I. Considerações iniciais

No cronograma revisado do projeto *Padrões rítmicos, fixação de parâmetros e mudança linguística, fase II*, encaminhado à Fapesp na ocasião do primeiro relatório, eram as seguintes as tarefas previstas para o segundo ano (2006):

- Extensão da codificação do *Corpus Tycho Brahe* em Linguagem XML;
- Inserção de textos brasileiros;
- Workshop nacional de divulgação e formação sobre as ferramentas de anotação do Corpus Tycho Brahe
- Re-treinamento do analisador sintático para o português;
- Marcação sintática de **400.000 palavras** (incluindo a correção da anotação morfológica);
- Análise do ritmo dos textos;

Algumas dessas tarefas foram realizadas plenamente. Trata-se em primeiro lugar da extensão do Corpus Tycho Brahe e da sua codificação em linguagem XML. Como apresentado no Relatório de Andamento do Módulo de Preparação de Textos (cf. Seção II), a adaptação à linguagem XML dos textos da primeira fase foi finalizada, foram acrescentados 8 novos textos num total de 501.442 palavras, e foi desenvolvida a anotação para edições de textos manuscritos. O Corpus contém atualmente 48 textos, num total de 2.279.455 palavras. Como previsto no projeto inicial, a extensão do Corpus se fez: no tempo, com a inclusão de textos dos séc. 14 e 15, no espaço, com a inclusão de textos brasileiros, e no tipo, com a inclusão de textos manuscritos.

Diretamente ligado à extensão do Corpus, o workshop sobre as ferramentas de anotação do Corpus aconteceu na Universidade Federal da Bahia, de 19 a 21 de abril, no âmbito deste projeto temático e do projeto CNPq/ Edital Universal 2004, CorPorA (Corpus do Português Anotado), coordenado por mim, numa promoção conjunta com o

grupo do Projeto para a História do Português (PROHPOR) coordenado pela Profa Tânia Lobo (cf. http://www.ime.usp.br/~tycho/corpora/).

A segunda parte do cronograma foi novamente prejudicada pela ausência de um pesquisador dedicado à tarefa de anotação sintática. Com efeito, a bolsa de pós-doutorado pleiteada para esse fim (cf. Considerações iniciais do Relatório do primeiro ano) não foi aprovada pela Fapesp. Não ficamos contudo de braços cruzados. A partir do texto já sintaticamente anotado e revisado em função das novas diretrizes de anotação do Corpus, o parser de Dan Bickel foi retreinado e aplicado a um texto do Séc. XIX, *As Memórias do Marquês de Alorna e Fronteira*, de 54 588 palavras, reetiquetado conforme as mesmas diretrizes, e com o novo etiquetador elaborado pelo aluno de Marcelo Finger, Fabio Kepler. Três pesquisadoras trabalharam durante o segundo semestre de 2006 na correção da anotação automática, que, como é de se esperar com uma ferramenta treinada com somente 50.000 palavras para uma tarefa tão complexa, foi bastante falha e portanto difícil e trabalhosa para corrigir. Esse trabalho está em fase de conclusão e prevê-se que a versão sintaticamente anotada do texto será incluida no Corpus no decorrer do mês de março. Os dois textos anotados servirão para um novo treinamento do Corpus, agora com mais de 100 000 palavras.

A lentidão do trabalho se deve a três fatores. O primeiro e o mais determinante é a falta de quem se dedique em tempo integral à tarefa, como acontece nos projetos similares – o *Penn-Helsinki Parsed Corpus of English* (Beatrice Santorini*)*, ou o *Canadian Parsed Corpus of Historical French* (Constanta Cornilescu*)*. As pesquisadoras que assumiram esta complexa e minuciosa tarefa este ano o fizeram em complemento a muitas outras. O segundo é a baixíssima eficiência, normal nesta fase inicial, do analisador automático. Cerca de 30% das sentenças são deixadas sem anotação, e as que são anotadas têm uma enorme quantidade de erros. O terceiro fator é dificuldade encontrada pelas anotadoras na instalação e na utilização, na fase inicial do trabalho, das ferramentas automáticas de correção.

Deve-se acrescentar a isso o fato de que, em razão da minha função de direção do IEL, não tive condições de me envolver pessoalmente nessa atividade. Deixei a direção do Instituto em 19 de janeiro passado e pretendo agora concentrar meus esforços no

encaminhamento dessa tarefa, que apresentei como prioritário na elaboração deste projeto. O balanço do primeiro biênio é longe de ser negativo, dada a expansão do Corpus e sua completa reestruturação num formato muito mais adequado, e dado o conjunto importante de trabalhos que foram produzidos no âmbito do projeto durante este período (cf. Seção II) . Porém não se cumpriu uma das metas mais importantes do projeto: privilegiar a anotação sintática como base para descrições mais seguras e análises mais empiricamente fundadas. A partir de meados deste ano, tenciono portanto participar eu mesma da realização desse trabalho, até que seja possível contar com um pesquisador dedicado. Passarei o mês de julho no Departamento de Linguística da Universidade da Pensilvânia, com a equipe de Anthony Kroch, em particular Beatrice Santorini, responsável pela anotação do *Penn-Helsinki Parsed Corpus of Early Modern English*.

No cronograma atualizado encaminhado em janeiro de 2006 na altura do primeiro relatório, reduzi a meta do número de palavras sintaticamente analisadas de 2 milhões para 1 milhão e meio, repartindo 1 milhão entre os anos 2006 e 2007, com a seguinte ponderação: "A meta de 1.000.000 de palavras anotadas sintaticamente no prazo de dois anos corresponde ao projeto de pós-doutorado encaminhado em setembro passado à Fapesp. Em função das dificuldades encontradas para a realização dessa tarefa no primeiro ano do projeto, não é mais possível manter, neste momento, a meta inicial de 2.000.000 de palavras". Uma vez que não foi possível contar com um pós-doutor para esse trabalho, é preciso revisar, de novo, as nossas metas. Uma opção para não haver uma redução drástica do que se pode razoavelmente prever para o fim do projeto será a sua extensão por mais um ano suplementar. No novo cronograma proposto no final deste relatório, levarei em consideração a possibilidade dessa extensão. A realização da nova meta ainda dependerá ainda da instalação de uma equipe mais permanente de anotadores.

O Corpus como mais que um repositório/ Sítio dinâmico, que roda.

O ano foi rico em divulgação do trabalho realizado no âmbito do projeto. Além do workshop já mencionado, devem enfatizar-se-se as duas participações em congressos internacionais de Linguística de Corpus de Maria Clara Paixão de Sousa, pós-doutora do projeto (O Seminário Internacional "Literaturas: del texto al hipertexto", na Universidade Complutense de Madrid, e a V Conferência internacional sobre "Language Resources and

Evaluation –LREC 2006 – em Genova). Ressalto também a minha participação no DIGS 9 em Trieste, com um trabalho conjunto com Zenaide Carneiro que terminou sua tese de doutorado, sob minha orientação, em dezembro de 2005. É importante mencionar que foi decidido no encontro de Trieste que o 11º DIGS aconteceria na Unicamp, em 2009, o que evidencia a visibilidade do trabalho que tem sido realizado em sintaxe histórica, no Brasil em geral e na Unicamp em particular.

Tanto os pesquisadores sêniores quanto os estudantes participaram em vários encontros nacionais, apresentando o resultado de suas pesquisas (cf. Seção II).

Queria enfim chamar a atenção para um evento que não está diretamente relacionado com o projeto mas que abordou temas importantes para a história do português, mais especificamente a constituição do português brasileiro, na sua relação com as línguas africanas. Trata-se do Colóquio "Caminhos da Língua portuguesa: África-Brasil", organizado por mim e realizado na Unicamp, de 6 a 9 de novembro de 2006, que contou com o auxílio da FAPESP, da CAPES e do FAEPEX-Unicamp. Uma das sessões temáticas foi consagrada ao trabalho conjunto com o grupo do Prohpor sobre um conjunto de documentos particularmente importantes para a história do PB, escritos na Bahia do séc. 19 por escravos brasileiros e africanos. Esses documentos estão sendo paralelamente incorporados ao Corpus Tycho Brahe.

Em 2006 e na primeira metade de 2007, vários trabalhos de dissertação e tese, desenvolvidos no âmbito do primeiro e do segundo projeto, estão sendo finalizados. Outros começaram, focalizando questões de sintaxe histórica do português. Na primeira categoria mencionarei o trabalho de Flaviane Fernandes, que se situa na interface da sintaxe e da fonologia, e já tem resultados publicados em revistas internacionais (cf. Seção II), bem como a tese de Cristiane Namiuti, que traz um novo olhar sobre o fenômeno da interpolação na história do português, e fortes evidências empíricas para uma nova periodização da língua. Ambas foram bolsistas da Fapesp desde a iniciação científica e participaram das atividades do projeto ao longo da sua formação. Ambas também se beneficiaram das relações privilegiadas que o projeto tem com pesquisadores do exterior, no caso, respectivamente, as pesquisadoras portuguesas Sonia Frota e Ana Maria Martins. Na categoria das novas teses ressaltarei também o projeto de doutorado

de André Antonelli, que diz respeito a uma das questões sintáticas levantadas no projeto temático: a questão da posição do verbo no português médio e no português europeu moderno.

Enfim, ressalta-se a publicação em 2006, no número 18 da revista *Probus,* do artigo "Secondary stress in two varieties of Portuguese and the Sotaq optimality based computer program", que traz uma importante contribuição à modelagem dos padrões prosódicos do português europeu e do português brasileiro, associada a uma implementação computacional.

A Seção IV apresenta os relatórios das duas viagens realizadas com os benefícios complementares do projeto, já comentadas acima.

A Seção V apresenta os gastos realizados com a Reserva Técnica do projeto durante os dois primeiros anos de vigência do projeto.

Termino este relatório com a elaboração de um novo cronograma (cf. Seção VI), levando em conta os progressos e atrasos dos primeiros anos. Note-se que se prevê a organização de um novo workshop nacional sobre as ferramentas de anotação em 2007, possivelmente na USP, nas instalações do *Núcleo de Modelagem Estatística e Complexidade* do qual o projeto faz parte.

Em anexo se encontra um resumo de atividades do ano da pós-doutora do projeto Maria Clara Paixão de Sousa, bem como um CD com a atual versão do Corpus.

Os trabalhos mencionados neste relatório são acessíveis no endereço, http://www.ime.usp.br/~tycho/papers