I. Considerações iniciais

No cronograma revisado do projeto *Padrões rítmicos, fixação de parâmetros e mudança linguística, fase II*, encaminhado à Fapesp na ocasião do terceiro relatório, são as seguintes as tarefas previstas para o quarto ano (2008) (que deixou de ser o último do projeto, uma vez que o ano suplementar solicitado na ocasião foi aprovado pela Assessoria):

- Anotação sintática de **7 textos**
- Inserção de novos textos (manuscritos editados)
- Análise sintática dos textos
- Modelagem da relação ritmo/sintaxe;
- Organização do workshop **Corpus Tycho Brahe, 10 anos** .

Em consonância com o cronograma, o ano de 2008 teve como uma das tarefas prioritárias a anotação sintática dos textos do Corpus Tycho Brahe. A previsão de 7 textos contava com a aprovação da bolsa de pós-doutorado solicitada para esse fim. Em primeira instância, o pedido foi rejeitado. Encaminhamos um pedido de reconsideração, que foi aprovado. Mas entretanto a candidata passou num processo seletivo na Universidade Estadual da Bahia, e, por razões pessoais, optou por renunciar à bolsa.
Formei então um pequeno grupo de doutorandos para trabalhar comigo na anotação. Foi assim possível anotar 3 novos textos:

- Barros, André (1675): Vida do apostólico padre Antonio Vieira, 52.055 palavras
- Sousa, Frei Luis de (1556): A vida de Frei Bertolameu dos Mártires, 53.986 palavras
- Cavaleiro de Oliveira (1702): Cartas, 51.234 palavras

Um quarto texto está sendo atualmente trabalhado e deve estar disponível no Corpus brevemente. Trata-se de Maria do Céu (Vida e Morte de Madre Helena da Cruz, 27.419 palavras).

Esses 4 textos somam em torno de 184.000 palavras, que, juntando aos dois textos corrigidos anteriormente, chegam a um total de em torno de 300.000 palavras sintaticamente anotadas.

Um ponto importante é que esse trabalho permitiu produzir uma primeira versão, disponibilizada on-line, do Manual de anotação (cf. .http://www.tycho.iel.unicamp.br/~tycho/corpus/manual/syn-frm.html). Um momento importante para a consolidação do Manual foi minha visita ao Departamento da Pensilvânia (cf. relatório na Seção IV) em agosto, para trabalhar com Beatrice Santorini, responsável pela anotação do *Penn-Helsinki Parsed Corpus of Early Modern English*.

Também aproveitei o convite da Profa Ana Maria Martins para uma banca de tese na Universidade de Lisboa, em março, para realizar uma sessão de trabalho sobre o sistema unificado de anotação do português com a equipe do projeto Cordial-Sin, que está anotando sintaticamente um conjunto de documentos de português europeu dialetal.

Alguns dos problemas que temos enfrentado continuam: baixa qualidade da anotação automática produzida pelo *Parser*, ausência de pessoas dedicadas à tarefa, dificuldades ainda encontradas pelo grupo de anotadores na aplicação do sistema de anotação, consequente trabalho de revisão muito demorado. Mas pouco a pouco, o Corpus sintaticamente anotado está se tornando realidade. Acreditando que podemos dobrar o número de textos no próximo ano, deve ser possível preparar mais oito textos até o final da vigência do projeto, perfazendo um total de 14 – em torno de 600.000 palavras. Ou seja, daqui a um ano, apesar das imensas dificuldades que esse projeto vem encontrando, teremos um conjunto de textos anotados que permitirão testar hipóteses sobre a história sintática do português europeu como nunca foi possível fazer. Minha intenção é continuar esse trabalho alem da vigência do projeto, uma vez que o mais difícil, que é lançar as bases, já está feito. Gerações de pesquisadores e estudantes (nos quais me incluo com meus orientandos) lucrarão com isso.

Em relação à inserção de novos textos, não foi possível avançar por duas razões. A primeira é a falta de implementação da bolsa TT3 solicitada na ocasião do último relatório, e aprovada pela Assessoria. Problemas técnicos na implementação do conjunto de bolsas TT solicitadas inviabilizaram essa parte do projeto até agora.

Se conseguirmos resolver esses problemas, em 2009, parte dos documentos das *Gazetas Manuscritas da Biblioteca Pública de Évora 1729-1734*, editados em livro por João Luis Lisboa, Tiago do Reis Miranda e Fernanda Olival (Edições Colibri, 2002 e 2005) poderão ser integrados ao Corpus, formatados em XML, em versão original e modernizada.

O que também inviabilizou a extensão do Corpus a novos textos - desta vez brasileiros - foi o fato de o projeto de pós-doutorado de Klebson Oliveira não ser aprovado, não por razões de mérito mas por ser o projeto julgado como devendo ser apoiado por organismos federais de fomento. É uma pena, porque era a oportunidade, como mencionei no relatório anterior, de juntar várias forças e competências num projeto só. Acredito que a historia do português brasileiro teria ganho com isso.

Apesar das dificuldades, uma série de melhorias foram realizadas no Corpus em vários níveis:

- Foi concluída a modernização dos textos *Monarquia Lusitana*, de Antonio Brandão (1584), e *Gazeta* de Manuel Galhegos (1597).
- Está sendo concluída a modernização do texto *Peregrinação* de Fernão Mendes Pinto (1510).
- Foi concluída a correção de etiquetas morfológicas do texto *Viagem à minha terra* de Almeida Garrett (1799).
- Está sendo corrigida a etiquetagem do *Teatro* do mesmo autor.

Enfim, está sendo integrado ao Corpus um conjunto de textos editados pela bolsista Ângela Kajita, no âmbito da sua dissertação de Mestrado, apoiada pela Fapesp (cf. Produção bibliográfica). Trata-se de documentos da Inquisição portuguesa do séc. 18, que vêm completar o Corpus das "Mãos inábeis", elaborado por Rita Marquilhas já integrado ao Corpus Tycho Brahe. Em 2009, prevê-se também completar o Corpus de *Cartas Brasileiras* editado por Zenaide Carneiro.

Além disso, avançamos muito nas potencialidades de trabalho, com a disponibilização da primeira versão da ferramenta E-dictor elaborada pelo bolsista Pablo Faria (cf. relatório em anexo) que permite formatar e modernizar os textos em XML de maneira muito mais rápida (nossa primeira avaliação é que o tempo necessário diminuiu de metade), e muito mais confiável, diminuindo fortemente os riscos de erros. O texto de Fernão Mendes Pinto, *Peregrinação*, mencionado acima está sendo usado para testar essa primeira versão.

O trabalho de análise sintática dos textos tem continuado por meio dos projetos dos doutorandos associados ao projeto (cf. Seção III). Duas teses de doutorado foram defendidas em fevereiro de 2008. Ambas trazem evidências empíricas de que a mudança do português clássico para o português europeu moderno se deu na passagem do século 17 para 18 (levando em conta a data de nascimento dos autores), indo ao encontro das conclusões de trabalhos anteriores desenvolvidos no âmbito do projeto. Nas teses em andamento, está colocada a questão da

natureza V2 da gramática clássica, e da interação da colocação de clíticos com a estrutura da oração e a natureza das categorias funcionais. Uma nova tese de doutorado se iniciou este ano, sobre o movimento do verbo em na história do espanhol. Apesar de fugir aparentemente do escopo do projeto temático, esse projeto nos propiciará elementos de comparação da história das duas línguas, muito próximas no período clássico, se afastando depois. A Seção III mostra que os alunos do projeto têm apresentado os resultados das suas pesquisas de maneira regular em eventos nacionais e internacionais. Pela primeira vez, o português clássico está sendo descrito de maneira detalhada, e sua gramática discutida com uma base empírica sólida. Esse estudo é relevante não só para a história do português europeu mas também para a do português brasileiro. Em dezembro foi defendida uma dissertação de Mestrado que discute a mudança do português em relação à propriedade do sujeito nulo, em textos de jornais mineiros, entre a primeira metade do séc. 19 e a primeira metade do séc. 20.

Em relação à modelagem da relação ritmo-sintaxe, foram produzidos três textos importantes e complementares durante o ano:

- O artigo "From Intonational Phrase to Syntactic Phase: the grammaticalization of enclisis in the history of Portuguese", em co-autoria com Filomena Sândalo, a ser submetido à revista Probus.
- O texto "Ler a fonologia: do português clássico ao português moderno", em co-autoria com Sonia Frota e Marina Vigário, no volume de Artigos selecionados do XXIII Encontro Nacional da Associação Portuguesa de Linguística.
- O artigo "Context-tree selection and linguistic rhythm retrieval from written texts" em co-autoria com Antonio Galves, Nancy Garcia e Florência Leonardi, submetido à revista JASA (Journal of The American Statistical Association)

O primeiro é uma reformulação e extensão de um trabalho publicado em 2004. Argumentamos que o domínio da restrição que força a ênclise é diferente no português clássico (PCl) e no português europeu moderno (PE). No primeiro, é a frase entoacional (IntP), e no segundo é o domínio sintático CP (considerado como 'fase' nos últimos desenvolvimentos do Programa Minimalista). Concomitantemente, a regra responsável pela afixação do clítico é diferente e se aplica em momentos diferentes da derivação: depois da linearização em ClP, antes da linearização no PE. Retomando um comentário de Anderson (2005) a nosso trabalho de 2004, sugerimos que se pode considerar essa mudança como um tipo de gramaticalização. Note-se que parte importante da argumentação empírica do trabalho diz respeito à posição na estrutura do sujeito pré-verbal no PE. Esse é um assunto longamente discutido em trabalhos anteriores do

projeto, uma vez que temos evidências de que a mudança afetando a colocação de clíticos afeta também a posição do sujeito.

Os dois outros artigos colocam a questão do ritmo da escrita. Baseado na ferramenta FreP, desenvolvida no Centro de Lingüística da Universidade de Lisboa, extrai-se de 8 textos do Corpus informações contidas na freqüência de determinados padrões segmentais, silábicos e acentuais. Emerge uma imagem da evolução da língua de tipo silábico para um tipo misto em que certas características das línguas acentuais se tornam mais proeminente. Do ponto de vista da datação, os resultados obtidos são compatíveis com a hipótese do projeto de que a mudança prosódica teria precedido a mudança sintática.

No terceiro trabalho, propõe-se uma modelagem probabilística do ritmo de textos escritos baseada na noção de árvores de contexto. Conseguimos discriminar o ritmo do português europeu e do português brasileiro. Essa metodologia está sendo atualmente aplicada a 15 textos do Corpus Tycho Brahe, de autores nascidos entre os séc. 15 e 19, à procura da datação da mudança, com base na hipótese de que a língua clássica era de tipo silábico, ou seja mais próxima do português brasileiro.

Enfim, não houve tempo hábil para a realização do workshop *Corpus Tycho Brahe, 10 anos*, previsto no cronograma de 2008. Foi contudo organizado um encontro intitulado: *Variação e Gramática: Diacronia e Aquisição*, no qual apresentei uma palestra intitulada: '10 anos do Corpus Tycho Brahe, balanço e perspectivas'. A programação do encontro está em anexo a esse relatório.

Queria acrescentar para concluir que o Projeto tem tido bastante visibilidade no Brasil e no Exterior. Ele certamente tem um papel no fato de que a Unicamp sediará em julho próximo a décima primeira edição do DIGS (Diachronic Generative Syntax Conference) que até agora estava restrito ao eixo Europa- América do Norte (cf. http://www.unicamp.br/~digs)

A Seção III traz a relação de todos os trabalhos publicados ou apresentados em congressos durante o ano 2008.

A Seção IV apresenta os relatórios das viagens realizadas com os benefícios complementares do projeto.

O relatório bem como os artigos e teses referidos nele são disponíveis na página do projeto: http://www.tycho.iel.unicamp.br/

Anexos ao relatório :

1. Relatório do bolsista Pablo Faria
2. Programa do workshop *Variação e Gramática: Diacronia e Aquisição*
3. Estatística de acesso ao Corpus