Balanço e perspectivas

Este relatório é ao mesmo tempo o do último ano do projeto e o do projeto inteiro, pelo menos a sua segunda fase, objeto do atual auxílio. Procurarei aqui contemplar simultaneamente esses dois aspectos.

O projeto submetido à FAPESP em 2004 apresentava assim os seus objetivos:

O primeiro objetivo do presente projeto é a extensão e consolidação do Corpus Tycho Brahe (http://www.ime.usp.br/~tycho/corpus), elaborado na fase anterior, tanto no que diz respeito aos textos que o compõem quanto à sua anotação morfológica e sintática.

O segundo objetivo é usar o Corpus para estudar a história do português a partir das seguintes questões:

- *Quais são as características da gramática intermediária entre a do português arcaico e do português europeu moderno?*
- *Qual é a trajetória no tempo dessa gramática?*
- *Como se dá a emergência do português europeu moderno?*

De uma maneira mais geral, este projeto se insere em duas grandes problemáticas da teoria da linguagem:

- *O que provoca a mudança lingüística?*
- *Como essa mudança se dá ao longo do tempo?*

Enfim, o projeto enfrenta a questão metodológica da detecção da mudança nos textos escritos. Para isso pretende articular análise qualitativa – no âmbito da teoria da gramática gerativa - e análise quantitativa, lançando mão da estatística descritiva e da modelagem estocástica.

No cronograma revisado do projeto *Padrões rítmicos, fixação de parâmetros e mudança linguística, fase II*, encaminhado à Fapesp na ocasião do quarto relatório, são as seguintes as tarefas previstas para o quinto, e último, ano (2009).

- Anotação sintática de **8 textos**

- Formatação e modernização dos Textos da *Gazeta Manuscrita de Évora* (a depender de bolsa TT3)
- Estudos sintáticos baseados nos textos do Corpus
- Consolidação da modelagem da relação ritmo/sintaxe
- Síntese final.

Seguirei aqui o roteiro acima, procurando trazer resposta, ao mesmo tempo, às grandes questões colocadas no projeto inicial.

1. A anotação sintática

A anotação sintática dos textos é a etapa final da construção do Corpus Tycho Brahe, primeiro objetivo deste temático bem como daquele que o precedeu. Nesta fase, foi uma prioridade absoluta, e na ocasião deste relatório, o Corpus já tem meio milhão de palavras anotadas sintaticamente, em 11 textos de autores nascidos entre 1510 e 1836.[1]

Este ano, ainda não foi possível alcançar a meta inicial de textos anotados (8), mas nos aproximamos bastante dela uma vez que foram 6 os textos produzidos ao longo deste ano:

Maria do Céu (1658)	Vida e Morte de Madre Helena da Cruz	27.419
Marquesa de Alorna (1750)	Cartas	49.900
Padre A. Vieira (1608)	Sermões	53.855
Fernão Mendes Pinto (1510)	Perigrinação	52.555
Manuel de Galhegos (1598)	Gazeta	28.926
Ramalho Ortigão (1836)	Cartas a Emília	32.441
TOTAL		245.096

Deve-se notar que conseguimos produzir este ano perto de metade do Corpus anotado atualmente disponível, que totaliza neste momento 513.038 palavras e onze

[1] No CD anexo, que contem os arquivos do Corpus disponíveis na data de entrega do relatório, as versões sintaticamente anotadas de Sousa, Vieira e Marquesa de Alorna estão incompletas, por faltar a revisão final de 30% aproximadamente de cada um desses textos. As versões completas serão disponibilizadas no decorrer do mês de abril.

textos.[2] Isso mostra uma aceleração do ritmo de trabalho, devida à consolidação do Analisador automático (cf. Seção II), e à experiência adquirida pelo grupo de anotadores.

O Corpus Tycho Brahe, que alcança agora sua 3ª fase, é uma realização extremamente importante para os estudos históricos do português, e constitui o único Corpus eletrônico histórico anotado dessa língua. Para as outras línguas, só se encontram Corpora históricos anotados do inglês – O Penn-Helsinki Parsed Corpus of Middle English, que foi nosso modelo, acrescido agora do Penn Parsed Corpus of Modern British English (cf. http://www.ling.upenn.edu/hist-corpora para os outros Corpora do Inglês desenvolvidos na Universidade de York em colaboração com a Universidade da Pensilvânia), e o Corpus do francês construído no âmbito do projeto canadense "Les voies du français", nos mesmos moldes (http://www.voies.uottawa.ca/projet_fr.html). Temos hoje à nossa disposição meio milhão de palavras cobrindo a história do português num período que vai do início do séc. 16 a meados do séc. 19 (data de nascimento dos autores), que nos permite obter informações morfossintáticas de todo o tipo instantaneamente, graças à ferramenta de busca Corpus Search. Como será relatado mais abaixo, estamos agora em condições de consolidar análises a respeito do rumo das mudanças acontecidas no período, que tinham sido percebidas com um conjunto de dados mais reduzido e mais limitado contextualmente, e eram portanto sujeitas a revisões. Alem disso, o Corpus está disponível na rede mundial de computadores e está portanto à disposição de todos os estudiosos do português. O anexo X mostra que ele é frequentemente visitado. Temos recebido este ano 140 solicitações de acesso aos textos (cf. Anexo X).

Como será enfatizado na Seção Perspectivas, o fato de o Projeto chegar ao fim não significa que o Corpus para por aqui. Com todas as dificuldades que têm sido relatadas ao longo dos anos, desencadeou-se uma dinâmica que vai no sentido de um crescimento e uma diversificação permanentes. Este ano, já se começou a anotar morfologicamente e

[2] A lista completa dos textos com anotação sintática –por ordem cronológica- é atualmente a seguinte:

Fernão Mendes Pinto (1510)*	Perigrinação	52.555
Luis de Sousa (1556)	A vida de Frei Bertolameu dos Mártires	53.986
Manuel de Galhegos (1598)*	Gazeta	28.926
Padre A. Vieira (1608)*	Sermões	53.855
Maria do Céu (1658)*	Vida e Morte de Madre Helena da Cruz	27.419
André de Barros (1675)	Vida do apostólico padre Antonio Vieira	52.055
Cavaleiro de Oliveira (1702)	Cartas	51.234
Matias Aires (1705)	Reflexões sobre a Vaidade dos Homens	56.479
Marquesa de Alorna (1750)*	Cartas	49.900
Marquês d'Alorna (1802)	Memórias	54.588
Ramalho Ortigão (1836)*	Cartas a Emília	32.441

sintaticamente textos brasileiros oriundos de projetos associados ao Projeto para a História do Português Brasileiro (PHPB). Trata-se das *Cartas Brasileiras*, editadas por Zenaide Carneiro na sua tese de doutorado defendida na Unicamp em 2005, já disponíveis no Corpus na versão etiquetada, atualmente em fase de correção da anotação sintática, e das *Atas dos Brasileiros*, editadas por Klebson Oliveira na sua tese de doutorado defendida na UFBA em 2006, em fase de correção da anotação morfológica, e com anotação sintática prevista em seguida. Nos dois casos, os textos anotados serão usados para auxiliar as pesquisas sobre esses documentos de grande relevância para a história do português brasileiro, que devem ser divulgadas em coletâneas de artigos a serem publicadas no próximo ano.

Desde 2009, graças ao apoio técnico do bolsista Pablo Faria, o analisador sintático (parser) é rodado na sala do projeto, no IEL-Unicamp. Isso agiliza muito o trabalho, uma vez que não dependemos mais da disponibilidade dos colegas da Pensilvânia. O parser tem sido retreinado a medida que os textos anotados são corrigidos, e já melhorou bastante. No anexo X, pode-se observar umas saídas, captadas pela interface Corpus Draw que usamos para a correção. Como mencionado acima, a melhoria do desempenho da ferramenta automática bem como a experiência adquirida pelos corretores, nos permitem agora trabalhar cada vez mais rapidamente. Essa consolidação da prática de trabalho bem como das ferramentas utilizadas está na origem de uma possibilidade crescente de transferência dessa tecnologia a outros grupos de pesquisa, que cada vez se mostram mais interessados. Retomo esse assunto na Seção Perspectiva.

Quero aqui registrar o nome dos doutorandos que estão participando do projeto de anotação sintática. A colaboração deles tem sido decisiva para chegar no ponto em que estamos hoje, ou seja com o Corpus anotado que se tornou realidade: Aline Gravina (bolsista CAPES), André Antonelli (bolsista FAPESP), Aroldo Andrade (bolsista FAPESP), Carlos Felipe da Conceição Pinto (bolsista FAPESP), Lilian Teixeira (bolsista FAPESP). Tem participado também ativamente do empreendimento a minha ex-orientanda Cristiane Namiuti, hoje professora na Universidade do Sul da Bahia, em Vitória da Conquista.

Deve-se ressaltar também a contribuição extremamente frutuosa de Cristina Schmitt, da Michigan State University, que participou decisivamente da revisão do sistema de anotação em 2007 e contribuiu ativamente à nova anotação dos textos *Reflexão sobre a vaidade dos homens* e *Memórias* do Marques de Alorna. Enfim, esse trabalho todo não existiria sem o modelo do Corpus do Inglês Médio da Universidade da Pensilvânia (*Penn-Helsinki Parsed Corpus of Middle English*) e do apoio e da orientação constantes do Prof. Anthony Kroch e da sua colaboradora Beatrice Santorini, que, durante a vigência deste

projeto me receberam 2 vezes em Philadelphia para sessões intensivas de trabalho (cf. relatórios anteriores).

2. A inserção de novos textos no Corpus

Fazia parte das tarefas deste ano a formatação e modernização das *Gazetas Manuscritas da Biblioteca Pública de Évora 1729-1734*. Isso foi integralmente cumprido pela bolsista Cynthia Yano (cf. relatório em anexo). Infelizmente, não podemos ainda inserir o texto no Corpus uma vez que não tivemos mais resposta para nossa proposta de protocolo de acordo, reiterada várias vezes, por parte da equipe portuguesa que publicou a edição da Gazeta em livro. Mas o arquivo do texto em XML consta do CD em anexo (http://www.tycho.iel.unicamp.br/~tycho/corpus/texts/xml/va_005)

.

Vale aqui mencionar as inserções novas que foram feitas no Corpus ao longo deste Período. No projeto da Fase II, previa-se o seguinte:

Além de reforçar com textos novos o período que os trabalhos anteriores apontam como aquele em que se deu a mudança para o português europeu moderno (entre 1650 e 1725), pretende-se estender o Corpus nas três direções apresentadas a seguir.

1.2.1 Recuando no tempo

Na fase I, o texto mais antigo disponível no Corpus era de um autor nascido nos últimos anos do século 15. Neste projeto, estenderemos o período contemplado para os séculos 15 e 14. A justificativa para isso é detalhada na seção 2.1.

Uma primeira lista de autores, sujeita a modificações, é a seguinte:

- ***Fernão Lopes**, 1380-1460 Crônica de D. João I (pesquisa preliminar: edição diplomática por Anselmo Braamcamp Freire, Lisboa, 1945. (A 1ª edição é de Lisboa, 1644.)*
- ***Dom Duarte**, 1391-1438 Leal Conselheiro, Livro da ensinança de cavalgar bem toda sela (edição a ser escolhida)*
- ***Rui de Pina**, 1440-1522 Crónica de D. Dinis (pesquisa preliminar: edição da Liv. Da Civilização, Porto, 1945)*
- ***Gil Vicente** (c. 1465-?), Obras completas (pesquisa preliminar: edição da Liv. Sá da Costa, 1951)*

- ***Bernardim Ribeiro*** *(1482-1552), Menina e Moça (pesquisa preliminar: edição quinhentista impressa por André de Burgos, Évora, 1557-58).*
- ***Pero de Magalhães de Gândavo*** *(?-1576) História da Provínvia Santa Cruz, a que vulgarmente chamamos Brasil (pesquisa preliminar: 1ª edição Lisboa, Antonio Gonçalves, 1576; edição da Academia Real das Ciências, 1858, disponível na Biblioteca Central, Unicamp).*

1.2.2 Textos não literários

1.2.3 Textos "brasileiros"

No final da Fase I, o Corpus continha 1.851.619 palavras (cf. relatório final disponível em http://www.tycho.iel.unicamp.br/~tycho/prfpml/fase2/98_04.html). São agora 2.398.715. Esse aumento de aproximadamente 450.000 palavras seguiu o roteiro proposto acima, apesar de algumas reorientações.

Dos autores mais antigos, foram incluídos Fernão Lopes, Rui de Pina e Pero de Magalhães Gândavo, além do autor quinhentista Duarte Galvão (1435-1517). A maior ausência me parece ser a de Gil Vicente, em função da pouca presença no Corpus de textos teatrais, e da imensa riqueza linguística desse autor. Num próximo projeto, o teatro será certamente um gênero a ser contemplado prioritariamente, e Gil Vicente um autor incontornável.

Foram também acrescentados dois autores portugueses nascidos no período identificado como o da mudança:

- Manuel dos Santos, n.1673, *História Sebástica*, 87.991 palavras.
- Tereza Margarida Silva e Orta, n. 1711, *Aventuras de Diófanes*, 68.985 palavras

Quanto aos *textos não literários*, ficaram essencialmente por conta dos textos brasileiros, grande inovação deste período da construção do Corpus. Graças à colaboração com grupos de pesquisa envolvidos na edição de textos produzidos no Brasil nos séc. 17 a 20, foi possível integrar dois conjuntos de textos representativos da complexa escrita brasileira novecentista:

- As *Atas da Sociedade Protetora dos Desvalidos*, escritos por ex-escravos nascidos na África (*Atas dos Africanos*) na primeira metade do séc. 19 ou no Brasil (*Atas dos Brasileiros*) na segunda metade do mesmo século - totalizando

em torno de 60.000 palavras – editados por Klebson Oliveira na ocasião de sua tese de doutorado na UFBA.

- As *Cartas Brasileiras* (escritas ao longo do Séc 19, editadas por Zenaide Carneiro como parte da sua tese de doutorado na Unicamp.

Como mencionado acima, esses dois conjuntos estão atualmente em fase de anotação morfossintática.

Para concluir esta seção, é importante enfatizar que a Gazeta de Évora foi formatada em XML usando a ferramenta E-DICTOR, criada no âmbito do projeto por Maria Clara Paixão de Sousa, pós-doutoranda do projeto de 2004 a 2006, em colaboração com Fábio Kepler, doutorando em computação na USP sob a orientação do Prof. Marcelo Finger, e implementada pelo bolsista Pablo Faria (cf. Relatório em anexo). A ferramenta permite fazer com eficiência, rapidez e segurança todas as operações necessárias para a padronização e modernização dos textos, bem como para a geração de catálogos, geração de léxicos, e correção das etiquetas morfológicas. Vários outros textos foram revisados este ano usando essa ferramenta, que torna a edição eletrônica muito mais fácil de realizar. Ela é certamente um dos grandes produtos deste projeto.

3. Estudos sintáticos baseados nos textos do Corpus

A seção II mostra que a produção decorrente do projeto continua densa. Podemos ver nela duas grandes linhas, uma consolidada e outra emergente:

3.1 Os estudos da gramática do português clássico

Os trabalhos de Aroldo Andrade (*A subida de clíticos em português: Um estudo sobre a variedade europeia dos séculos XVI a XX* – tese de doutorado com defesa marcada para o dia 28 de abril de 2010) e Ana Luiza Lopes (*A ênclise em orações dependentes na história do Português Europeu (Séc. 16 a 19)* – dissertação de mestrado defendida dia 26 de fevereiro de 2010) fecharam o ciclo, muito denso, dos estudos da posição e colocação de clíticos na história do português europeu a partir do séc. 16. O primeiro estuda o fenômeno da subida de clíticos, que envolve a complexa questão sintática da "restruturação" de orações. A segunda considera o fenômeno muito marginal de um ponto de vista quantitativo, porem bastante relevante para a discussão da estrutura da oração, e da compreensão do que está em jogo na alternância ênclise/próclise, da ênclise

em orações subordinadas. Ambos trabalhos reforçam a conclusão de que o chamado português clássico correspondem a uma gramática (que intitulei de "média" no texto inicial deste projeto, e de "hispânica" na mesa redonda do Congresso em homenagem à Rosa Virgínia – ROSAE) de tipo V2, cuja posição pré-verbal é ocupada por uma sintagma tópico ou foco, que não é forçosamente o sujeito da oração. Particularmente no trabalho do Aroldo, também se verifica a hipótese apresentada em Galves, Brito e Paixão de Sousa (2005) de que é com as gerações nascidas na primeira metade do séc. 18 que emerge uma outra gramática, na qual há uma posição pré-verbal reservada para o sujeito.

A dissertação de Juliana Trannin *A sintaxe do infinitivo com verbos causativos na diacronia do Português Europeu*, é complementar da tese sobre a subida de clíticos porque ela estuda orações que são contextos quase categóricos de subida de clíticos, uma vez que formam com a oração superior uma "união de orações". O enfoque da Juliana não são os clíticos – que aparecem somente como diagnósticos das estruturas subjacentes - mas a natureza dessas estruturas (denominadas na literatura de "faire-infinitif", "Marcação excepcional de caso", e "infinitivo flexionado") cuja distribuição varia ao longo do tempo.

A tese de Alba Gibrail, *As estruturas de topicalização do Português Médio ao Português Europeu Moderno, um estudo diacrônico e teórico***,** cuja defesa está também prevista para este ano, estuda uma questão central para a história do português, a saber, as diversas formas de topicalização e sua distribuição ao longo do tempo. De novo, a natureza V2 do português clássico está no centro desse trabalho, bem como a perda dessa propriedade no português europeu moderno, associada a novas formas de topicalização.

A tese de André Antonelli *Sintaxe de Posição do Verbo e Mudança Gramatical na História do Português Europeu* olha para o português clássico e a mudança para o português europeu moderno do ponto de vista da posição do verbo. Isso passa pela análise da chamada periferia esquerda da sentença, e o estudo das orações interrogativas, alem das orações afirmativas. Essa tese fecha o ciclo, uma vez que ela deve produzir uma representação formal da gramática do português clássico, ao determinar de que tipo de língua V2 se trata, ou seja, quais são as categorias envolvidas na posição do verbo, dos elementos que o precedem, e do sujeito quando está em posição pós-verbal. Por isso, é uma tese que levanta também de maneira crucial a questão da ordem, em particular da ordem Verbo – Sujeito, retomando questões que tinham sido já levantadas na tese de Maria Clara Paixão de Sousa, em 2004.

Esses projetos de tese correspondem aos temas propostos no Projeto inicial para o estudo da gramática do português clássico e a mudança para o português europeu moderno (cf. http://www.tycho.iel.unicamp.br/~tycho/prfpml/fase2/projeto_completo.html#ahistoriadoPE) : a sintaxe de topicalização, a posição do sujeito, a posição do verbo. O outro assunto mencionado no projeto era a questão do uso do determinante em sintagmas possessivos. Esse foi o tema da tese de Simone Floripi intitulada *Estudo da variação do determinante em sintagmas nominais possessivos do Português Médio ao Português Europeu Moderno*, defendida em 2008.

Como pode ser verificado na Seção III, os trabalhos de tese desenvolvidos no âmbito do projeto e com base no Corpus Tycho Brahe, têm sido divulgados em congressos locais (Seminário de Tese em andamento – SETA), nacionais/internacionais no Brasil (Abralin, ANPOLL, Rosae, ICHS) e internacionais no exterior (Incontro di Gramática Generativa, Sienna; TUCSPS, Philadelphia; Linguistic Evidence, Tubingen; SIMELP, Évora). Estão sendo também publicados em artigos de revistas locais e nacionais com boa circulação.

Note-se também a divulgação em artigos e comunicações em congressos nacionais e internacionais das pesquisadoras que, tendo já defendido seu doutorado, continuam a participar do projeto nas suas instituições, e a criar novos núcleos de pesquisa mantendo ligação, tanto de um ponto de vista metodológico quanto no que diz respeito às grandes questões de pesquisa, com o projeto Tycho Brahe. É o caso de Cristiane Namiuti, atualmente docente na Universidade Estadual da Bahia, em Vitória da Conquista, que divulga ativamente sua pesquisa de doutorado defendido em 2008, em torno dos temas da negação e da interpolação na história do português. Pode-se citar também a mesa-redonda organizada por Maria Clara Paixão de Sousa no ROSAE, intitulada “A aliança entre abordagens quantitativas e perspectivas formais: repercussões sobre o estudo das mudanças gramaticais do português”.

Uma questão essencial do projeto tem sido a da periodização. No que diz respeito à mudança do português clássico para o português europeu moderno, como já mencionado, os trabalhos desenvolvidos graças ao Corpus Tycho Brahe permitiram sustentar a hipótese de que aconteceu no início do século 18, e não um século antes como proposto por Ana Maria Martins na sua tese de doutorado. Mas discutimos bastante a questão da emergência da gramática chamada por mim de *média* no Projeto. O texto que retoma sistematicamente essa questão é Galves, Namiuti e Paixão de Sousa (2006), que argumenta que a gramática media, subjacente ao português clássico, emerge nos autores nascidos na segunda metade do séc. 14, quando tradicionalmente se considera

que se entra numa segunda fase do português antigo, chamada de português médio por Lindley Cintra e Ivo Castro. Argumentamos que esse Período não deve ser considerado como o fim do português antigo, mas antes como o início da fase gramatical seguinte, que durará até o início do séc. 18. Essa visão da história do português tem conseqüências para a compreensão do português brasileiro, que discuti no meu artigo de 2007 “Periodização do português europeu e origem do português brasileiro”, às quais voltarei mais abaixo. Ela constitui também uma linha de pensamento para a análise da história do português europeu, que foi retomada em vários dos trabalhos citados aqui, em particular na comunicação que apresentei no ROSAE intitulada “Periodização e competição de gramáticas, o caso do português médio”. Essa visão tem conseqüências para futuras linhas de pesquisa, porque ela postula no português quatrocentista e quinhentista a emergência da gramática que vai plenamente se expressar nos textos dos autores nascidos na segunda metade do séc. 16 e no séc. 17. Muito trabalho empírico ainda deve ser feito sobre isso, em particular nos mais antigos dos autores concernidos, como Fernão Lopes (já no Corpus Tycho Brahe), Dom Duarte e outros.

3.2 Os estudos da história do português brasileiro de um ponto de vista comparativo

Os avanços no conhecimento da língua escrita entre o séc. 16 e 19 nos permitiram inaugurar uma nova linha de pesquisa na lingüística histórica do português: a história comparada do português europeu e brasileiro. Essa comparação, alem de ser interessante em si, é essencial para a compreensão da complexa variação que aparece nos textos escritos no Brasil no século 19. Vários trabalhos seguem essa linha dentro do projeto, que também está associada à inclusão de textos brasileiros no Corpus Tycho Brahe mencionada acima (cf. Galves e Lobo (2009), Galves (2009), e os trabalhos de Galves e Carneiro. Essa nova direção de pesquisa já abriga um projeto de doutorado: o de Aline Gravina que propõe um estudo diacrônico comparado da inversão do sujeito em português europeu e brasileiro.

4. Consolidação da modelagem ritmo/sintaxe

Uma característica saliente deste projeto é propor um articulação entre a fonologia e a sintaxe na mudança, partindo da hipótese, formulada na primeira fase, de que a

mudança sintática observável no séc. 18 teria sido provocada por uma mudança prosódica, que lhe seria portanto anterior. Essa hipótese está na origem de uma pesquisa muito frutuosa no interior do projeto, e de uma colaboração com matemáticos e estatísticos que tem produzido resultados inovadores, tanto do lado lingüístico quanto do lado da modelagem matemática.

No interior da própria teoria lingüística, essa linha de pensamento achou uma formulação particularmente interessante na Morfologia Distribuída, que define, na saída da sintaxe, vários níveis em que regras se aplicam para produzir o output morfológico final. No artigo "From Intonational Phrase to Syntactic Phase: the grammaticalization of enclisis in the history of Portuguese", em co-autoria com Filomena Sândalo, apresentado no Going Romance 23 em dezembro passado, e submetido à revista *Língua* em fevereiro de 2010, assumimos que a ênclise é o resultado da aplicação dessas regras pos-sintáticas. A diferença ente o português clássico (PCl) e no português europeu moderno (PE) seria então o domínio da restrição que força a ênclise a ser produzida. Argumentamos que 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 PCl, 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. É também um modelo de reanálise: o que era percebido como o efeito de uma restrição de natureza prosódica, passa a ser interpretado como uma restrição mais sintática, que se aplica mesmo quando o sujeito precede o verbo, o que tem por corrolário uma reanálise da posição do sujeito, associada à perda de V2. Vê-se como essa análise é consistente com a idéia de que foi uma mudança prosódica que desencadeou a mudança sintática: a partir de um certo momento, a ênclise não é mais associada a uma fronteira entoacional imediatamente antes do verbo, e as crianças adquirindo o português, não percebendo mais essa associação reanalisam a regra que produz a ênclise, e reanalisam consequentemente a posição do sujeito pré-verbal.

Como abordagem complementar, temos procurado achar nos textos escritos os "vestígios" do ritmo. O texto "Ler a fonologia: do português clássico ao português moderno", em co-autoria com Sonia Frota e Marina Vigário, da Universidade de Lisboa, publicado no volume de *Artigos selecionados do XXIII Encontro Nacional da Associação Portuguesa de Lingüística* em 2008 procura evidências estatísticas de mudança na fonologia do ritmo ao longo da história do português (de um tipo mais silábico para um

tipo mais accentual), com base nos textos do Corpus Tycho Brahe. Estamos atualmente trabalhando numa nova versão desse artigo, com o titulo “The phonology of rhythm from Classical to Modern Portuguese” que procura validar os resultados muito encorajadores do primeiro texto, indo em duas direções: o aumento dos autores considerados, 16 em lugar de 8, e a modelagem estatística mais confiável, graças à colaboração de uma estatística da Unicamp, Veronica Gonzalez-Lopez.

O artigo “Context-tree selection and linguistic rhythm retrieval from written texts” em co-autoria com Antonio Galves, Nancy Garcia e Florência Leonardi, está atualmente em fase de revisão para publicação na revista *Annals of Applied Statistics*. Nesse 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. Infelizmente, essa metodologia ainda não deu resultado nos textos do Corpus Tycho Brahe, para discriminar um grupo anterior à mudança e um grupo posterior à mudança. Não conseguimos replicar, usando essa abordagem, os resultados positivos que obtivemos no trabalho comentado no Parágrafo anterior. Mas continuamos a trilhar por esse caminho.

5. Considerações finais

Para terminar, queria ressaltar os avanços extremamente positivos que este projeto, na sequência do anterior, nos permitiu realizar:

- O Corpus Tycho Brahe se consolidou como um instrumento de trabalho incontornável para quem quer estudar o português europeu a partir do séc. 16. É um dos poucos corpora anotados acessíveis na rede mundial de computadores a disponibilizar os textos na sua integralidade. Apesar da anotação sintática ainda ser limitada à parte do Corpus, já constitui uma massa considerável de dados, inexistente em qualquer outra base relativa à língua portuguesa desse período e de outros períodos históricos.[3]
- Graças ao Corpus Tycho Brahe, produziu-se um saber que não existia sobre essa fase do português, até então conhecida de maneira superficial no que diz respeito à sua sintaxe e à dinâmica das mudanças que levaram à gramática do português europeu moderno.

[3] Deve-se mencionar a existência de um Corpus dialetal anotado nos mesmos moldes, no âmbito do projeto Cordial-Sin coordenado na Universidade de Lisboa pela Profa. Ana Maria Martins. Temos mantido uma importante interação com esse projeto.

- Esse conhecimento fornece aos estudiosos do português brasileiro uma base sólida para entender melhor a língua que veio ao Brasil no início da colonização, bem como a língua que funcionou como modelo quando emergiu uma escrita brasileira.
- O projeto contribuiu a implantar no Brasil uma área de modelagem probabilística de fenômenos lingüísticos, ao exemplo do que acontece em outras áreas, como a biologia. Deu origem a novos projetos contemplados como “Padrões rítmicos, domínios prosódicos e modelagem de grande corpora” (Edital Universal CNPq 2007-2010).
- O projeto adaptou e produziu ferramentas para a construção de grandes corpora eletrônicos anotados, que estão sendo agora transferidos para outros grupos de pesquisa trabalhando sobre a história do português, numa perspectiva cada vez mais efetiva de interação e de trocas de competências.
- O projeto funcionou como um viveiro de pesquisadores. Na segunda fase, foram produzidas até agora 6 teses de doutorado e 5 dissertações de mestrado. 4 teses de doutorado e uma de mestrado estão em andamento (com uma defesa de mestrado e uma de doutorado ainda previstas para 2010). As pesquisadoras doutoras oriundas do projeto passaram em concursos em universidades públicas (2 na USP, 1 na Universidade Federal de Uberlândia, 1 na Universidade Federal do Rio de Janeiro e 1 na Universidade Estadual do Sul da Bahia) e várias delas estão implantando linhas de pesquisa derivadas da sua atuação e formação no Projeto.
- O projeto contribuiu fortemente a reforçar e internacionalizar a área de lingüística histórica no Brasil. Testemunho disso se acha no fato de a Unicamp ter sediado em 2009 o 11º Colóquio do DIGS (Diachronic Generative Syntax), que só tinha acontecido antes em universidades americanas européias ou canadenses.

6. Perspectivas

Os avanços elencados acima criaram um quadro e instauraram uma dinâmica que obviamente não se esgotam no Projeto que se conclui agora. O Corpus demanda extensões de conteúdo e de funcionalidades para se adequar cada vez melhor aos seus objetivos. E os resultados da pesquisa criam novas perguntas a serem respondidas em novas interações.

Dentro desse quadro, são várias as grandes perspectivas ligadas ao Projeto neste momento:

- **A extensão permanente do Corpus Tycho Brahe.**

Como tem sido até agora, essa extensão é conduzida por questões de pesquisa. Devemos continuar a produzir a anotação sintática de novos textos. Não só o grupo de anotadores que se constituiu se mostra disposto continuar o trabalho, como também novas pessoas têm me procurado para fazer parte da equipe. Num primeiro tempo, algumas lacunas devem ser preenchidas para constituir um corpus anotado mais representativo em termos cronológicos, para a pesquisa em geral, e para o projeto de livro que apresento a seguir em particular. A mais longo prazo, como mencionei acima, há também necessidade de incluir mais textos representativos de certos gêneros, como o teatro em geral e Gil Vicente em particular. Mas talvez a extensão mais espectacular do Corpus seja agora para o português brasileiro, em função de projetos que começaram a tomar corpo na vigência do atual projeto, e que têm a ver tantos com aspectos metodológicos, como com questões de pesquisa sobre a história do português brasileiro, no âmbito de uma colaboração sistemática com grupos afiliados ao Projeto para a História do Português Brasileiro, já concretizada em publicações realizadas ou em preparação.

- **A publicação de um livro sobre o português clássico (ou "médio" na definição proposta no texto inicial do projeto, ou "hispânico", como sugerido em mesa-redonda do Rosae).**

Esse livro está atualmente em fase de planejamento e deve tomar corpo ainda este ano, para ser publicado em 2011. A idéia é tomar como base os trabalhos que foram realizados nas teses e dissertações ligadas ao projeto, e complementá-los com novos dados oriundos do Corpus anotado. Será uma coletânea escrita por vários autores, mas procurando caracterizar os textos mais como verdadeiros capítulos de um livro do que como trabalhos individuais. A obra será dividida em duas partes: 1. A gramática do português clássico 2. A mudança para o português europeu moderno.

- **A elaboração de um novo projeto, voltado para uma história comparada do português europeu e brasileiro.**

Esse projeto é a continuação natural dos anteriores, e redireciona o Corpus Tycho Brahe em duas direções: 1. a criação de uma sessão brasileira do Corpus (que alguns já chamam de Tycho Brahe Brasil), em colaboração com os projetos mencionados acima, que deve abrigar vários tipos de textos produzidos no Brasil (literários e não literários), uma grande parte dos quais são textos editados no âmbito dos projetos

afiliados ao PHPB, 2. a inclusão no Corpus de textos portugueses comparáveis com esses. Isso coloca uma questão que não está encaminhada ainda: a eventualidade de cooperação com projetos portugueses engajados na edição e divulgação de textos dessa natureza, como – por exemplo - o projeto Cards coordenado por Rita Marquilhas no centro de Lingüística da Universidade de Lisboa.