Novos Rumos para a Pesquisa Lingüística no Brasil

Charlotte Galves,
IEL-UNICAMP

Em primeiro lugar, eu queria agradecer enfaticamente o convite que o professor Claudio me fez para fazer esta conferência de abertura. É uma honra muito grande, eu espero estar à altura das expectativas. Devo dizer que eu achei a iniciativa deste Seminário de Pesquisa extremamente interessante, e eu vou tentar, na minha apresentação, mostrar o quanto eu acho, pessoalmente, que o futuro da lingüística está na integração das pesquisas de vários domínios, de várias áreas.

O professor Cláudio me deixou livre para propor o título, e eu achei que era um tema adequado falar das novas perspectivas que a gente tem atualmente em lingüística, na pesquisa lingüística no Brasil. Eu tomei a liberdade de fazer isso e de ilustrar essa conferência com um trabalho que está sendo realizado no projeto de pesquisa de equipe que eu coordeno, do qual eu participo há vários anos, porque eu acho justamente que o projeto se caracteriza por uma grande pluridisciplinariedade, multidisciplinariedade. Trata-se de um projeto temático da FAPESP, que se caracterizam por articular vários horizontes, várias áreas, várias disciplinas, dentro da lingüística e também fora dela.

Então, essa pesquisa priorizou áreas que eu queria apresentar, argumentando que se encontram aí várias direções de pesquisas integradas extremamente interessantes para o futuro, em particular sobre a língua portuguesa, porque o objeto desse projeto é a língua portuguesa, a história da língua portuguesa, a comparação do português falado no Brasil e falado em Portugal. Pretendo mostrar como a gente está usando várias abordagens para tentar responder a velhas perguntas sobre o português.

Organizei a minha apresentação em três partes. Eu acho que as duas primeiras serão muito mais detalhadas, a terceira será mais uma indicação, e eu gostaria de ter tempo para a gente interagir um pouco. Um dos grandes desafios, e uma das tarefas importantíssimas que a gente tem agora pela frente, é uma tarefa que não é de hoje, mas temos hoje recursos fantásticos para realizá-la: é a elaboração e a exploração de grande *Corpora* de língua. Eu vou apresentar aqui um Corpus histórico do português. Podemos ter Corpora de todos os tipos, pois os recursos computacionais atuais extremamente poderosos nos permitem não só construir esses Corpora, elaborar esses Corpora, mas explorá-los de maneira interessante.

O segundo ponto da minha fala tem a ver com aquilo que eu chamei de articulação *língua externa/língua interna* e aqui vem uma proposta de integração teórica. *Língua externa*, são os enunciados, *língua interna* é a gramática. *Língua interna* é a gramática entendida de um certo ponto de vista, do ponto de vista da gramática chomskiana, que considera a gramática como órgão mental. Eu sou uma lingüista chomskiana, e durante muito tempo eu sei que eu fui olhada por colegas como quem não trabalhava com dados. Antes eu só trabalhava com algumas poucas frases, a partir das quais a gente fazia grandes elaborações teóricas. Eu acho que isso está mudando e para mim também. Eu aprendi a trabalhar com muitos dados, e eu acho que neste momento o lugar de integração é justamente entre trabalho com dados e trabalho cuja meta é formular gramáticas abstratas. O que eu queria argumentar aqui é que isso não deve ser mais visto como antagônico, como contraditório, mas, pelo contrário, é um lugar de integração extremamente produtivo para os estudos da linguagem.

E o terceiro ponto tem a ver com uma integração muito ambiciosa, eu já disse que dessa falarei menos e falarei um pouquinho mais hoje à tarde, aliás, eu quero agradecer ao professor Cláudio, que me convidou não só para fazer esta longa fala hoje de manhã, mas ainda para hoje à tarde, junto com membros da minha equipe, para falar numa mesa redonda. O terceiro ponto é um ponto que ainda é um pouco lingüística-ficção, mas nem tanto assim, é uma integração entre a lingüística e a modelagem matemática. A matemática modela tudo. Ela modela a física, ela modela a biologia, ela modela a química. Ela pode também nos ajudar a responder a perguntas sobre a linguagem. E o ponto que eu vou rapidamente mencionar, é a questão da identificação de padrões que vêm escondidos na linguagem e que a gente pode identificar, graças a abordagens de natureza matemática.

O projeto temático em questão se chama *Padrões rítmicos, fixação de parâmetros, e Mudança Lingüística*. O endereço da página do projeto é http://www.ime.usp.br/~tycho . Eu vou começar, portanto, falando desse primeiro ponto que eu chamei de *elaboração e exploração de grandes Corpora de língua*, apresentando a vocês um Corpus, que é um Corpus anotado, e eu vou explicitar o que isso significa, do português histórico. A esse Corpus, a gente deu o nome não de um lingüista, mas o nome de um astrônomo dinamarquês do século XVI, porque Tycho Brahe foi o primeiro astrônomo que resolveu fazer um catálogo exaustivo do movimento dos planetas no céu. Essa idéia de tentar fazer um catálogo exaustivo é que nos levou a usar o nome de Tycho Brahe. O Corpus Tycho Brahe tem atualmente 41 textos que vocês podem acessar livremente na rede virtual.

A questão inicial que nos levou a construir esse Corpus é a mudança ocorrida na colocação de clíticos entre o século XVI e o século XIX, nos textos portugueses. Partimos da hipótese de que essa mudança está relacionada com uma mudança fonológica que afetou a língua nalgum ponto do século XVIII.

Com efeito, pronúncia portuguesa atual não é a pronúncia portuguesa mais antiga, possivelmente a pronúncia brasileira atual está mais próxima da pronúncia do português do século XVI do que a pronúncia portuguesa atual. Então, nós temos na história do português, na história recente, uma mudança prosódica de natureza rítmica, que tem a ver com essa maneira que os portugueses têm de não pronunciar as sílabas não acentuadas, que torna para os brasileiros, às vezes, a compreensão do português europeu difícil.

Então, essa mudança rítmica aconteceu em algum momento entre o século XVI e o século XIX, no século XVIII, digamos. E a essa mudança rítmica se soma uma outra mudança, uma mudança sintática, que afeta alguma coisa que também no português brasileiro foi bastante afetada ao longo do tempo, que é a colocação de clíticos, a colocação desses pronomes que têm acento próprio e que nas línguas podem preceder ou seguir o verbo.

No português europeu moderno, em muitos contextos, o que a gente encontra é o clítico seguindo o verbo. Então, o português europeu é uma língua muito enclítica, enquanto o português brasileiro é uma língua muito proclítica. O clítico, quando ele aparece, aparece antes do verbo, e o que é interessante é que até o século XIX, nos textos, a gente tinha as duas coisas, uma variação entre próclise e ênclise.

Do ponto de vista da lingüística, é muito interessante, porque aí estamos na interface entre a fonologia e a sintaxe, entre a prosódia e a sintaxe, e não é tão óbvio que a teoria lingüística nos dê instrumentos para estudar isto. A teoria lingüística é muito compartimentada, tem os fonólogos, os sintaticistas, os semanticistas. Então, essa questão da articulação entre fonologia, prosódia, ritmo e a sintaxe é uma coisa que, de certa maneira, era mais fácil de trabalhar nos estudos da linguagem do final do século XIX e início do século XX do que agora, porque a teoria lingüística, de uma certa maneira, perdeu

essa possibilidade de dar conta das interfaces. Isso faz parte da evolução das disciplinas, já que elas têm uma tendência natural à especialização, mas de novo eu acho que a gente está num momento de integração.

Então, a primeira coisa que precisávamos fazer era ter uma visão clara de como evolui essa sintaxe da colocação de clíticos do século XVI ao século XIX. Temos algumas informações sobre isso, alguns lingüistas portugueses, italianos, enfim, algumas pessoas trabalharam sobre isso aqui no Brasil também, mas nunca com base numa grande quantidade de dados. Nós fizemos esse Corpus para termos essa grande quantidade de dados. Ele reúne autores nascidos desde o comecinho do Século XVI (O primeiro é João de Barros, nascido em 1497, que escreveu a gramática da língua portuguesa) até autores nascidos em meados do século XIX, como Eça de Queirós e Oliveira Martins.

No Corpus, temos um lugar de integração, articulação, interface, com a literatura portuguesa e com a filologia também, porque quando a gente faz um Corpus dessa natureza, uma das primeiras questões que se colocam é a questão da escolha das edições. Nós, ao longo dos quatro anos que esse Corpus demorou a ser construído, passamos por vários momentos em relação à questão das edições. Primeiro, tivemos um grande auxílio para escolher edições confiáveis de textos dessa época, foi o auxílio do professor Ivo Castro, da Universidade de Lisboa, e da professora Ana Maria Martins, da Universidade de Lisboa também, que é filóloga e sintaticista. Estávamos preocupados em ter textos que fossem confiáveis do ponto de vista sintático, obviamente, porque a nossa questão é essencialmente uma questão sintática e sobretudo em que o revisor eventual não tivesse mudado a posição dos clíticos, o que seria dramático, o que os revisores atuais fazem nas editoras. Então, alguém que vai querer estudar a posição dos clíticos, no século XX, do português brasileiro, vai ter muitas dificuldades, porque os revisores, a primeira coisa que eles fazem é mudar o clítico de lugar. A nossa preocupação era termos textos que tivessem a colocação que o autor tivesse posto.

No início só usamos edições modernas, com grafia modernizada ou não. Por exemplo, passando rapidamente pelos autores, temos D. João III com as suas cartas, a peregrinação do Fernão Mendes Pinto, um tratado de pintura, as décadas de Diogo do Couto, a biografia de Frei Luís de Sousa, de Frei Bertolameu dos Mártires, enfim, textos de historiografia, biografias, cartas, textos religiosos, textos filosóficos, um dos primeiros jornais portugueses, senão o primeiro, esta gazeta de Manuel de Galhegos, que é escrita logo depois da restauração da monarquia portuguesa, e, alguém de quem está se falando bastante aqui neste encontro, o Padre Vieira.

O Padre Vieira é muito importante nesse Corpus, porque como eu vou depois mostrar para vocês, nos seus Sermões ele tem uma colocação de clíticos muito diferente dos outros autores da época dele. Isso foi uma das questões iniciais: por que é que o Padre Vieira é tão enclítico nos Sermões? Nos Sermões, ele usa muita ênclise, ou seja, o clítico depois do verbo, quando os seus contemporâneos são muito proclíticos. Em compensação nas suas cartas, ele é muito proclítico, tão proclítico quanto os seus contemporâneos. Descobrimos então que Vieira não só era diferente dos outros, mas era diferente de si mesmo nos Sermões e foi durante muito tempo um grande ponto de interrogação, e como acontece muitas vezes com os pontos de interrogação, acabou se tornando uma explicação. Voltarei a isso depois.

A questão da grafia antiga é um problema grande para quem quer fazer um Corpus eletrônico anotado do português antigo. Por quê? Porque nós, como eu vou mostrar depois para vocês, vamos anotar esse Corpus, porque todo Corpus grande precisa ser anotado, senão é muito difícil recuperar a informação. Não adianta ter vários milhões de palavras e não ter como extrair a informação desses milhões de palavras, o que à mão se torna totalmente impossível. Então, precisamos ter uma anotação, e precisamos de

ferramentas automáticas de anotação, porque anotar à mão milhões de palavras é uma tarefa inglória.

Então temos um etiquetador, ou seja, um programa que atribui a cada palavra dos textos do Corpus uma etiqueta. Aí se encontra uma interface importante com a computação. Precisamos elaborar, e ter quem elabore para a gente, ferramentas computacionais que nos permitam trabalhar com essas grandes quantidades de dados. Então, vejam aqui, a versão anotada do texto da Maria do Céu, onde podem ver que cada palavra vem com uma barra, e uma etiqueta, por exemplo, /VB-G para "verbo no gerúndio", /D-F-P para "determinante feminino plural", /N-P para "nome no plural", /VB-AN-F para "particípio passado no feminino", /P para "preposição", /ADJ-F para "adjetivo no feminino", P+D para "preposição mais determinante", etc...

```
Misturando/VB-G as/D-F-P lagrimas/N-P com/P a/D-F tinta/N pella/P+D-F saudade/N ,/, dar
```

É um etiquetador automático que faz isso. Ele atualmente acerta 95% dos casos, o que é muito bom, mas quer dizer que ele erra ainda em 5% dos casos, e 5% dos casos, em 50 mil palavras, são 2.500 palavras, e geralmente 2.500 palavras importantes. Então, ainda fazemos todo um trabalho de correção manual, mas mesmo assim a correção manual é muito menos trabalhosa do que se tivessemos que fazer a anotação manual inteira, tanto mais que a vantagem do homem sobre a máquina é que ele raciocina, a desvantagem é que ele erra de maneira não-sistemática. Quando a pessoa está cansada, ela começa a errar, a escrever, por exemplo, etiquetas impossíveis. O etiquetador nunca escreve etiquetas impossíveis, podem ser erradas, mas impossíveis nunca são, porque ele se baseia uma lista finita. Isso é a vantagem da máquina.

Vejam o início do texto da Maria do Céu, que é um texto muito bonito, *misturando as lágrimas com a tinta pela saudade, dando vôos à pena pelo assunto, pedindo oração ao papel pela memória*. Bom, o prazer do texto a gente tem também, isso é uma coisa interessante. *Escreve em digna vida desta ilustríssima serva de Deus.* Agora, vejam uma coisa, o etiquetador tem muitas dificuldades criadas pela variação gráfica. É por isso que eu estava dizendo que é um problema trabalhar com um Corpus eletrônico anotado automaticamente quando existe variação gráfica.

Nós estamos trabalhando com um teórico da computação da USP, Marcelo Finger. Foi ele que elaborou esse etiquetador para gente, e estamos trabalhando com ele para dar conta da variação gráfica, que na realidade é um problema terrível computacionalmente, porque a variação gráfica do português é muito grande. O português é uma língua que demorou muitíssimo para fixar sua grafia, aliás, ainda nem está totalmente fixada, mas, até fins do século XVIII, a gente ainda tem textos com uma grafia totalmente diferente da grafia moderna, e o etiquetador só conhece uma forma para cada palavra. Por exemplo, vocês têm *pela*, *pela* ele sabe que é /P+D-F enquanto está escrito *p-e-l-a*, agora quando ele vê *p-e-l-l-a,* para ele já não existe, então, ele procura pelo contexto, porque ele trabalha com o contexto e com a forma, mas, enfim, ele vai errar muito mais.

Nesse texto, a eficiência do etiquetador baixou muitíssimo, e eu que corrigi esse texto, demorei o dobro do tempo que eu demoro para um texto modernizado. Então, o que fizemos mais recentemente? Nós, recentemente, resolvemos usar edições originais.

Esse é o caso, por exemplo, do texto de André de Barros, que é muito interessante, porque é a vida do padre Antônio Vieira. É a primeira biografia do Vieira. E André de Barros nasceu numa época que nos interessa muito, que é a segunda metade do século XVII. Nós pegamos na Biblioteca Nacional de Lisboa o xerox da primeira edição desse texto, que data de 1746, e fizemos a modernização do texto nós mesmos.

E a idéia é trazer depois, isso não está ainda no Corpus que está disponível na rede, o fac-símile do texto original, ou simplesmente pôr um ponteiro para a Biblioteca Nacional de Lisboa.

A Biblioteca Nacional de Lisboa está atualmente disponibilizando muitos textos em fac-símiles. Isso facilitou muitíssimo a nossa tarefa - com o defeito que a gente nem precisa mais ir a Lisboa buscar os textos... - mas facilitou muito. Pegamos também um outro texto, que é uma outra gramática que vocês podem achar no Corpus, a gramática de Jerônimo Contador de Argote, que é uma gramática extremamente interessante, publicada na primeira metade do século XVIII, sob a forma de diálogo entre um mestre e seu discípulo, e fizemos a mesma coisa, modernizamos o texto, ou seja, fizemos o seguinte: modernizamos a grafia, porque esse é nosso problema, mas não mexemos em mais nada, em particular, não modificamos a pontuação. Fazemos o contrário do que fazem muitos editores de textos antigos, eles deixam muitos aspectos da grafia antiga, não todos, alguns, e eles mexem sistematicamente na pontuação. E eu sempre me perguntei por que é que mexiam na pontuação, porque a pontuação nos traz informações muito preciosas.

Certamente com esses textos eu entendi porque eles mexem na pontuação. A pontuação dos séculos XVII e XVIII é muito diferente da pontuação moderna, e ela chega a tornar difícil a leitura dos textos, mas eu acho que é muito interessante, porque a gente vê que quando lê em voz alta, de repente, a coisa flui. Isso mostra que se trata de uma pontuação mais entoacional, prosódica, retórica, do que a pontuação moderna, que é uma pontuação de tipo lógico-semântico, sintático-semântico. E de fato, é uma coisa que complica às vezes a leitura da gente, mas vale a pena ter acesso a essa pontuação original.

Darei agora alguns elementos do nosso sistema de anotação, remetendo os auditores/leitores interessados numa apresentação exaustiva para o *Manual de anotação morfológica do Corpus* (cf. http://www.ime.usp.br/~tycho/manual). A anotação é uma tarefa muito interessante, porque é uma tarefa que nos lembra muito tarefas bem tradicionais de análise morfossintática. Em grande parte, é relativamente fácil, mas em alguns lugares não é óbvio escolher as melhores categorias para descrever adequadamente o português. Elaboramos um sistema de anotação morfológica no qual vocês vão reencontrar muitos termos que são os termos tradicionais, e mais alguns, que são mais modernos. Ou seja, nesse trabalho, estamos integrando abordagens mais tradicionais e abordagens mais modernas, nomeadamente advindas da gramática gerativa. Notem que não se trata, quando se anota o texto, de fazer uma análise do texto. Análise do texto quem vai fazer são as pessoas que vão vir buscar os dados dentro do texto. O que se trata de fazer é dar a possibilidade aos pesquisadores de recuperar informações facilmente.

Por exemplo, se eu estou interessada na história dos clíticos, eu quero poder recuperar muito rapidamente todas as orações que contêm um pronome clítico. E, melhor, eu quero até poder recuperar muito facilmente todas as orações que contêm um pronome clítico à direita do verbo e todas as orações que contêm um pronome clítico à esquerda do verbo, e já jogar essas orações para arquivos separados. E com algumas pequenas ferramentas computacionais, é muito fácil fazer isso, ou seja, em três minutos a gente consegue extrair de um texto de 50 mil palavras dois arquivos distintos: um que tem toda a ênclise e um que tem toda a próclise.

Os verbos são divididos em cinco tipos: *ser, estar, haver, ter* e os outros verbos todos. Então, alguém que quer fazer um trabalho, por exemplo, sobre o uso do gerúndio por oposição ao infinitivo nas locuções *estar fazendo* versus *estar a fazer*, na história do português, pode recuperar muito facilmente todas as orações que têm "*estar* mais gerúndio" ou "*estar a* mais verbo no infinitivo". É por isso que deixamos esses verbos auxiliares com uma etiqueta diferente. Como mencionei e já exemplifiquei antes

para outras categorias, também temos sub-etiquetas para os verbos, que são as etiquetas que anotam a morfologia verbal. O infinitivo flexionado, o imperativo presente, o subjuntivo presente, o passado, então, tudo isso para cada tipo de verbo a gente vai ter.

Temos os pronomes, e o que nos interessa particularmente, atualmente, são os pronomes clíticos. Por isso, os clíticos têm etiquetas diferentes dos pronomes tônicos, a etiqueta /CL. E o clítico *se,* por sua vez, sendo um assunto de sintaxe do português muito trabalhado, tem uma etiqueta diferente dos outros clíticos, a etiqueta /SE. Assim podemos recuperar facilmente todos os *se* que estão no Corpus. Os outros pronomes, como *eu, tu, ele, ela,* vão aparecer com a etiqueta /PRO. Enfim, para as formas compostas de uma preposição mais um pronome, usamos o símbolo +, que aparece em toda as palavras que são o resultado da contração de duas categorias distintas . Por exemplo, *comigo* vai ser /P+PRO, "preposição mais pronome", *fazê-lo* vai ser /VB+CL, "verbo no infinitivo mais clítico" etc....

As partículas de foco, como *só*, *mesmo*, *até,* são marcadas com uma etiqueta especial /FP, porque a focalização também é um assunto muito interessante. Notem que todas essas palavras podem ter outras funções. A*té*, por exemplo, pode ser também uma preposição, *mesmo* e *só* podem também ser adjetivos. Isso acontece freqüentemente, a mesma forma pode ter várias funções e portanto várias etiquetas possíveis: *como* pode ser uma conjunção (/CONJS) ou uma palavra interrogativa ou relativa (/WPRO), *melhor* e *pior* podem ser advérbio comparativo (ADV-R) ou adjetivo comparativo (/ADJ-R). *Que* é uma palavra particularmente complicada, que pode ser muitas coisas, pode ser uma conjunção integrante, que a gente marca com /C, pode ser um *que* explicativo, que a gente encontra muitíssimo nos textos, e marca como /CONJ ou pode ser um pronome relativo ou interrogativo e vai ser marcado /WPRO.

Para algumas palavras, em compensação, decidimos manter uma só etiqueta, por exemplo, *mais* e *menos* sempre etiquetamos como um advérbio de comparação (/ADV-R), mesmo quando *mais* parece exercer uma função nominal como em *os mais*, ele vai ser marcado só com essa etiqueta.[1]

Nós passamos um certo tempo, a partir dos textos, lendo os textos, etiquetando os textos, foi uma equipe que fez isso e elaborou esse sistema. Ele não é completamente satisfatório, mas pelo que a gente achou até agora, representa um compromisso bastante satisfatório entre a descrição lingüística e a complexidade computacional.

Num segundo momento, o Corpus vai receber um segundo tipo de anotação, da qual falarei pouco, porque será apresentada por Helena Britto na mesa redonda desta tarde.[2] É uma anotação sintática. Porque já é ótimo poder recuperar todas as frases com clíticos, mas precisamos poder extrair automaticamente do Corpus mais informações ainda. Por exemplo, queremos todas as frases que tenham um verbo seguido de um clítico e precedido de um sujeito, só que sujeito aí não está marcado, porque sujeito não é uma categoria, sujeito é uma função sintática. Então, o ideal é ter o texto também marcado quanto às funções, e isso é muito mais complexo e muito mais difícil de ser feito com ferramentas automáticas. Só darei aqui um exemplo de como isso pode ser feito. Trata-se das primeiras palavras da primeira frase das *Reflexões sobre a vaidade dos homens* de Matias Aires (1705-1763):

Ofereço a Vossa Majestade as reflexões sobre a vaidade dos homens:

```
(1)  Senhor/NPR :/. Ofereço/VB-P a/P Vossa/PRO$-F Majestade/NPR[i][5]

(2)  (IP-MAT (NP-VOC (NPR Senhor))
             (. :)
            (NP-SBJ *pro*)
            (VB-P Ofereço)
            (PP (P a)
                 (NP (PRO$-F Vossa)
```

(NPR Majestade)))

(In Matias Aires (1705-1763))

Primeiro, em (1), temos as etiquetas: /V-P "verbo no presente", /P "preposição", /PRO$ "pronome possessivo", /NPR "nome próprio", etc. A partir disso o analisador vai nos dar a análise sintática da frase como em (2). Não se trata de uma análise aprofundada, não se trata aqui, de novo, de substituir o sintaticista, trata-se de dar ao sintaticista a possibilidade de extrair do Corpus o tipo de sentença que ele está interessado em analisar.

As funções da oração são inseridas nesse nível. No exemplo acima, temos três sintagmas nominais (NPs), um é marcado vocativo, o segundo sujeito, o terceiro objeto de preposição. Introduzimos também informações sobre o tipo de orações presentes na frase. IP é um termo que vem da gramática gerativa, e quer dizer *oração*, IP-mat quer dizer *oração matriz*, *oração principal*. Uma outra informção essencial inserida nesse nível é a presença de um sujeito nulo, marcado *pro*. O português é uma língua que pode omitir o pronome sujeito, e é muito importante que possamos recuperar essa informação quando analisamos sintaticamente uma oração.

Como podem observar, o analisador sintático nos dá uma árvore, alguma coisa que tem uma disposição gráfica, que representa as relações hierárquica dentro da oração. Por que fazemos isso? De novo, eu insisto, é para poder recuperar as informações sintáticas pertinentes a partir de todas as orações do Corpus. Por exemplo, alguém que trabalha sobre a posição do sujeito - para um sintaticista a posição do sujeito em português é uma questão muito interessante - pode muito rapidamente extrair de todos os textos todas as frases que têm o sujeito anteposto ou o sujeito posposto.

Então, esse trabalho é um longo trabalho, mas de uma certa maneira a gente trabalha para o futuro, porque uma vez que está feito, a possibilidade de trabalhar com muitos dados se torna uma realidade.

Apesar de o Corpus ainda não ter todas as funcionalidades que a gente gostaria que ele tivesse, a gente, como diz um colega meu, está navegando construindo o barco. Então, estamos usando o Corpus ao mesmo tempo que continuamos a construí-lo, tentando enfrentar essa questão que eu coloquei no início, que é a questão da evolução da colocação de clíticos no português europeu, e a relação dessa colocação dentro dessa evolução com uma possível mudança rítmica que aconteceu entre o século XVI e o início do século XIX.

Um dos desafios é conseguir, a partir dos dados que a gente tem no Corpus, e olhando para a evolução da colocação de clíticos, ter uma idéia de quando aconteceu a mudança. Isso é obviamente um objetivo muito ambicioso, mas é isso que estamos tentando fazer. Para isso, criamos uma base de dados a partir dos 20 textos anotados, dos 41 que já estão no Corpus. Poderíamos anotar todos, porque o etiquetador automático demora 5 minutos para realizar essa tarefa, mas a questão é a correção. Então, nós já temos 20 textos já corrigidos, isso não quer dizer que não tenham nenhum erro, mas têm poucos. A equipe que tem trabalhado na base é indicada no index da página[3], Helena Britto, pós-doutoranda, Maria Clara Paixão de Sousa, doutoranda, Sílvia Regina Cavalcante, doutoranda, Cristiane Namiuti, que começou como bolsista de iniciação científica, agora mestranda, e Lucianne Chociay, bolsista de iniciação científica. O que é muito interessante nesse trabalho é que tem espaço para trabalhos em vários níveis. Eu acho que é um trabalho de equipe que é muito bom para a formação dos alunos, e eles podem entrar rapidamente e aprender muita coisa e fazer um trabalho que é muito importante para o projeto como um todo.

Criamos então essa base de dados.

The Tycho Brahe Corpus Database

[main] [manual] [data files] [related papers] [corpus main page]

arquivos de dados

modificado em 30/10/2002
M.C. Paixão de Sousa

---

formato html (em quadros)

[V1]

[V2 e V3]

[subordinadas]

---

somente texto e planilhas:

.txt	.xls
[V1.txt]	[V1.xls]
[V2_V3.txt]	
[suj-clv.txt] [suj-Vcl.txt]	[V2_V3.xls]
[adv-clv.txt] [adv-Vcl.txt]	[subordinadas.xls]
[pp-clv.txt] [pp-Vcl.txt]	
[or-clv.txt] [or-Vcl.txt] (em construção)	
[x-clv.txt] [x-Vclv.txt]	
[subordinadas.txt]	

Nessa base, classificamos inicialmente as orações em função da posição do verbo (V1, V2, V3). V1 quer dizer que são orações que começam pelo verbo. E por que é que as orações que começam pelo verbo ficam à parte? Porque nesse caso vocês sabem todos que na história do português europeu, desde o século XII ao século XX, quando o verbo está em primeira posição, o clítico está sempre depois do verbo. Isso é a famosa lei de **Tobler Mussafia**, que a norma brasileira ainda quer impor, mas que na realidade no Brasil se perdeu na fala, mas em Portugal não se perdeu, as pessoas realmente falam assim mesmo. Portanto, as frases que têm o verbo em primeira posição não nos interessam, porque nesse caso não há variação na posição do clítico, ele está sempre depois do verbo. As subordinadas também não nos interessam, porque fora algumas exceções, que são interessantes, mas são marginais, nas subordinadas, em toda a história do português, o clítico vem sempre antes do verbo, então não nos interessa, porque a gente quer ver a variação e aí não há variação.

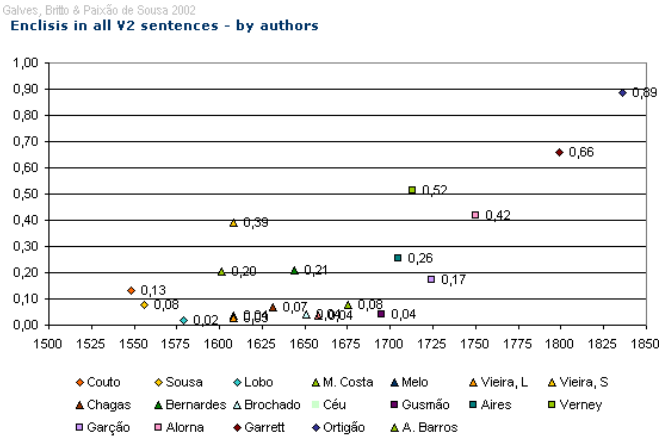
O que vai nos interessar são os casos que a gente chama de V2 e V3, ou seja, quando o verbo está em segunda ou terceira posição e temos variação[4]. Vejam essa variação no P^e^ António Vieira: *Eles conheciam-se como homens, Cristo conhecia-os como Deus.* E aqui a gente tem a colocação enclítica, que

é bastante freqüente nos sermões, de novo: *Deus julga-nos a nós por nós, os homens julgam-nos a nós por si.* Aí o que a gente tem em primeira posição é o sujeito: *Eles/ Cristo/ Deus/os homens.* Mas, a gente também pode ter em primeira posição, logo antes do verbo, um sintagma preposicional, como em: *Entre as feras tomava-se com os leões e entre os homens com gigantes*.

A gente pode ter também um advérbio antes do verbo, vocês podem ver isso em: *Aqui vêem-se as suas dependências*, também com ênclise. Mas, em todos esses contextos, podemos ter também próclise, ou seja, a outra colocação, a colocação pré-verbal. *O evangelho o diz*, isso, por exemplo, é uma colocação impossível no português europeu moderno, atualmente, e quando eu digo impossível não é porque a norma diz que não pode, é que as pessoas não fazem isso, realmente não fazem. Então, essa próclise que eu tenho aqui em *O Evangelho o diz*, *o mesmo texto o declara*, ou seja com sujeitos pré-verbais, mas também com sintagmas preposicionais: *Doutras se lavram, semeiam e plantam os mesmos lugares*. Vê-se portanto em Vieira essa variação na colocação de clíticos, quando o verbo está em segunda posição e, seja qual for o que preceda o verbo, pode ser um sujeito, pode ser um sintagma preposicional, pode ser um advérbio, pode ser uma oração também. E nós fizemos, portanto, uma base de dados exaustiva com todos esses casos de variação nos 20 textos que já estão etiquetados. Ao todo são 3030 dados.

Então o Corpus nos permite fazer um trabalho exaustivo desse tipo, e é isso que é a vantagem de trabalhar assim. E é por isso que invocamos Tycho Brahe, que olhava para os planetas todas as noites.

A partir disso, nós fizemos uma quantificação, e vocês podem ver aqui a imagem, a partir dos textos que a gente tem, da evolução dessa variação.



As datas que estão aqui são as datas de nascimento dos autores. A gente faz referência à data de nascimento por duas razões: primeiro, porque a gente acredita que a gramática é alguma coisa que se constrói na aquisição, então, a data de nascimento é importante, apesar de a gente saber que quando escrevemos usamos uma língua que incluí saberes adquiridos depois, na escola em particular, no contato com textos escritos, que são mais conservadores. Mas, isso dito, é de fundamental interesse saber qual é a data de nascimento dos autores. A outra razão é porque nesse Corpus, atualmente, a gente só tem textos de autores e, muitas vezes, a única maneira objetiva que temos de datar esses textos é a própria data de nascimento dos autores. Às vezes, não sabemos exatamente quando o texto foi escrito e, às vezes, usamos correspondências que se estendem por vários anos. Por exemplo, a correspondência do próprio Vieira, que vai de 1642 a 1697, o ano em que ele morreu, ou a correspondência de um autor português bem mais recente: Ramalho Ortigão, que se estende 1875 a 1915.

Então cada ponto corresponde a um autor, em função da sua data de nascimento, e, aqui, é a percentagem de ênclise que a gente tem nos contextos em que existe essa variação entre ênclise e próclise.

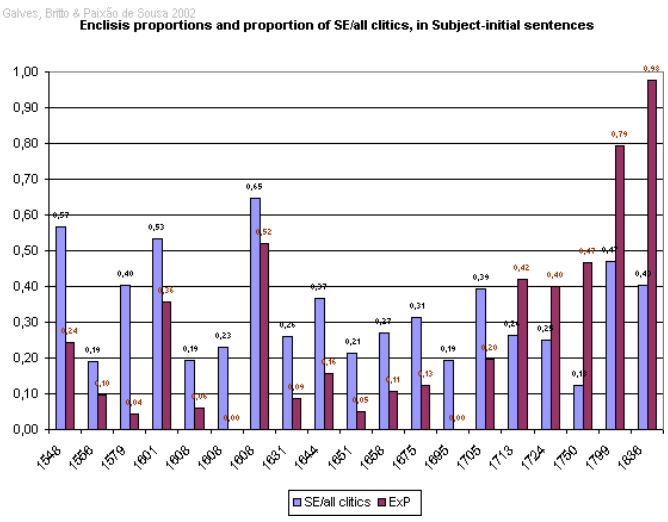
Vocês podem ver uma coisa que eu não vou ter tempo de comentar em detalhe, mas que se verifica em todas as codificações que a gente tem feito até agora. No período considerado, podemos claramente distinguir duas partes. Na primeira parte, que vai mais ou menos até 1700, vêmos que a maior parte dos autores e dos textos tem uma taxa de ênclise muito baixa. Ou seja, os autores são, geralmente, extremamente proclíticos, mas temos alguns que são mais enclíticos que os outros, obviamente. O ponto mais alto é Vieira nos seus sermões. Vocês vêem realmente que Vieira usa muito mais a ênclise nos sermões, mas se olharem para Vieira nas suas cartas, então, verão que, até para um mesmo autor, podemos ter essa grande diferença que eu estava mencionando no início.

Mas observem que Vieira nas suas cartas é mais condizente, digamos, com a maior parte de seus contemporâneos. Agora a partir do início do século XVIII, o que a gente vê? A gente vê uma subida, obviamente essa subida é a subida para o português europeu moderno, no fim da qual um autor como Ramalho Ortigão, nascido em 1836, vai usar ênclise em 90% dos casos, nos contextos em que o português do século XX ou XXI encontramos 100% de ênclise.

Portanto o que nós temos nessa história da colocação de clíticos, no português europeu, entre autores nascidos entre 1550 e 1850, são dois momentos: um momento em que a variação não configura uma mudança e um momento em que a gente vê uma mudança acontecer.

Estamos trabalhando com um conceito de mudança muito interessante, inicialmente proposto por Anthony Kroch**,** que é um lingüista que é, ao mesmo tempo, sociolingüista e gerativista, e que é alguém que trabalha muito sobre a história das línguas em geral, e do inglês em particular. A idéia que ele defende é que uma mudança não acontece quando a gente vê o fim da mudança. A mudança aconteceu quando a gente vê o início da mudança, porque na realidade, nos textos, o que se tem é uma competição entre a nova gramática e a gramática antiga. Por quê? Porque os textos são conservadores. Pensem como nós escrevemos e como nós falamos. Nos textos de autores brasileiros do século XX, XXI ainda vemos aparecer a gramática do século XVIII. Então, o que a gente vê neste gráfico é que é possivelmente no finzinho do século XVII ou início do século XVIII que se dá a mudança na colocação de clíticos, porque é o início dessa curva ascendente que nos leva para a situação moderna.

O caso do clítico *se* é muito interessante, porque a gente vê muito mais ênclise com *se* do que com os outros clíticos. No gráfico a seguir, onde as colunas vermelhas mostram a proporção de ênclise versus a próclise e as colunas azuis representam a proporção de *se* em relação aos outros clíticos, isso aparece claramente:



O que se vê é que, pelo menos até o final do século XVII, quando a proporção de *se* é maior do

que a média, temos também mais ênclise. Isso mostra claramente uma correlação entre ter muito *se* nos textos e ter ênclise. Obviamente, a partir do século XVIII essa correlação se perde . Vocês podem ver que o *se* aqui se mantém em 40%, mas a ênclise vai subindo, subindo, subindo, obviamente, já não tem nada a ver com a escolha do clítico, tem a ver com a mudança que está se implementando nos textos.

Eu disse no início: temos que reconciliar língua externa e língua interna. Tudo o que eu mostrei até agora é o que a gente pode chamar de língua externa, são dados, muitos dados quantificados. Agora, queremos interpretar esses dados à luz de hipóteses sobre as gramáticas, que estão subjacentes aos dados. Vou apresentar agora a hipótese com a qual o projeto trabalha. Essa hipótese é que vai nos permitir ligar a questão sintática com a questão prosódica, porque a hipótese com a qual trabalhamos é que na gramática, que eu vou chamar de português clássico, que é a gramática que a gente tem no século XVI e XVII, temos variação entre próclise e ênclise, porque temos duas estruturas subjacentes possíveis. No caso da próclise, todos esses elementos que podem aparecer antes do verbo (sujeito, sintagma preposicional, advérbio, oração etc...), ocupam uma posição interna à oração. Eu não estou agora querendo dizer qual é exatamente essa posição, mas é uma posição *interna à oração*.

Agora, no caso da ênclise, temos esse mesmo elemento, sujeito ou qualquer outra coisa, que ocupa uma posição *externa à oração*. Estas duas possibilidades podem ser representadas da seguinte maneira, onde XP representa qualquer sintagma e [é a fronteira de oração:

[XP cl-V

XP[V-cl

Ou seja, a idéia é a seguinte: quando no século XVI, XVII, há uma variação entre ênclise e próclise, é porque eu posso eventualmente pôr esse sujeito ou esse advérbio ou sintagma preposicional fora da oração. Fazendo referência às orações de Vieira mostradas anteriormente, é como se eu tivesse *Deus,* ou *Os homens,* e só depois é que a oração começa, com o verbo em primeira posição. Então, eu tenho ênclise, porque o verbo está em primeira posição na oração, e aí vejam, agora estou falando de estrutura, de gramática, de língua interna. O que eu vejo é o verbo em segunda posição, mas eu estou dizendo, sim, eu vejo o verbo em segunda posição, mas, na realidade, ele está em primeira posição e, o que vem antes, está fora dos limites da oração. Obviamente, isso é uma hipótese, mas é uma hipótese que nos permite explicar, entre outras coisas, o que a gente vê nos sermões do Vieira. Eu disse que os sermões de Vieira, inicialmente, eram um mistério, e que depois eles nos deram uma chave. É que olhando de maneira mais detalhada para os sermões, o que eu percebi é que TODOS os casos em que encontramos sujeito, verbo, clítico, são casos claros em que Vieira está contrastando dois termos.[5] Nos exemplos que eu mostrei para vocês, é bem claro: *Eles conheciam-nos como homens, Cristo conhecia-os como Deus*. Aqui, a gente tem uma oposição entre *eles* e *Cristo*. *Deus julga-nos a nós por nós, os homens julgam-nos a nós por si*. Oposição entre *Deus* e *os homens*. O mesmo acontece com os sintagmas preposicionais iniciais. *Entre as feras tomava-se com os leões e entre os homens com gigantes*. Oposição *Entre as feras/entre os homens*. Isso acontece em 100% dos casos, e é muito raro achar 100% quando se procura alguma coisa. Em 100% dos casos em que um sujeito é seguido do verbo seguido do clítico, nos sermões do Vieira que temos no Corpus, encontramos um caso de oposição entre dois termos. Podemos então concluir que esse sujeito, ou esse sintagma preposicional, é o que se chama de tópico contrastivo. É um tópico porque é o termo sobre o qual se faz a asserção, mas ele é contrastivo, porque é contrastado com um outro termo do texto.

Então, em Vieira, nos sermões do Vieira, o que explica a ênclise é a topicalização contrastiva. Esse tipo de topicalização é muito recorrente nos sermões, porque são textos de estilo barroco, e aí temos um

ponto de contato interessantíssimo com a literatura. O barroco é baseado em oposições. O que descobrimos então é que Vieira usa a colocação de clíticos como um recurso estilístico.

Obviamente, na gramática do português europeu moderno, nada disso se verifica, a ênclise se torna o padrão absoluto, e a interpretação é que nessa gramática sujeitos e tópicos não ocupam a mesma posição. Os tópicos são externos, mas o sujeito é sempre interno à oração.[6] E mesmo assim temos ênclise, o que configura uma gramática totalmente diferente. Resumindo, desse ponto de vista, a mudança na gramática afeta a posição do sujeito e pode ser resumida da seguinte maneira no que diz respeito às orações com ênclise (onde [, novamente, simboliza a fronteira da oração):

Gramática 1 (português clássico): Sujeito [V-cl

Gramática 2 (português europeu moderno): [Sujeito V-cl

Tenho dois minutos para falar da procura dos padrões escondidos, e eu já tinha dito que eu teria muito pouco tempo para falar disso, já que nessa fala, mesmo longa, não dá para mostrar por onde tentamos ligar essa análise sintática com uma análise prosódica, que tal maneira que possamos responder à questão inicial, que era: é verdade que a mudança rítmica do português provocou uma mudança sintática? E uma primeira coisa que a gente tem que fazer é definir de que mudança rítmica se trata, e por isso nós fizemos, dentro desse projeto, todo um trabalho, comparando o português europeu moderno e o português brasileiro moderno. Usamos o português brasileiro como imagem do português falado do século XVI. Atenção, no que diz respeito à sintaxe ele é muito diferente, mas, possivelmente, o ritmo da fala brasileira é muito mais próximo do ritmo da fala do século XVI do que o ritmo português moderno. Citarei aqui um foneticista português do século XIX, Gonçalves Viana, que dizia que os atores da época dele eram incapazes de ler Camões de maneira correta, porque eles comiam algumas sílabas a caminho, e os decassílabos do Camões se tornavam heptassílabos, no melhor dos casos. Em compensação, os brasileiros lêem Camões muito bem, ou seja, o ritmo da fala brasileira é, certamente, muito mais próximo do ritmo da fala do século XVI.

Fizemos um grande trabalho de comparação de ritmo português europeu/ português brasileiro. Mas num segundo momento, precisamos de modelagem, que não vem atualmente da lingüística, e pode ser que nunca venha da lingüística, porque fazer fonologia histórica é muito difícil, no sentido que não temos nenhuma evidência a partir do texto escrito de como as pessoas falavam. Mas, esperamos ser capazes, com métodos estatísticos, de reconhecer padrões, ou seja, tipos de seqüências que a gente encontra nos textos portugueses modernos e brasileiros modernos, fazer um modelo matemático desses padrões, e depois aplicar esses modelos a Corpora antigos. Estamos atualmente numa fase ainda preliminar, já com algum sucesso. Se conseguirmos avançar nesse caminho, talvez a gente consiga achar a prosódia perdida, que é o objetivo mais ambicioso de toda essa pesquisa.

Muito obrigada.

Claudio C. Henriques (mediador) – Passo a palavra para o professor André Valente, que deseja fazer algumas considerações.

André Valente – Primeiro, quero parabenizar a professora Charlotte pelo trabalho e pela generosidade de socializá-lo. Tenho duas considerações: a constatação de que, na segunda metade do século XVI, predominava a próclise, o que colabora para desmistificar algumas afirmações precipitadas. Comprovou-se que Vieira, nos sermões, trabalhava mais com a ênclise e, nas cartas, com a próclise. Na interpretação dos dados haveria alguma dependência dessa busca prosódica ou vocês já estão trabalhando a

justificativa dessa evolução? Essa é a primeira pergunta. A segunda: ficou claro que o objetivo é trabalhar a oposição ênclise e próclise, e me chamou a atenção a presença da mesóclise em dois casos apresentados: *dar-te-ei* e *entregar-me-ei*. Isso é tratado de que forma na pesquisa?

Charlotte – Em relação à sua primeira pergunta, a análise sintática é uma análise independente, mas, obviamente, a hipótese, que parece ser interessante para dar conta dos dados do ponto de vista sintático, é uma hipótese que pode ser interessante também para essa questão da articulação sintaxe-fonologia, porque o que estamos dizendo é que o que vai mudar entre o português clássico e o português europeu moderno, do ponto de vista da sintaxe, é a posição do sujeito com a ênclise. Então, quando a gente tem sujeito pré-verbal com ênclise, no português clássico, a gente está dizendo que esse sujeito, na realidade, está fora dos limites da oração, mas no português europeu moderno, em que isso é, aliás, a única possibilidade de colocação, o sujeito está dentro dos limites da oração. Essa questão de estar fora e estar dentro é alguma coisa que tem uma ligação muito forte com a questão da entoação, porque esse limite, que eu marquei com um colchete, é o limite da oração, é o limite interpretado fonologicamente.

E isso não é tão novo, há outros casos conhecidos em que alguma coisa parecida acontece. Em francês, por exemplo, aconteceu no sentido inverso. No francês antigo, quando havia um tópico inicial, ele estava dentro da oração, porque era uma língua de tipo germânico, que pode ter um tópico interno, e, a partir de um certo momento, esse tópico foi reinterpretado como um elemento externo à oração. E isso está ligado também a questões de ritmo, porque as línguas que têm um tópico interno inicial, como o alemão, por exemplo, tem um acento inicial. Veja que essa análise pode se sustentar em termos puramente sintáticos mas é um lugar interessantíssimo também para trabalhar com a interface sintaxe/fonologia.

Em relação à mesóclise, o projeto não tematiza particularmente essa questão, porque implicitamente consideramos a mesóclise como um caso de ênclise. Vê-se aliás que a mesóclise é alguma coisa que, no português brasileiro, desaparece junto com a ênclise. Isso dito, parece que a mesóclise também está desaparecendo no português europeu, ou seja, as criancinhas portuguesas já não fazem mesóclise como antigamente. Nós não fizemos isso até agora, mas é possível estudar a evolução da mesóclise no Corpus muito facilmente, já que ela tem uma etiqueta especial, porque a gente já imaginou que alguém podia se interessar por essa construção. Então, alguém que queira fazer um estudo da evolução do uso da mesóclise, nesse Corpus, é só pegar a etiquetinha da mesóclise, que tem um ponto de exclamação, e procurar nos textos, e ver como é que ela acontece.

Outros trabalhos que foram feitos do quais não falei. Cristiane Namiuti, já mencionada, tem um trabalho sobre a interpolação. A interpolação é alguma coisa que aparece pouco nesse período, aliás, aparece muito, mas só com a negação. A Cristiane mostrou, de maneira interessante, que não só ela é restrita à negação, como ela é usada em muito mais contextos. Então, por um lado a interpolação se restringe à negação, mas, por outro lado, ela vai ser usada em qualquer contexto em que é possível ter próclise, enquanto que, no português antigo, era só em casos de próclise obrigatória. Então há outros trabalhos paralelos, que já foram feitos, e que podem ser feitos, porque o Corpus traz a informação necessária.

Espero ter respondido às suas perguntas.

Claudio – Obrigado. Faço também uma pergunta: Na exemplificação dos casos do pronome *o,* o sistema também permite que se faça distinção entre os casos em que esse *o* é pronome pessoal oblíquo e os casos em que ele é um pronome demonstrativo, por exemplo? O exemplo do Vieira, que é assim "*o diz*", e ele diz depois "*o declara*". Haveria outras considerações a fazer, mas pergunto-lhe se o levantamento já

poderia distinguir o caso dos *os* com o valor de *eles*, e do *o* com o valor de *ele* ou de *isto*?

Charlotte – Sim, claro, tudo isso é uma questão de etiquetagem. Esse é um caso em que a gente tem que tomar decisões no campo da anotação, porque na anotação a gente já pode marcar certas distinções, fazendo isso, a gente vai facilitar o trabalho da busca, mas a gente complica o trabalho do etiquetador. Quando uma palavra tem várias funções, e a gente puser uma só etiqueta, isso vai facilitar a etiquetagem e vai dificultar a busca. A gente procurou equilíbrio entre as duas coisas e, em certos casos, a gente tomou a decisão no sentido de multiplicar as etiquetas e, em outros casos, em deixar uma etiqueta só para facilitar a etiquetagem.[7]

Quanto ao *o,* a gente basicamente faz a distinção entre o *o* determinante e *o* pronome clítico. O *o* determinante podia também ser considerado como demonstrativo, quando a gente tem: *o que você fez*, que seria a mesma coisa que: *aquilo que você fez* e, nesse caso, a gente deixou determinante. É pela construção sintática, porque vai ser sempre seguido por uma oração ou por um pronome relativo, que vai ser possível perceber que não se trata do determinante mesmo. Mas isso são escolhas que têm que ser feitas no sistema de anotação e, às vezes, não é óbvio fazer essas escolhas.

Claudio – Obrigado. Temos uma pergunta por escrito: *Você pode explicar um pouco mais o sistema de anotação que você chama de melhores categorias e quais os critérios utilizados para essa classificação?*

Charlotte – A gente tem que chegar a um equilíbrio entre um sistema econômico e um sistema descritivo. Porque se ele não for econômico do ponto de vista computacional vai ser dramático, aliás, para exemplificar isso, fazer um sistema de etiquetagem para o português é muito mais complicado do que fazer para o inglês, por causa da morfologia, porque o inglês é uma língua de pouca morfologia, e o português é uma língua de muita morfologia. Então, nós, no ponto de partida, trabalhamos com o sistema de etiquetagem que foi feito para o inglês médio, no projeto de Anthony Kroch. Eles têm 35 etiquetas, e nós, no final das contas, temos 360. Isso, computacionalmente, faz crescer a complexidade de maneira enorme, quer dizer, para treinar um etiquetador automático que tenha 360 etiquetas, demora meses. Então, o que a gente fez? Marcelo Finger, já mencionado antes, inventou para a gente um sistema em vários passos. Vocês viram que a gente tem etiquetas e sub-etiquetas, e isso resolve o problema em grande parte. Então, é isso. Você tem que ter um sistema econômico, mas você quer ter um sistema que seja descritivamente adequado, então, respondendo à pergunta do Claudio, direi que é esse equilíbrio permanente entre a descrição lingüística e as restrições computacionais, que faz com que você tente achar um sistema ótimo, no sentido da teoria da otimalidade, que é o melhor possível. É isso. Depois, tem alguns problemas de descrição, e é muito interessante fazer esse trabalho, porque a gente vê que há certas coisas pelas quais a gramática tradicional nunca se interessou, e pelas quais a gramática gerativa também não se interessa, e a gente fica meio sem ferramentas, e aí tem que inventar.

Então, o melhor é no sentido do melhor possível, e os critérios são esses. O sistema acaba tendo um pouco a nossa cara, porque, obviamente, nós somos um grupo interessado particularmente na sintaxe, na morfossintaxe, privilegiando coisas como os clíticos, por exemplo. Talvez um outro grupo não fizesse uma distinção entre clíticos e não clíticos, entre clíticos, em geral, e *se*. Nós fizemos certas distinções, porque a gente sabia que elas eram importantes para a gente, e para gente como a gente, então, há um lado subjetivo também.

Claudio – Temos uma pergunta sobre os textos do Corpus: *Vocês fizeram a atualização da ortografia e não alteraram a pontuação. Existe algum projeto de se estudar pontuação, segundo padrões rítmicos no*

português atual?

Charlotte – Esse estudo da pontuação é fascinante. Acontece que, no português atual, a pontuação normativa não é uma pontuação de fôlego, eu diria, não é uma pontuação de leitura em voz alta, é uma pontuação mais lógica. Mas eu acho que há umas interfaces interessantes com a pontuação usada mais espontaneamente. Eu tenho colegas que trabalham com produções de crianças ou produções de gente pouco escolarizada, que não passou justamente por esse processo de normatização. Talvez aqui haja gente trabalhando com isso, para ver como é que as pessoas usam a pontuação. Talvez a tendência mais natural seja usar a pontuação como uma marca de grupos prosódicos, e é o que achamos ainda em textos do século XVIII. Ainda está para ser feita uma história da pontuação no português. Existem trabalhos com a pontuação antiga, mas aí é que está, a história do português foi muito trabalhada até o século XVI e depois achou-se que já era a língua moderna, por isso houve muito menos trabalhos. Mas isso é falso. A língua portuguesa ainda passou por sérias modificações depois do século XVI.

A pontuação dos séculos XVII e XVIII é bem diferente da nossa. Ela é mais retórica, mais ligada à prosódia, e seria muito interessante estudá-la e comparar eventualmente com o que fazem pessoas pouco escolarizadas na língua moderna, porque a escolarização moderna leva as pessoas a usarem a pontuação de maneira semântica, sintática e lógica eu diria. Então, nós vamos, em algum momento, trabalhar essa pontuação. Até agora, a gente tinha pouquíssimos textos com a pontuação original, porque os editores mudam a pontuação, isso é uma coisa assim, não tem jeito, você vê, lê a primeira página, compara com a edição original e vê que a primeira vírgula já mudou de lugar, é uma coisa que os editores não deixam de fazer, mesmo quando eles mexem pouco na ortografia.

Claudio – Outra pergunta da platéia: *Na sua pesquisa existe alguma etiquetagem para o estudo do contexto transfrástico, por exemplo? Vocês estão pensando alguma coisa assim? Você falou agora da pontuação. Há alguma etiqueta para isso?*

Charlotte – Sim, a pontuação vem etiquetada. Mas não há etiquetas remetendo a funções transfrásticas. O nosso sistema de etiquetagem é um sistema que se dá no nível da frase, mas eu penso mesmo, e isso me aconteceu com os clíticos, que muitas vezes a gente vai ter interesse em olhar para o texto. A vantagem de ter o Corpus à disposição é que você pode ir para ele. Nós não damos as ferramentas imediatamente, porque não fazemos uma análise textual, a própria anotação sintática é uma anotação muito básica para recuperar informações, mas, depois, eu acho que a gente, até para análise sintática tem que voltar ao texto. É por isso que disponibilizamos o texto inteiro, porque o pesquisador não pode prescindir do texto. Agora, eu acho que esse Corpus, apesar de ser pensado para análise sintática, ele já pode ser um Corpus interessante para análise textual e para a interface entre a sintaxe e o texto, a sintaxe e a prosódia.

Claudio – Outra pergunta da platéia: *Ao explicar a presença da ênclise e sujeito, você usou o conceito de tópico, que é de discurso. Como é que você concilia os aspectos gerativistas e discursivos?*

Charlotte – Eu acho que a gramática é um órgão biológico, mas ela é, obviamente, usada para funções discursivas. Então, eu sempre achei que uma abordagem discursiva e a abordagem sintática, mesmo gerativista, não eram incompatíveis, mas complementares. E para mim esse trabalho com Vieira foi fascinante, porque eu vi que a ênclise nos sermões tem a ver com o estilo dele. E isso passa por uma análise sintática, em que a noção de tópico é importante, mas a noção de tópico é uma noção de interface sintaxe/discurso, porque o tópico é também uma noção discursiva. Acontece que cada língua trata sintaticamente o tópico de maneira diferente.

Então, tipicamente, existem línguas nas quais o tópico é sempre um elemento externo à oração, e línguas nas quais o tópico pode ocupar uma posição interna à oração. Línguas como o português clássico, o português antigo possivelmente também, as línguas de acento inicial, são línguas que podem ter um tópico interno à oração. Então, o tópico, do ponto de vista discursivo, vai ser o mesmo, mas cada sintaxe vai tratar esse tópico de maneira diferente, e pode ser um gancho justamente na mudança das línguas uma vez que a reanálise da posição do tópico é um lugar de mudança sintática.

Claudio – Eu vou pedir a compreensão de todos que enviaram perguntas e solicitar que venham conversar com a professora Charlotte reservadamente, porque nosso tempo está mais do que esgotado. Quero novamente me congratular com a professora pela belíssima exposição e agradecer pela sua presença em nosso Seminário.

Charlotte – Obrigada a todos. Foi um prazer estar com vocês.

@ @ @ @ @ @ @

N. do Org.: Versão escrita pela autora a partir de transcrição feita pela monitora Márcia de Oliveira Gomes, do Instituto de Letras da UERJ.

---

[1] Uma discussão dessa proposta se encontra em Britto, H. e C. Galves "A construção do Corpus anotado do português histórico Tycho Brahe: o sistema de anotação morfológica". *IV Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 99),* Évora, 20-21 de setembro de 1999.

[3] Cf. http://www.ime.usp.br/~tycho/corpus/database

[4] Uma primeira apresentação dessa variação encontra-se em Galves, C., H. Britto e M.C Paixão de Sousa "First Results from the Tycho Brahe Corpus" disponível em http://www.ime.usp.br/~tycho/what/index.html

[5] Esta análise está desenvolvida no meu artigo "Sintaxe e estilo: a colocação de clíticos nos Sermões do Padre Vieira" a sair no volume comemorativo dos 25 anos do IEL-UNICAMP (cf. a versão em inglês em http://www.ime.usp.br/~tycho/what/index.html)

[6] Existe atualmente uma polêmica sobre a posição do sujeito no português europeu moderno. Cf., entre muitos outros, o artigo de João Costa e Charlotte Galves " External subjects in two varieties of Portuguese: evidence for a non-unified analysis",publicado em *Portuguese Syntax*, João Costa (org.), Oxford 2000.

[7] A esse respeito ver Britto e Galves (1999) citado em nota acima.
