# CAMINHOS DA LINGUÍSTICA DE CORPUS

*Série Espaços da Linguística de Corpus*
Editor: Tony Berber Sardinha – Pontifícia Universidade Católica de São Paulo – Brasil

TANIA M. G. SHEPHERD
TONY BERBER SARDINHA
MARCIA VEIRANO PINTO
(ORGANIZADORES)

# CAMINHOS DA LINGUÍSTICA DE CORPUS

MERCADO® LETRAS

**Dados Internacionais de Catalogação na Publicação (CIP)**
**(Câmara Brasileira do Livro, SP, Brasil)**

Caminhos da linguística de corpus / Tania M. G. Shepherd, Tony Berber Sardinha, Marcia Veirano Pinto (organizadores) . – Campinas, SP : Mercado de Letras, 2012. – (*Série Espaços da Linguística de Corpus*)

ISBN 978-85-7591-158-7

1. Análise linguística 2. Linguagem e línguas - Ensino auxiliado por computador 3. Linguagem e línguas - Estudo e ensino 4. Linguística - Metodologia 5. Linguística - Processamento de dados I. Shepherd, Tania M. G.. II. Sardinha, Tony Berber. III. Pinto, Marcia Veirano. IV. Série.

12-00700 CDD-410.285

**Índices para catálogo sistemático:**
1. Linguística de corpus : Análise lingüística via computador : Linguística aplicada 410.285

*capa e gerência editorial*: Vande Rotta Gomide
*revisão*: Editora Mercado de Letras

*Série Espaços da Linguística de Corpus*
coordenação: Tony Berber Sardinha


Rua João da Cruz e Souza, 53
Telefax: (19) 3241-7514
13070-116 – Campinas SP Brasil
www.mercado-de-letras.com.br
livros@mercado-de-letras.com.br

1ª EDIÇÃO
*texto adaptado para as novas normas de ortografia da língua portuguesa*

**2012**

IMPRESSÃO DIGITAL

*Ao Richard*
in memoriam

*À Tania*

*Ao Walter*

## AGRADECIMENTOS

Os organizadores gostariam de agradecer a todos aqueles que contribuíram para o presente volume.Sem os esforços dos monitores, que atuaram no Encontro de Linguística de Corpus em 2009 no Rio de Janeiro, sem os trabalhos dos autores e dos pareceristas anônimos que os selecionaram para apresentação tanto na fase do Encontro como agora neste livro e sem a ajuda incansável do Grupo de Estudos em Linguística de Corpus, nada teria sido possível.

Agradecemos também à FAPERJ/UERJ, CAPES e CNPq pelas bolsas de pesquisa recebidas e pelo auxílio para a realização do evento que deu origem a este livro.

Pela autorização da tradução para língua portuguesa dos três artigos seminais incluídos nesta coletânea, agradecemos aos seguintes editores: a John Benjamins Publishing Company pelo artigo "Starting with the small words: patterns, lexis and semantic sequences", de Susan Hunston, publicado no *International Journal of Corpus Linguistics* 13: 3 (2008, pp. 271-295); a Mouton de Gruyter pelo artigo "A grammar of linguistic metaphors", de Alice Deignan publicado em A. Stefanowitsch e S. T. Gries (eds.) *Corpus-based Approaches to Metaphor and Metonymy,* (2006, pp. 106-122) e finalmente à Association for Computational Linguistics, pelo artigo "The Human Language Project: Building a universal corpus of the World's languages", de Steven Abney e Steven Bird, publicado originariamente nos *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (2010, pp. 88-97), Uppsala, Sweden.

Finalmente, agradecemos à Maria Elisa, que acreditou no presente projeto e vislumbrou um futuro para a Linguística de Corpus através da série Espaços.

*Tania, Tony e Márcia*

## SUMÁRIO

CONSTRUÇÃO E CODIFICAÇÃO DE CORPUS



QUESTÕES DE LINGUAGEM E LINGUÍSTICA APLICADA

# PREFÁCIO

*Ana Frankenberg-Garcia*

A ideia de coligir coleções de textos naturais com o objetivo de os submeter à análise linguística remonta ao trabalho dos estruturalistas norte-americanos da década de 1950, tais como Harris (1951) e Fries (1952). Com o Brown *Corpus* (Francis e Kucera 1964), surgiria o primeiro *corpus* eletrônico compilado para este fim. Embora até hoje este *corpus* seja largamente utilizado, na altura praticamente não existiam textos escritos em formato digital, os computadores eram máquinas enormes e caras, que ocupavam salas inteiras, e os programas informáticos demoravam horas ou até dias a correr. Além disso, ofuscada pelo racionalismo de Chomsky, a abordagem essencialmente empírica do estudo das línguas abraçada por pesquisadores que então começaram a trabalhar com *corpora* permaneceria ainda por vários anos nos bastidores. Foi apenas com a proliferação dos computadores pessoais, de textos em formato digital e de ferramentas acessíveis de análise de *corpora*, tais como o WordSmith Tools (Scott 1996), que a Linguística de Corpus pôde finalmente, a partir dos anos noventa, começar a se desenvolver de fato.

No Brasil, o primeiro Encontro de Linguística de *Corpus* (ELC) teve lugar em 1999. Dele não participaram mais do que um grupo reduzido de

pesquisadores, mas estava lançada a semente. Com o objetivo de "abrir um espaço de discussão para as questões relativas à elaboração e manutenção de *corpora*, ao intercâmbio de recursos e ideias referentes à pesquisa baseada em *corpus* e à formação de parcerias entre pesquisadores e instituições" (Sardinha 2008, p. 19), estes encontros, inicialmente bienais, passaram a ser anuais e a contar com cada vez mais participantes.

Este volume é produto da oitava edição do ELC, organizado pela Universidade Estadual do Rio de Janeiro em novembro de 2009. Infelizmente, não pude estar presente. De qualquer forma, é uma grande honra para mim poder escrever este prefácio, pois os dezoito trabalhos escolhidos e reunidos nesta coletânea são uma amostra tanto das oportunidades que a Linguística de *Corpus* oferece aos pesquisadores, como daquilo que de melhor vem sendo feito no Brasil neste domínio. Em comum, temos a observação empírica de fenômenos da linguagem natural a partir de conjuntos de textos digitais representativos de uma língua ou sublíngua. A diversidade de enfoques que se pode privilegiar a partir daí é incomensurável. Vemos aqui novos *corpora*, novas abordagens de codificação, ferramentas de análise inovadoras, discussões sobre conceitos básicos e pesquisas específicas envolvendo metáforas, expressões fixas, textos históricos, linguagens especializadas, linguagem de aprendizes, linguagem oral, tradução, lexicografia, terminologia, análise do discurso e ensino de línguas. A multiplicidade de temas patentes neste volume não é uma coincidência, mas sim um sinal de que a Linguística de *Corpus* é um campo fértil e em franca expansão para a pesquisa.

Conforme também se reflete nos capítulos presentes neste livro, a Linguística de *Corpus* apresenta-se, simultaneamente, como uma nova metodologia (que utiliza textos naturais e ferramentas informáticas para descrever a língua) e uma nova disciplina (no sentido de uma nova abordagem à descrição linguística). Por um lado, os métodos básicos utilizados - a visualização de palavras-chave-em-contexto, a ordenação das palavras em termos da sua frequência e o cálculo do grau de proximidade entre palavras através de estatísticas de coocorrência - coadunam-se com qualquer campo de investigação baseado na análise textual, incluindo, entre outros, o ensino-aprendizagem de línguas, a lexicografia, a análise do discurso histórico, político

e jornalístico, os estudos literários, os estudos de tradução, a sociolinguística e o desenvolvimento de novas ferramentas de processamento da linguagem natural, tal como sistemas de tradução automática e de detecção de plágio. Por outro lado, esses métodos abriram as portas a uma leitura vertical do texto e a uma consequente visão de padrões de uso da língua sem precedentes, chegando a pôr em causa certos pressupostos linguísticos nunca antes contestados. Segundo Tognini Bonelli (2010, pp. 17-18)

What started as a methodological enhancement but included a quantitative explosion (I am referring to the quantity of data processed thanks to the aid of the computer) has turned out to be a theoretical and qualitative revolution in that it has offered insights into language that have shaken the underlying assumptions behind many well-established theoretical positions in the field […] It is strange to imagine that just more data and better counting could trigger philosophical repositionings, but that is indeed what has happened.

Ao lermos o conjunto de artigos apresentados nestes *Caminhos da Linguística de Corpus,* temos precisamente a oportunidade de acompanhar de perto esta tendência no Brasil, o que é uma evidência feliz de que a semente lançada no primeiro ELC, há mais de uma década, germinou e frutificou.

*Referências*


FRANCIS, W. e KUCERA, H. (1964). *Manual of Information to Accompany a Standard Corpus of Present-Day Edited American English, for Use with Digital Computers*. Providence, RI: Brown University, Department of Linguistics (revisto em 1971; revisto e ampliado em 1979). Disponível em: http://icame.uib.no/brown/bcm.html.

FRIES, C. (1952). *The Structure of English: An Introduction to the Construction of Sentences*. Nova York: Harcourt-Brace.

HARRIS, Z. (1951). *Methods in Structural Linguistics*. Chicago: University of Chicago Press.

SARDINHA, T. (2008). "A Linguística de Corpus no Brasil", *in*: TAGNIN, S. e VALE, O. (eds.) *Avanços da Linguística de Corpus no Brasil*. São Paulo: Humanitas.

SCOTT, M. (1996). *WordSmith Tools.* Oxford: Oxford University Press.

TOGNINI BONELLI, E. (2010). "Theoretical overview of the evolution of corpus linguistics", *in*: O´KEEFFE, A. e McCARTHY, M. (eds.) *The Routledge Handbook of Corpus Linguistics.* Londres e Nova York: Routledge.