I. Considerações iniciais

No cronograma do projeto *Padrões rítmicos, fixação de parâmetros e mudança linguística, fase II*, encaminhado a Fapesp, são as seguintes as tarefas previstas para o primeiro ano:

- Treinamento do analisador sintático para o português;
- Marcação sintática (correção) de **350.000 palavras**;
- Extensão da anotação morfológica do *Corpus Tycho Brahe;*
- Codificação do *Corpus Tycho Brahe* em Linguagem XML;
- Inserção de novos textos;
- Análise do ritmo dos textos.

O projeto enfatizava a necessidade de priorizar a marcação sintática dos textos do Corpus Tycho Brahe como passo inicial para uma análise sintática extendida, uma vez que os resultados da primeira fase sobre a colocação de clíticos apontavam para a necessidade de aprofundar outros aspectos da mudança.

Infelizmente, não foi possível concretizar essa intenção durante este primeiro ano. A razão é que a pessoa prevista, e formada, para a marcação das 350000 palavras mencionada acima, aluna de doutorado do programa de linguística da Unicamp, teve que se afastar do projeto, e adiar sua defesa, por razões pessoais. Procurei muito ativamente quem a substituisse, mas não consegui achar quem tivesse a formação necessária, e se dispusesse a assumir a tarefa. No final do ano, a aluna teve condições de terminar sua tese, cuja defesa se realizou no dia 27 deste mês, e se ligar de novo ao projeto. Ela encaminhou um projeto de pós-doutorado centrado na anotação sintática do Corpus em

setembro de 2005 (cf. Anexo 19). Isso nos leva a propor um novo cronograma para a anotação do Corpus (cf. Seção VI)

Assim, as grande maioria das atividades do primeiro ano ligadas ao Corpus Tycho Brahe concerniram a reorganização do Corpus em linguagem XML, bem como a inserção de novos textos (cf. Seção II e CD em anexo) . Nesse campo, julgo que as metas foram satisfatoriamente cumpridas: a página do Corpus foi inteiramente reorganizada, 5 novos textos foram incluídos, 23 textos estão codificados em XML e os outros 22 estão em processo adiantado de codificação. Todos possuem cabeçalho de maneira a compor um catálogo que permite diferentes categorias de busca dos textos.

Por outro lado, foi um ano rico em publicações e trabalhos oriundos do projeto, bem como de divulgação da metodologia utilizada e das ferramentas produzidas (cf. Seção III). Com sua extensão para o período anterior (século 15), para textos não literários (Corpus das Mãos inábeis de Rita Marquilhas) e para textos brasileiros (Corpus de cartas de Zenaide Carneiro, em andamento), o Corpus Tycho Brahe está se afirmando cada vez mais como um instrumento essencial para a escrita da história da língua portuguesa em Portugal e no Brasil. Com seu novo formato, ele também atende à necessidades de outras categorias de pesquisadores.

Os trabalhos listados e resumidos na Seção III trazem resultados nas duas grandes áreas do projeto.

Por um lado, consolidam-se e divulgam-se, em publicações internacionais, os resultados, e consequências para a história do português, da evolução da colocação de clíticos baseada nos dados extraídos da primeira fase do Corpus. Esses resultados estão na base 1) de uma nova proposta de periodização do português, 2) de uma nova reflexão sobre a origem do português brasileiro, 3) da caracterização de três gramáticas distintas, cujas características básicas vão sendo empiricamente fundadas a cada novo trabalho.

Por outro lado, em colaboração com o Projeto temático "Comportamento estocástico, fenômenos críticos e identificação de padrões rítmicos nas línguas naturais" continua-se o trabalho de descrição e a modelagem dos padrões prosódicos das três vertentes do português consideradas na pesquisa, com o objetivo principal de elaborar metodologias

de detecção desses padrões nos textos escritos. O uso das cadeias de Markov de alcance variável (*Variable Length Markov Chain*, VLMC) tem se revelado muito interessante para a comparação de textos contemporâneos brasileiros e portugueses, e deverá ser no próximo período aplicado aos textos históricos. A outra vertente de trabalho consiste em entender a relação entre a prosódia e a sintaxe das línguas – no caso o português -, em particular no que diz respeito à coincidência de fronteiras prosódicas e sintáticas. Baseado em Galves e Sândalo (2004), o trabalho "Clitics in European Portuguese: X-bar and prosodic words phrasing" das mesmas autoras, formula um programa de pesquisa complementar da abordagem matemática baseada nas VLMC.

Foi também um ano de reorganização da infra-estrutura de equipamento, com a compra de novo servidor, e a reorganização da rede interna do projeto, baseada em Linux (cf. Seção IV).

Termino este relatório com a elaboração de um novo cronograma (cf. Seção VI), que leva em conta os progressos e atrasos do primeiro ano.

Em anexo se encontram cópias de todos os trabalhos mencionados na Seção III, bem como um CD do novo Corpus, com ponteiros para a nova página do projeto (http://www.ime.usp.br/~tycho). Juntam-se também o resumo de atividades relativo ao ano 2005 da pós-doutoranda Maria Clara Paixão de Sousa, bem como o projeto de pós-doutorado de Sílvia Regina de Oliveira Cavalcante, encaminhado à Fapesp em setembro de 2005.