BERBER SARDINHA, Tony. Lingüística de Corpus: uma entrevista com Tony Berber Sardinha. *Revista Virtual de Estudos da Linguagem - ReVEL*. Vol. 2, n. 3, agosto de 2004. ISSN 1678-8931 [www.revel.inf.br].

# LINGÜÍSTICA DE CORPUS – UMA ENTREVISTA COM TONY BERBER SARDINHA

Tony Berber Sardinha
Pontifícia Universidade Católica de São Paulo – PUCSP

**ReVEL – Com o que se preocupa atualmente a Lingüística de Corpus?**

**Tony** – Antes de mais nada, quero agradecer a oportunidade de dar essa entrevista, e te congratular pela ReVEL. A Lingüística de Corpus se ocupa de quase todas as áreas de investigação lingüística. O léxico é a que mais recebe a atenção dos lingüistas de corpus e é a que mais se projeta para o mundo, basta ver os dicionários de inglês atuais, que são produzidos com base em corpus. Além dos léxico, o estudo da gramática começa a se tornar baseado em corpus. A gramática da editora Longman de 1999, sob direção de Douglas Biber, é uma delas. Outras são as do COBUILD, tanto a 'geral' do inglês quanto as específicas (verbos, adjetivos e substantivos), de autoria de Susan Huston e Gill Francis. Ainda podemos citar a sintaxe, a morfologia e a fonologia como áreas que possuem extensa participação de pesquisas baseadas em corpora. Afora essas áreas 'tradicionais' da lingüística, há outras mais novas, como os Tradução, de um lado, e Metáfora, de outro. Esses dois são mais recentes. O ensino de línguas estrangeiras como corpus também é relativamente recente, embora já venha há mais tempo do que tradução ou metáfora.

**ReVEL – Quais foram os primeiros estudos baseados em corpora lingüísticos de que se tem conhecimento? E quais foram os primeiros estudos baseados em corpora eletrônicos?**

**Tony** – Talvez sejam os relativos à Bíblia. Aqui me refiro às compilações de citações dos Livros Sagrados, compilados por monges, provavelmente na Idade Média. Não sei se podemos chamá-los de estudos, tal como entendemos essa palavra hoje, mas são certamente concordâncias extraídas de um grande corpus. Essa compilação de citações ainda hoje é comum, inclusive como auxílio a pastores e pregadores, que precisam encontrar rapidamente as partes das Escrituras que desejam. É interessante tentar entender por que se fazia e se faz esse tipo de trabalho. A razão é bem simples – não se pode 'inventar' ou adaptar a palavra de Deus – ela devia ser transcrita tal qual aparecia no texto original. Não se cogitava alguém ter 'intuição' da palavra de Deus. Talvez os que a tivessem eram tidos como profetas ou santos, mas não um ser humano qualquer. Já com a palavra do homem, a situação foi bem diferente, como nos mostra a história dos estudos lingüísticos.

**ReVEL – Que tipos de aplicativos podem resultar a partir de pesquisas baseadas em estudos de corpora eletrônicos?**

**Tony** – Há muitos. No âmbito dos grupos de pesquisa, praticamente não há limite. Aliás, nem podemos conhecer todos que existem, porque há incontáveis grupos de pesquisa que utilizam corpora, e os programas que esses grupos criam são particulares e de acesso restrito. Na esfera do que poderíamos chamar de 'consumidor final', ou seja, aqueles programas a que podemos ter acesso, podemos destacar alguns, como corretores ortográficos, resumidores (por exemplo, a função 'Auto Resumo' do Microsoft Word), sintetizadores de voz, tradutores e digitadores (Via Voice, por exemplo). Todos esses programas hoje estão disponíveis para o usuário final de um sistema operacional como Windows

ou Mac. Isso tudo sem falar nos programas para manuseamento de corpora, como concordanciadores, extratores de freqüência e etiquetadores.

**ReVEL – Como o senhor avalia a Lingüística de Corpus no Brasil? Já possuímos grandes recursos para trabalhar com corpora em língua portuguesa?**

**Tony** – Ela está se desenvolvendo bastante rápido nos últimos anos. Já contamos com muitos recursos para a pesquisa, a começar com corpora eletrônicos disponíveis à comunidade em geral. O Banco de Português tem parte de seu acervo na Web. O Lácio Web já se encontra na Web e tende a crescer. O Tycho-Brahe, de português histórico, também está na Web há muitos anos. Fora do Brasil, a Linguateca já disponibiliza vários corpora em português, inclusive o do NILC, de português brasileiro, há um certo tempo. Temos software para análise de corpus em português, como etiquetadores. Temos também literatura sobre corpora em português, artigos, dissertações, um livro, muitas apresentações nos mais variados encontros científicos relacionados à linguagem, como os Encontros de Corpora (www.nilc.icmc.usp.br/iiiencontro), o GEL (www.gel.org.br), o InPLA (lael.pucsp.br/inpla), o CIATI (www.unibero.br), o CBLA (lael.pucsp.br/cbla), entre outros. Há também vários grupos de pesquisa cadastrados no CNPq que utilizam corpora. Contudo, a literatura sobre corpora é quase toda em inglês; embora muitos saibam ler nessa língua, ainda assim acho necessário termos um diálogo em português sobre corpus. Só assim podemos nos apropriar de fato da Lingüística de Corpus e dizer que temos uma Lingüística de Corpus brasileira. Não quero dizer, claro, que não devamos produzir em inglês, ir a congressos fora do Brasil, etc. Muito pelo contrário. O que eu quero enfatizar é que a literatura e os discursos compartilhados em português são o alicerce para criarmos uma disciplina e uma comunidade no país. Mesmo com todo esse desenvolvimento, ainda precisamos de várias coisas, que virão a seu tempo. Por exemplo, acho que estamos caminhando para termos, a certa altura, um Corpus Nacional de Português Brasileiro, talvez nos moldes do British National Corpus.

Esse é um projeto de grande envergadura, que precisaria de muito investimento e de parceiros comerciais, além do financiamento público. Outro elemento que precisamos é de ferramentas simples para o usuário iniciante. Isso eu digo por experiência própria, lecionando e orientando dissertações. Não podemos nos esquecer que boa parte de nosso público é de alunos de cursos de Letras, de Tradução, por exemplo, que possuem conhecimento básico de informática. Programas que exijam conhecimento de programação, com instruções via linhas de comando, por exemplo, são inviáveis para esses alunos. Não quero dizer que não devamos ter programas assim, claro que não, até porque alunos 'computeiros' não se intimidam com linhas de comando e coisas assim. Mas não podemos nos esquecer de nossos alunos sem grande conhecimento de informática e de como podemos fornecer meios para incluí-los nas pesquisas com corpora.

**ReVEL – Como especialista na área, que livros o senhor poderia indicar para aqueles que estão começando seus trabalhos com corpora lingüísticos?**

**Tony** – Já que você levantou a bola... ;-) não poderia deixar de mencionar o meu 'Lingüística de Corpus', que saiu este ano (2004) pela editora Manole. Outro livro, 'A Língua Portuguesa no Computador', é uma coletânea organizada por mim sobre Lingüística de Corpus, PLN e áreas afins, que vai sair este ano também, pela editora Mercado de Letras, co-edição com a FAPESP.

Em inglês, temos várias ótimas introduções à Lingüística de Corpus (McEnery e Wilson, Biber et al., Kennedy, Hunston). Para quem quer ter uma visão ampla e histórica da Lingüística de Corpus, recomendo a antologia 'Corpus Linguistics: readings in a widening discipline', organizado por Geoffrey Sampson e Diana McCarthy. Para o futuro, deixo a recomendação de 'Corpus Linguistics – Critical Readings', a ser organizado por Wolfgang Teubert. Para o pessoal mais voltado à PLN, creio que 'Foundations of Statistical Natural Language Processing', de

Manning e Schütze seja leitura obrigatória, além de possivelmente 'Probabilistic Linguistics', de Bod , Hay e Jannedy.

Mas além dos livros (que são caros!), lembro que muitas revistas publicam artigos sobre corpora e podem ser acessadas pela Internet sem precisar pagar pela aquisição dos artigos. O Portal de Periódicos da CAPES é um recurso extraordinário (www.periodicos.capes.gov.br) e traz muitas revistas com trabalhos sobre corpus. Dá um certo trabalho 'pescar' os artigos, porque o portal não permite busca direta, de entrada, por título ou assunto *do artigo,* mas apenas do periódico. Mas uma vez encontrado o periódico ou editora, fica fácil baixar muitos artigos preciosos sobre corpora. Lembro que o Portal CAPES só permite baixar artigos se for acessado de dentro de uma universidade conveniada. Acessar diretamente de casa não funciona – você apenas vê o título e o resumo, mas não o artigo inteiro. A SciELO (www.scielo.br), outro recurso público financiado pela FAPESP, dispõe a revista DELTA online, também de graça, onde é possível encontrar vários artigos sobre Lingüística de Corpus. Para o pessoal mais computacional, indico o site da ACL (Association for Computational Linguistics) que traz o ACL Anthology, com artigos sobre Lingüística Computacional, também de graça, em http://acl.ldc.upenn.edu; é um verdadeiro tesouro de artigos atuais e antigos sobre computação, muitos dos quais sobre corpora. Outro *site* na mesma linha é o http://xxx.lanl.gov/cmp-lg/, com milhares de artigos.