elaboração do relatório da execução do projeto.

7 REFERÊNCIAS BIBLIOGRÁFICAS

LABOV, William. "On the use of the present to explain the past". In: Heilmann, L.(ed.). Bologna-Florence, Società Editrice il Mulino Bologna, 1972

MAIA, Clarinda. **História do Galego-Português**. Coimbra, Instituto Nacional de Investigação, 1986.

PARKINSON, Stephen. "Um arquivo computadorizado de textos medievais portugueses", **Boletim de Filologia**, XXVIII, 1-4, 241-252, 1983.

ROMAINE, Suzanne. **Sociohistorical Linguistics: its status and methodology**. Cambridge, Cambridge University Press, 1982.

SANKOFF, David. Variable Rules. In: Ammon, U., Dittmar, N. e Mattheier, J. K. **Sociolinguistics** - An internacional handbook of science of language and society. 1988.