I. Considerações iniciais

No cronograma revisado do projeto *Padrões rítmicos, fixação de parâmetros e mudança linguística, fase II*, encaminhado à Fapesp na ocasião do segundo relatório, são as seguintes as tarefas previstas para o terceiro ano (2007):

- Marcação sintática de **200.000 palavras** (incluindo a correção da anotação morfológica);
- Inserção de novos textos;
- Análise sintática dos textos;
- Modelagem da relação ritmo/sintaxe.
- Workshop nacional sobre as ferramentas de anotação.

As duas primeiras tarefas dizem respeito à extensão do Corpus Tycho Brahe. Quanto à inserção de novos textos, o trabalho deste ano foi no sentido de introduzir textos brasileiros, em consonância com os objetivos iniciais da Fase II do projeto. Dos 6 novos textos (cerca de 210 000 palavras), 5 são brasileiros, e o sexto é da autoria de Teresa Margarida Silva e Orta, nascida em São Paulo em 1711 e considerada por muitos como a primeira romancista brasileira. Esse trabalho foi realizado em parceria com dois outros projetos: *Caminhos do Romance*, coordenado por Márcia Abreu na Unicamp (cf. http://www.caminhosdoromance.iel.unicamp.br) , e *Variedades L2 e L1 do português escrito por africanos e afro-descendentes na Bahia do séc. XIX*: *estudos morfossintáticos*, integrante do Prohpor e coordenado por Tânia Lobo, na UFBA (http://www.prohpor.ufba.br/projetos.html). Com esse trabalho se inaugura uma nova fase para o Corpus, que passa a abrigar transcrições de textos impressos no Brasil no séc. 19 – preferencialmente primeiras edições – e transcrições de textos manuscritos brasileiros. Na seqüência da formatação em XML e inclusão no Corpus das Atas da Sociedade Protetora dos Desvalidos, o Projeto temático deve acolher o projeto de pós-doutorado de Klebson Oliveira, editor dessas atas, intitulado "Procuram-se fontes para a

história do nosso latim vulgar", em que ele pretende editar e incluir no Corpus textos representativos da escrita popular no Brasil do séc. 19 e primeira metade do séc. 20: tábuas votivas e cartas de cangaceiros.

Um outro projeto de integração ao Corpus de textos manuscritos recentemente editados diz respeito às *Gazetas manuscritas da biblioteca pública de Évora* (vol.1 1729-1731; vol.2 1732-1734). Como menciono no relatório de viagem a Évora (cf. Seção IV), estou em contato com um dos editores, João Luis Lisboa, para oficializar a inclusão dessa edição no Corpus. Trata-se de textos extremamente interessantes para a história do português europeu, uma vez que o séc. 18 é o século da mudança, e temos já no Corpus uma gazeta do séc. 17.

A tarefa de anotação sintática foi enfim implementada de modo produtivo durante o ano. Contamos com isso com uma ferramenta nova, *CorpusDraw*, criada na Universidade da Pensilvânia por Beth Randall, também autora da ferramenta de busca *CorpusSearch*. *CorpusDraw* e *CorpusSearch* são distribuídos no mesmo pacote no endereço corpussearch.sourceforge.net. *CorpusDraw* permite uma visualização gráfica da anotação que facilita enormemente o trabalho de correção da saída do analisador automático. Com essa ferramenta, foi corrigido integralmente o texto *Memórias*, do Marquês de Alorna e Fronteira (n. 1802), de 54.588 palavras. Também foi revisado, segundo as novas diretrizes de anotação, o texto *Reflexão sobre a vaidade dos homens*, de Matias Aires (n. 1705), de 56.479 palavras. Os dois textos anotados, totalizando 111.067 palavras já estão disponíveis no Corpus. Com base nesses dois textos, foi possível retreinar para o português o analisador automático (parser) de Dan Bickel. Um novo texto (*A vida do apostólico Padre Antonio Vieira*, de André de Barros, n. 1675), de 52.055 palavras, foi em seguida anotado automaticamente. A sua correção deve começar em breve.

Durante o período, paralelamente à correção do texto de Alorna, trabalhamos ativamente na revisão do manual de anotação elaborado por Helena Britto, pós-doutora da primeira fase do projeto (cf. http:// www.ime.usp.br/~tycho/corpus/manual/syntax.html, também acessível no CD anexado a este relatorio). Esse trabalho contou com a colaboração intensiva da Profa Cristina

Schmitt, da Universidade do Estado de Michigan, que esteve na Unicamp de 20 de maio a 30 de junho, com o apoio da Fapesp (processo 2007/01222-6). É preciso ressaltar também que o sistema de anotação está sendo desenvolvido conjuntamente com a equipe do projeto português Cordial-Sin (Corpus Dialetal para o estudo da Sintaxe) coordenado pela Profa Ana Maria Martins, do Centro de lingüística da Universidade de Lisboa, com financiamento da Fundação de Ciência e Tecnologia portuguesa (cf.http://www.clul.ul.pt/sectores/variacao/cordialsin/projecto_cordialsin.php).

Visitei o projeto da Profa Ana Maria Martins de 13 a 20 de maio deste ano, a convite dela. Foi a ocasião de revermos juntas todo o sistema de anotação. Dessas reuniões de trabalho também fez parte a pós-doutora do Cordial-Sin, Ernestina Carrilho, responsável pela anotação sintática do Corpus Dialetal. Na ocasião dessa visita, apresentei duas conferências: "Colocação de clíticos e posição do sujeito na história do português europeu", na Faculdade de Letras da Universidade de Lisboa, no dia 15 de maio, e "A competição de gramáticas na história da colocação de clíticos no português brasileiro", no Complexo inter-disciplinar da Universidade de Lisboa, no dia 17 de maio.

Também visitei o Departamento de Linguística da Universidade da Pensilvânia, de 1 a 18 de Agosto, para aperfeiçoar meu uso das ferramentas de anotação sintática e de busca e discutir detalhes da anotação para o português com a equipe de Anthony Kroch, em particular Beatrice Santorini, responsável pela anotação do *Penn-Helsinki Parsed Corpus of Early Modern English*. (cf. Seção IV)

No cronograma atualizado encaminhado em janeiro de 2007 na altura do segundo relatório, redefini a meta de anotação sintática para 200.000 palavras em 2007, 500.000 em 2008 e 500.000 em 2009, ano suplementar a ser solicitado. Só conseguimos um pouco mais de metade da meta de 2007, mas o trabalho está agora solidamente engatado. Parece-me mais realista fixar o objetivo agora em 1 milhão de palavras, ou 20 textos para o final de 2009. Se o desempenho do analisador automático e dos corretores aumentar significativamente, talvez consigamos mais. A correção do primeiro texto anotado automaticamente foi muito lenta, apesar da praticidade da nova ferramenta, por várias razões: pouquíssimo acerto do parser; frases imensas, que não tinham sido cortadas adequadamente (pela grande quantidade de coordenadas no texto de Alorna, que deveriam ter sido separadas previamente como orações independentes, e não foram – o

que fizemos agora para o terceiro texto); problemas com a aplicação da ferramenta no início, justamente por causa da separação das orações coordenadas – função prevista na ferramenta, mas que estava com um problema computacional que não tinha sido detectado até então; computadores inadequados porque lentos demais ou laptops com tela inadaptada para a ferramenta; dúvidas ainda sobre o sistema de anotação que estava sendo revisto; dificuldades dos anotadores nesse tipo de trabalho, raramente feito; lentidões devidas à estrutura da língua, por exemplo a grande quantidade de sujeitos nulos que são categorias demoradas de anotar. Em relação a esse último ponto, trabalhei muito na minha estadia na Pensilvânia com a funcionalidade da ferramenta *Corpus Search* para a revisão/anotação automática de certos aspectos recorrentes da língua. Escrevi um protocolo de revisão que anota automaticamente os sujeitos nulos, além de revisar as etiquetas das orações e de outras categorias. Esse tipo de procedimento agiliza bastante tanto a anotação (correção do parser) como a revisão geral (2ª correção).

O problema maior, que venho mencionando em todos os meus relatórios, é a dificuldade de achar as pessoas adequadas para fazer esse trabalho, e conseguir os recursos para remunerá-los. Trata-se de um trabalho minucioso que requer competência na análise gramatical e uma grande faculdade de atenção. Neste ano, contei com uma bolsa TT3, mas o aluno, apesar de ter mostrado entusiasmo no início, desistiu depois de 4 meses. Tratava-se de um excelente aluno, que me parecia ter todas as qualidades requeridas para a tarefa. Fiquei surpreendida com as dificuldades que ele teve ao fazer esse trabalho, e percebi que os alunos carecem de preparação para esse tipo de atividade. Encaminhei em novembro à Fundação um pedido de pós-doutorado que articula o trabalho de anotação com um projeto de pesquisa que precisa de textos anotados sintaticamente para ser bem sucedido (ver anexo V.). Espero que esse projeto seja aprovado. Ele prevê a correção de 5 textos em 2008 e 8 textos em 2009. Com os 2 textos já prontos, serão 15 textos anotados no final de 2009. Os outros 5 ficarão a cargo de um bolsista TT3 até o final da vigência estendida do projeto.

Queria ressaltar a importância crucial que teve para o desenvolvimento do Corpus, desde agosto deste ano, a participação do bolsista Pablo Faria, que conta com uma bolsa TT4A. Pelo relatório dele que segue em anexo, pode se verificar a extensão das suas atividades. Dentre dessas atividades, queria ressaltar uma particularmente importante

quanto ao uso do Corpus pela comunidade brasileira e internacional. O Pablo adaptou e melhorou sensivelmente uma ferramenta de busca on-line em textos etiquetados realizada para o Francês por Anthony Kroch, com base numa adaptação de *CorpusSearch* à busca em textos etiquetados morfologicamente (mas não sintaticamente anotados). Essa ferramenta, livremente acessível no Corpus, permite realizar buscar com base nas etiquetas e/ou nas palavras, manualmente usando a sintaxe de *CorpusSearch*, ou usando uma interface gráfica, nos 24 textos etiquetados corrigidos do Corpus, e mais 3 cuja etiquetagem automática não foi corrigida. Essa ferramenta representa um passo gigantesco para o uso do Corpus pela comunidade trabalhando sobre a história da língua portuguesa, e a compreensão por essa mesma comunidade do alcance de um Corpus desse tipo. Mas a importância da participação do Pablo, como se pode ver pelo extenso relatório em anexo, não se limitou a essa tarefa. Foi essencial para toda a dinâmica de construção e atualização do Corpus. Hoje em dia, contrariamente ao que aconteceu na primeira fase, não se trata mais de um simples repositório estático de arquivos. O uso do Corpus agora é dinâmico: depende de programas que geram as versões escolhidas pelo usuário, o catálogo também é gerado automaticamente e todas as atualizações e melhorias envolvem lidar com esses programas. A presença de um especialista da computação torna o trabalho dos outros participantes muito mais produtivo e eficiente, e permite a criação de um instrumento de trabalho cujas funcionalidades não cessam de aumentar. Acompanha portanto este relatório uma solicitação de recursos para a prorrogação da bolsa TT4A por mais 18 meses (cf. Seção VIII).

Não houve tempo hábil este ano para a realização do workshop nacional sobre o uso das ferramentas, mas estas foram apresentadas numa mesa redonda do Congresso Internacional da Abralin em Belo Horizonte em março de 2007, coordenada por Maria Clara Paixão de Sousa (cf. Seção III.3.), e organizei em agosto um workshop do projeto intitulado "O Corpus Tycho Brahe, construção e uso", com um dia consagrado às ferramentas (cf. Anexo 1 a esta Seção). Tive também a ocasião de divulgar os trabalhos baseados no Corpus Tycho Brahe no âmbito de um seminário de doutorado e de três conferências que dei em fevereiro no Departamento de Lingüística da Universidade de Paris VII a convite da Profa Carmen Sorin.

O workshop nacional sobre o Corpus ficará para 2008, ano em que o Corpus completará 10 anos.

O trabalho de análise sintática dos textos tem continuado por meio dos projetos dos doutorandos associados ao projeto (cf. Seção III). Em 2007 foram finalizadas duas teses, que serão defendidas respectivamente nos dias 25 e 26 de fevereiro. 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. O trabalho de Cristiane Namiuti sobre a história da interpolação também reforça a hipótese da existência de uma gramática média, entre a gramática arcaica e a gramática moderna, tal como propus no projeto de 2004. A Seção III mostra que os alunos do projeto têm apresentado os resultados das suas pesquisas de maneira regular em eventos nacionais e internacionais.

Quanto à modelagem da relação ritmo-sintaxe, foram feitos alguns avanços importantes. Na ocasião da defesa da tese de Flaviane Fernandes, em que participava a fonóloga portuguesa Sonia Frota, que foi a supervisora do estágio da Flaviane em Portugal, foi organizado o workshop *Domínios*, que reuniu os matemáticos e linguistas envolvidos na questão dessa modelagem (cf. Anexos 2 e 3 a esta Seção). Nessa ocasião, estabelecemos um protocolo de parceria com o projeto *Padrões de Frequência na Fonologia do Português - Investigação e Aplicações*, do Laboratório de Fonética da Universidade de Lisboa (cf. http://www.fl.ul.pt/LaboratorioFonetica/projectos.htm) . O texto dessa parceria se encontra no Anexo 4 desta Seção. Os primeiros frutos dessa parceria se encontram no texto apresentado no Encontro da *Associação Portuguesa de Lingüística* em outubro de 2007 em Évora, sob o titulo, Ler a fonologia: do português clássico ao português moderno , cuja versão definitiva já foi submetida para publicação no volume de artigos selecionados. Encontra-se também em fase adiantada de redação o artigo From Prosodic Inversion to Local-dislocation : the grammaticalization of enclisis in the history of European Portuguese , em co-autoria com Filomena Sândalo, do qual uma versão preliminar tinha sido apresentada no workshop.

A Seção III traz a relação e os resumos de todos os trabalhos publicados ou apresentados em congressos durante o ano 2007.

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

A Seção V apresenta os gastos realizados com a Reserva Técnica do projeto durante o ano.

Termino este relatório com a elaboração de um novo cronograma (cf. Seção VI), que já leva em conta a extensão da vigência do projeto por mais um ano, solicitada na Seção VII. Na Seção VIII, solicito mais bolsas TT, em consonância com as novas diretivas da Fundação.

Em anexo ao relatório se encontram os relatórios da pesquisadora principal Filomena Sândalo, da pós-doutora Maria Clara Paixão de Sousa (até setembro 2007), dos bolsistas TT Marco Antonio Pereira Domingues (agosto a dezembro de 2007) e Pablo Picasso de Faria (agosto de 2007 a janeiro de 2008), bem como os projetos de pós-doutorado de Cristiane Namiuti e Klebson Oliveira.

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

Anexos a esta Seção:

1. Programa do workshop *O Corpus Tycho Brahe, construção e uso*
2. Programa do workshop *Domínios*
3. Cartaz do workshop *Domínios*
4. Parceria com o projeto *Padrões de Frequência na Fonologia do Português*