RESENHA/REVIEW

SAMPSON, Geoffrey e Diana MCCARTHY (eds.) 2004. *Corpus Linguistics – Readings in a Widening Discipline*. Londres e Nova York: Continuum, 524 págs.*

Resenhado por Tony BERBER SARDINHA
*(Pontifícia Universidade Católica de São Paulo)***

KEY-WORDS: *Corpus Linguistics; Corpora; History of Linguistics.*
PALAVRAS-CHAVE: *Lingüística de Corpus; Corpora; História da Lingüística.*

1. Introdução

Corpus Linguistics – Readings in a Widening Discipline, compilada por Geoffrey Sampson e Diana McCarthy, é uma antologia de trabalhos que empregaram corpora, eletrônicos ou não, na pesquisa lingüística ao longo de mais de 50 anos de história. A obra impressiona por vários motivos. Primeiramente, pelo número de trabalhos que reúne: 42. Em segundo lugar, pelo período de tempo que cobre: 50 anos, entre 1952 e 2002. E em terceiro lugar, ela impressiona pelo seu tamanho: 524 páginas.

Geoffrey Sampson é professor da universidade de Sussex, departamento de Informática. Defensor da pesquisa lingüística empírica, é autor de muitos livros, com ênfase no ataque ao gerativismo, como em *Empirical Linguistics* e *The Language Instinct Debate*. Diana McCarthy é uma jovem professora do mesmo departamento e se dedica à pesquisa em Processamento de Linguagem Natural, especialmente sobre aquisição lexical (*lexical acquisition*) e desambiguação de sentido de palavra (*word-sense desambiguation*).

* Agradeço aos pareceristas pelos valiosos comentários.
** Agradeço ao apoio do CNPq, mediante a bolsa de número 350455/03-1.

Devido às dimensões da obra, esta resenha se limitará a enfocar em detalhe apenas alguns dos capítulos da obra. Os capítulos escolhidos para essa resenha detalhada serão aqueles que parecem ser os de maior destaque histórico no desenvolvimento da Lingüística de Corpus e também aqueles que ainda possuem influência contemporânea, por terem começado uma tendência ou enfrentado uma questão atual. Devido a esse recorte, a seleção foi inevitavelmente enviesada, direcionada àquilo que considero ser algumas das questões criticas da Lingüística de Corpus. Entretanto, não reivindico que minha seleção seja a definitiva.

2. Conteúdo

O primeiro capítulo é a introdução à obra, escrita pelos organizadores (Sampson e McCarthy). Ela é escrita no tom direto que é marca registrada de Sampson. Logo cedo, eles anunciam uma visão de mundo que aparece claramente na seleção dos capítulos que compuseram a obra; segundos os organizadores, haveria dois tipos de lingüista de corpus: os da área de Ciências Humanas ('humanistic scholars') e os de Engenharia de Linguagem ('language engineers'). Os primeiros, voltados aos que eles chamam de 'pesquisa lingüística com base nas Ciências Humanas' ('humanities-based linguistic scholarship') (p. 2) parecem ser os pesquisadores voltados à descrição e ao ensino de línguas, normalmente ligados a departamentos de línguas e de lingüística aplicada. Os segundos, por sua vez, praticariam 'engenharia lingüística economicamente útil' ('economically useful language engineering') (p. 2) e se preocupariam em desenvolver programas de computador para atender finalidades práticas, muitas vezes comerciais. Essa visão ressoa na montagem da obra, pois há muitos capítulos (possivelmente 15) que se encaixariam em engenharia de linguagem. Há ainda na introdução ótimas passagens sobre o estatuto e o desenvolvimento da Lingüística de Corpus, tais como: "os 'átomos' de uma língua são as palavras" ("the 'atoms' of a language are its words") (p.2); "cinqüenta anos atrás, Lingüística de Corpus era uma atividade minoritária, especializada e obscura" ("fifty years ago, corpus linguistics was an osbcure and highly specialized minority activity") (p.5); "agora é uma boa época para ser um lingüista de corpus" ("this is a good time to be a corpus linguist") (p.5).

O segundo capítulo é uma reprodução de um trecho do livro *The Structure of English*, de Charles Fries, mais especificamente a introdução e o

capítulo 3. *The Structure of English* foi lançado em 1952 (Fries, 1952) e a sua inclusão no livro está relacionada ao fato de que Fries compilou um corpus de cerca de 250 mil palavras, de conversas telefônicas, gravadas em disco, num equipamento rudimentar para os dias de hoje. Se hoje em dia a captura e transcrição de dados orais ainda é o tendão de Aquiles dos lingüistas de corpus, imagine o que foi isso há mais de 50 anos! Mesmo depois da coleta, o trabalho de análise foi monumental, visto que Fries não teve computadores disponíveis para si, já que os poucos computadores que existiam eram de difícil acesso. A leitura desse trecho de Fries revela, ao leitor de hoje, uma preocupação com os dados empíricos, algo que viria a cair de moda mais tarde naquela década, com o lançamento da obra de (Chomsky, 1957) e o gerativismo. O autor apresenta dados de freqüência dos vários enunciados ('utterance units') realizados nas conversas telefônicas, análise dos padrões de entoação, bem como a organização dos enunciados no discurso. Em suma, o primeiro capítulo do livro é um resgate histórico de um dos pioneiros dos estudos de corpus.

O terceiro capítulo traz mais uma contribuição de valor histórico, qual seja, o artigo de Nelson Francis intitulado 'A standard corpus of edited present-day American English', onde ele relata a compilação do famoso 'Brown Corpus'. Esse corpus é tido como o pioneiro em termos de corpora eletrônicos. O corpus é uma compilação de um milhão de palavras de inglês americano, retiradas de jornais, revistas, manuais, enfim, publicações variadas que saíram nos Estados Unidos em 1961. O artigo de Francis, que saiu em 1965 (Francis, 1965), interessa ao leitor contemporâneo por vários motivos, mas talvez principalmente com os detalhes relativos ao uso da tecnologia disponível no início dos anos de 1960. Por exemplo, na p. 31 o autor discute o problema de desenvolvimento de um padrão para processamento de dados textuais via computador; segundo ele, os compiladores do corpus tentaram usar o padrão usado pelo governo americano, mas não conseguiram, porque a reprodução do material impresso requeria funcionalidades que o sistema existente não fornecia (ele havia sido desenvolvido em 1959 para registrar patentes!). O autor revela ainda que o sistema de registro digital seria então fornecido aos usuários junto com as fitas do corpus; esse é outro detalhe que salta aos olhos do leitor contemporâneo: o sistema de armazenamento original do corpus era em fita de computador. Esses aspectos podem parecer apenas pequenos detalhes, mas são na verdade lembretes importantes de que a Lingüística de Corpus está intima-

mente ligada à tecnologia. Os corpora que tivemos no passado e os corpora que temos hoje têm essa feição devido à tecnologia existente na época.

O quarto capítulo é uma reprodução de artigo publicado na revista *Lingua* por F. Aarts em 1971 (Aarts, 1971). Nele, o autor mostra resultados de uma pesquisa em um corpus de 75 mil palavras retiradas do Survey of English Usage (SEU); essa amostra não havia ainda sido digitalizada (ela o seria décadas depois, quando se tornou o London-Lund Corpus). O SEU é um corpus pioneiro de inglês britânico fundado em Londres por iniciativa de Randolph Quirk. A pesquisa focou a composição de sintagmas nominais. Ele descobriu que os grupos nominais (NP) possuíam composição diferente de acordo com a posição deles na estrutura da frase e com o tipo de texto. Dessa forma, ele colocou em questão a aplicabilidade de regras sintáticas do tipo $S \Downarrow \ NP\ VP$ e $VP \Downarrow \ V\ NP$, pois segundo seus resultados, a estrutura do NP nessas duas regras era diferente; além disso, a estrutura do NP também era diferente em ambas as regras de acordo com o tipo de texto em que ocorriam. Esse capítulo é um dos primeiros exemplos de pesquisa em que os achados baseados em corpus confrontam diretamente posições teóricas da lingüística. Ele ainda mostra como um pequeno corpus pode ser grande o bastante para responder uma pergunta específica relativa à estrutura da língua.

Passando adiante para a década de 1980, a contribuição mais significativa, a meu ver, parece ser a de Sinclair, intitulada 'Corpus Creation' (capítulo 8), originalmente publicada em 1987 e depois reproduzida no seu livro *Corpus Concordance Collocation* de 1991 (Sinclair, 1991), livro esse que é um dos grandes marcos da área. John Sinclair é um dos criadores da Lingüística de Corpus contemporânea, tendo sido diretor do projeto Cobuild, uma pareceria da Universidade de Birmingham e da editora Collins; esse projeto nos deu os dicionários Cobuild (Sinclair & et al., 1987, 1995), que revolucionaram a lexicografia mundial, fazendo com que se tornasse impensável criar um dicionário sem apoio total em dados de corpora eletrônicos. Além dos dicionários, Sinclair chefiou a gramática do Cobuild (Sinclair, Fox, & et al., 1990; Sinclair & et al., 1996) e uma série de livros didáticos de inglês. Mas foi na pesquisa com corpora que a marca de John Sinclair se fez mais sentir no mundo acadêmico; ele criou uma escola de Lingüística de Corpus que é talvez a mais praticada e a mais dinâmica do mundo. Nesse trabalho, ele apresenta vários critérios de criação de corpus

que permanecem válidos até hoje (até porque são em grande parte gerais). Já se pode perceber nesses critérios a marca da 'Escola de Birmingham' de Lingüística de Corpus, principalmente talvez na diretriz de número 6, intitulada 'Clean text policy'. Nela, Sinclair advoca que os textos (arquivos) que compõem o corpus devem ser mantidos livres de anotação (morfológica, sintática, etc); em termos simples, a versão principal do corpus deve ser de textos 'limpos'. Um texto limpo é mais útil para um grupo maior de pesquisadores em projetos diferentes. Outra diretriz importante é a de número 7 ('Basic provisions'), que preconiza, entre outras coisas, que os programas de gerenciamento do corpus devam ser disponibilizados online. Atualmente, com o maior acesso à Internet, essa diretriz se torna cada vez mais uma realidade nos projetos de corpora.

Em relação aos anos de 1990, três contribuições chamam a atenção.

A primeira é um artigo de Doug Biber de 1993 (Biber, 1993), intitulado 'Representativeness in Corpus Design' (capítulo 17 da obra). Doug Biber é hoje um dos mais respeitados lingüistas de corpus do mundo, sendo o nome de maior expressão na América do Norte; é um dos pesquisadores mais prolíficos da área, com dezenas de artigos e livros importantes da área. O artigo reproduzido na coletânea é um texto denso, repleto de cálculos matemáticos, que deve ser de difícil leitura para a maioria dos praticantes da Lingüística de Corpus. Nele, o autor apresenta maneiras estatísticas de calcular a quantidade mínima de textos que um corpus deve ter. Essa tem sido uma das questões mais difíceis que o lingüista de corpus enfrenta, já que ele deseja que seus achados sejam possíveis de serem generalizados ao máximo, e para isso é preciso que sejam baseados em uma quantidade representativa de dados. Todo corpus é uma amostra que provém de uma população. A população, no caso, é a língua como um todo, mas como essa população não possui um tamanho ou estrutura definidos, é muito difícil afirmar se essa população está sendo representada pela amostra. Biber apresenta no artigo um método estatístico para resolver esse impasse, que se baseia no cálculo da distribuição (média, desvio padrão e erro amostral) de categorias sintáticas. A esses dados, o autor aplica uma fórmula que determina o tamanho mínimo que uma amostra deva ter para representar a população de uma dada categoria lingüística (p.ex. substantivos).

O segundo artigo de destaque dos anos de 1990 é 'A corpus-driven approach to Grammar – Principles, Methods, and Examples') de uma seguidora de Sinclair e pesquisadora sênior do projeto Cobuild, Gill Francis (Francis, 1993). O artigo é um capítulo de uma *Festschrift* para John Sinclair, lançada por seus seguidores (Baker, Francis, & Tognini-Bonelli, 1993). Nesse capítulo, Francis detalha os princípios que guiariam a criação das gramáticas de padrão do projeto Cobuild, a *Collins Cobuild Grammar Patterns 1: Verbs* (Francis & Hunston, 1996) e a *Collins Cobuild Grammar Patterns 2: Nouns and Adjectives* (Francis & Hunston, 1998), ambas publicadas por Francis e sua colega Susan Hunston, e depois descritas teoricamente em (Hunston & Francis, 2000). Os princípios sobrevivem até hoje e fizeram escola ao redor do mundo, influenciando muitos pesquisadores com suas idéias de partir da palavra e chegar aos padrões gramaticais, em vez de usar as palavras para somente ilustrar os padrões gramaticais identificados independentemente do léxico. Eles consistem essencialmente na identificação das colocações e coligações das várias palavras da língua e na análise de sentidos em comum entre essas colocações ('semantic sets'). A partir desses conjuntos de sentidos em comum, a pesquisadora levanta os padrões gramaticais. Por exemplo, um desses conjuntos é 'general nouns', como 'reason', 'basis' e 'danger'. Todos esses compartilham a coligação 'substantivo + that'. 'Reason' aparece em padrões do tipo 'The reason that' e 'for the simple reason that'. A categoria 'general noun' é uma que entraria na gramática de substantivos.

A terceira contribuição de destaque dos anos 1990 é 'Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies', capítulo 20 da obra, de autoria de Bill Louw, publicada anteriormente (Louw, 1993) na mesma *Festschrift* para John Sinclair em que o capítulo de Francis aparecera. Esse trabalho ficou famoso por cunhar um nome para um tipo de padrão que tem se mostrado cada vez mais útil na análise de corpus: a prosódia semântica. Prosódia semântica é um tipo de padrão que indica uma conotação favorável ou desfavorável. O nome prosódia é na verdade um empréstimo da fonologia e indica um sentido que impregna uma série de palavras que normalmente ocorrem juntas. Louw inicia seu artigo dando crédito a John Sinclair por ter descoberto esse tipo de padronização, em suas análises de 'happen' e 'set in'; porém, é Louw que merece o crédito de ter cunhado o termo que se firmou na área. Louw mostra na sua análise o caso, agora clássico, de 'utterly', que ao se juntar com adjeti-

vos do tipo 'grand' e 'vulnerable' resulta em ironia. Outro caso clássico apresentado pelo autor é 'bent on', que imprime uma prosódia negativa ao seu contorno (p.ex. 'Labour 'bent on destroying education' says NUT Staff and agencies'[1]).

Em relação ao século XXI, dentre os cinco trabalhos que aí se encaixam o que mais chama a atenção devido a sua crescente atualidade é o de Adam Kilgarriff, intitulado 'Web as Corpus', de 2001 (capítulo 42). Esse é um pequeno artigo de três páginas publicado originalmente nos anais da conferência *Corpus Linguistics* de Lancaster de 2001. Nele, o autor afirma que o BNC foi o corpus dos anos de 1990, e a Web é o corpus do século XXI. Ninguém pode discordar disso, visto que a web veio para ficar e cresce a cada dia. O tamanho da web em qualquer das maiores línguas do mundo é muitas vezes maior do que o maior dos corpora de qualquer uma dessas línguas. Enquanto antigamente era preciso digitar textos impressos ou pedir fitas ou CDs de editoras com textos para corpora, hoje em dia com pouco esforço é possível baixar os mesmos textos diretamente da web. As facilidades que a web oferece, porém, são muitas vezes contrabalanceadas pelos desafios que ela impõe, muitos dos quais ainda sem solução. Por exemplo, como respeitar os direitos autorais de textos retirados da rede? Como limpar os arquivos baixados da rede de tal modo que o processamento não seja afetado por códigos e palavras não pertencentes ao texto em si? Como classificar textos 'nativos' da rede, como chats, blogs e emails? Como evitar textos repetidos da web? Como limpar o conteúdo de emails que têm várias outras mensagens incluídas? Todas essas questões e muitas outras se abrem agora que a web se firmou tanto como fonte de material para corpus como um corpus em si. O maior desafio, a nosso ver, está nessa segunda modalidade: usar a web como um corpus online, assim como fazem os buscadores de informação (p.ex. Google). Alguns programas já existem para fazer concordâncias diretamente de textos residentes na web, mas a falta de velocidade e de flexibilidade desses programas ainda não os fazem substitutos dos programas offline analisadores de corpus, como WordSmith Tools. Mas o caminho da web está aberto e não tem mais volta.

---

[1] education.guardian.co.uk/ schools/story/0,5500,1190952,00.htm [acessado em 31/1/2006]

3. Apreciação da obra

O nome Corpus Linguistics, para identificar os trabalhos que compõem o volume, não é tão óbvio quanto parece. Primeiramente, porque Lingüística de Corpus é um nome recente (que apareceu por volta do final dos anos de 1980) e se firmou apenas mais recentemente. Desse modo, muitas obras que ali estão não foram concebidas pelos seus autores como fazendo parte da Lingüística de Corpus pelo simples fato de essa área não existir na época. Em segundo lugar, muitos trabalhos recentes também não parecerem ter sido concebidos por seus autores como pertencentes à Lingüística de Corpus, mas sim à Lingüística Computacional e ao Processamento de Linguagem Natural. O fato de os organizadores da obra terem usado o nome Lingüística de Corpus como rótulo que encampa toda essa variedade de comunidades de pesquisa nos parece relevante e aponta para a projeção crescente da Lingüística de Corpus na pesquisa lingüística. Na introdução, os organizadores declaram ter evitado dividir o conteúdo em seções e tópicos para não dar uma visão compartimentalizada da área. Eles conseguiram não segmentar a área, mantendo a seqüência cronológica dos capítulos, mas isso torna a leitura seqüencial da obra muito mais difícil, pois não há coerência entre um capítulo e o que se segue. Por exemplo, o capítulo de Resnik e Yarowsky sobre *word-sense disambiguation* (31) é seguido de um outro sobre escrita de alunos de L1 e L2 (de autoria de Hyland e Milton). Vale lembrar que a relação entre as áreas que empregam tecnologia computacional para pesquisar a linguagem é pouco clara, em parte devido ao caráter dinâmico da computação. As linhas divisórias dessas áreas estão cada vez mais porosas, e o trânsito de pesquisadores entre as áreas 'humanas' (os chamados 'lingüistas') e as 'técnicas' (os chamados 'computeiros') tende a aumentar.

A leitura dos capítulos é facilitada em muito pelos resumos colocados no início de cada capítulo pelos editores. Esse resumo ajuda o leitor a antecipar o conteúdo do capítulo e também a apreciar a importância do trabalho no conjunto da obra. Embora úteis, esses resumos poderiam ter sido mais uniformizados, porém, visto que em alguns casos eles são extremamente curtos (como no caso do capítulo 41, de Campione e Véronis, com quatro linhas), enquanto outros são muito mais longos (o capítulo 37, de Böhmová e Hajicová tem um resumo de 24 linhas).

Embora a obra contenha um grande número de trabalhos, a seleção não é uniforme. A grande maioria dos trabalhos (32) situa-se nas décadas de 1990 e de 2000. Restam, assim, apenas dez trabalhos para os demais 38 anos de história; desses, há um trabalho dos anos de 1950 (Fries, 1952), um dos anos de 1960 (Francis, 1965) e um dos anos de 1970 (Aarts, 1971). O ano com maior número de contribuições é 1996, com sete artigos. A explicação mais óbvia para essa falta de equilíbrio é o fato de que os estudos de corpora eletrônicos são recentes, sendo, portanto, mais numerosos a partir de 1990 (quando, segundo os organizadores, teria havido uma 'explosão' dos estudos com corpora, p.5).

Como em toda obra de tal amplitude, pode-se apontar um outro autor que deveria ter sido incluído, mesmo sabendo que não é possível agradar a todos. De qualquer modo, sente-se falta na coletânea de trabalhos de uma figura histórica como Michael West (1953) e de contemporâneos como Susan Hunston (2002) e Mona Baker (1993), todos com trabalhos influentes sobre corpora.

4. Conclusão

Corpus Linguistics – Readings in a Widening Discipline é uma antologia interessante, pois resgata alguns trabalhos de difícil acesso (de livros que saíram de catálogo ou de números antigos de periódicos). Além disso, consegue reunir em um só volume uma grande quantidade de trabalhos. Não é, contudo, uma obra que apresenta um panorama consistente e coerente da Lingüística de Corpus, devido às razões que se seguem: a seleção de trabalhos dá excessiva ênfase aos anos de 1990; deixa de fora personagens importantes tanto históricos (Michael West) quanto contemporâneos (Susan Hunston e Mona Baker); mistura na seqüência trabalhos de descrição de língua com outros técnicos de engenharia de linguagem. Essa não é uma obra que um leigo poderia ler para se inteirar do que é a Lingüística de Corpus. Nem é uma obra para usar como leitura obrigatória em cursos avançados de Lingüística de Corpus. É uma obra para ter como referência caso se precise de um ou outro artigo de difícil acesso.

Recebido em fevereiro de 2006
Aprovado em janeiro de 2007
E-mail: tony4@uol.com.br

Referências

Aarts, F. 1971. On the distribution of Noun Phrase types in English clause structure. *Lingua, 26*: 281-293.

Baker, M. 1993. Corpus Linguistics and translation studies: Implications and applications. In M. Baker, G. Francis & E. Tognini-Bonelli. Eds. *Text and technology: In honour of John Sinclair*. Philadelphia/Amsterdam: John Benjamins.

Baker, M., G. Francis, & E. Tognini-Bonelli. 1993. *Text and technology - In honour of John Sinclair*. Philadelphia/Amsterdam: John Benjamins.

Biber, D. 1993. Using register-diversified corpora for general language studies. *Computational Linguistics, 19*: 219-241.

Chomsky, N. 1957. *Syntactic Structures*. The Hague: Mouton.

Francis, G. 1993. A corpus-driven approach to grammar -Principles, methods and examples. In M. Baker, G. Francis & E. Tognini-Bonelli. Eds. *Text and technology: In honour of John Sinclair*. Philadelphia/ Amsterdam: John Benjamins.

Francis, G., & S. Hunston. 1996. *Grammar Patterns 1: Verbs*. London: HarperCollins, COBUILD.

_____ 1998. *Grammar Patterns 2: Nouns and Adjectives*. London: HarperCollins, COBUILD.

Francis, W. N. 1965. A standard corpus of edited present-day American. English. *College English, 26*: 267-273.

Fries, C. C. 1952. *The Structure of English*. New York: Harcourt, Brace, and Company.

Hunston, S. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Hunston, S., & G. Francis. 2000. *Pattern Grammar - A Corpus-Driven Approach to the Lexical Grammar of English*. Amsterdam/ Philadelphia: John Benjamins.

Louw, B. 1993. Irony in the text or insincerity in the writer: the diagnostic potential of semantic prosodies. In M. Baker, G. Francis & E. Tognini-Bonelli. Eds. *Text and technology - Essays in honor of John McH Sinclair*. Philadelphia/Amsterdam: John Benjamins.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, J., G. Fox, & et al. 1990. *Collins COBUILD English Grammar*. London: Collins.

SINCLAIR, J. M., & ET AL. 1987. *COBUILD English Dictionary*. London and Birmingham: Collins COBUILD.
_____ 1995. *COBUILD English Dictionary*. London: Collins COBUILD.
_____ 1996. *Grammar patterns*. London: Collins COBUILD.
WEST, M. 1953. *A General Service List of English Words*. London: Longman.