Resenha: Carmen Scherer. *Korpuslinguistik*. Heidelberg: Winter, 2006 (98 p.)

Félix Bugueño Miranda

recebido em 27/07/2009 e aceito em 09/09/2009

Nascido Carmen Scherer, compiladora do Mainzer Zeitungskorpus, nos apresenta, nesta excelente "Einführung", um panorama claro, preciso e muito didático da interessante área da linguística de *corpus*. A disposição do texto foi claramente concebida para um leigo na matéria, sem perder, no entanto, em rigor. O viés didático diz respeito não somente à disposição dos tópicos, como logo será exposto, mas também à presença, ao longo de todo o texto, de quadros-resumo, atividades práticas e uma bibliografia complementar para aprofundamento nos diferentes temas.

O primeiro capítulo do livro ("Korpuslinguistik – was ist das?") começa com uma reflexão sobre a complementariedade entre o "theoretischer" e o "empirischer Ansatz" como fundamento da linguística. Carmen Scherer (doravante CS) refuta a aparente dicotomia que esses "Ansätze" representariam e salienta que eles perseguem o mesmo objetivo, "nämlich die Sprache zu beschreiben" (p.1). Amparada na pirâmide coseriana, CS considera que o "theoretischer Ansatz" lida com o sistema da língua, enquanto o "empirischer Ansatz" preocupa-se com o uso linguístico.

A autora define, a seguir, *corpus* como um conjunto de textos ou de excertos de texto [Textteilen] recolhidos e ordenados segundo determinados critérios linguísticos (p.3), salientando que os *corpora* podem ser escritos ou orais. Essa definição, aliás, é a que permitirá, no capítulo 5, uma aproximação ao controvertido papel da *web* como *corpus*.

Segundo CS, os fatores que determinam a definição de um *corpus* são o tamanho, o conteúdo, a consistência e a representatividade. No que diz respeito à representatividade, Carmen Scherer a julga "oberstes Ziel" na compilação de um corpus. A condição fundamental para alcançá-la é uma definição clara do objeto da linguagem que se deseja pesquisar. A consistência, por sua vez, diz respeito aos tipos de texto e à proporcionalidade com que eles constituirão um *corpus*.

Ao referir-se ao tamanho, a autora se encarrega de abalar a crença segundo a qual há uma relação direta entre o tamanho de um *corpus* e a sua qualidade ("so hört man häufig, je größer ein Korpus ist, desto besser" (p.6)). Em sua opinião, o importante é ter clareza sobre o objeto e o objetivo de uma pesquisa com *corpora*, de maneira que *corpora* com um número entre 10.000 e 20.000 ocorrências [Textwörter] podem oferecer dados confiáveis. Junto ao número total de ocorrências, CS julga que o tamanho dos textos e dos excertos que os compõem são também fatores fundamentais.

O quarto fator que determina um *corpus* é o seu conteúdo. O conteúdo corresponde aos tipos e gêneros textuais empregados na constituição do corpus.

Instituto de Letras / UFRGS, Av Bento Gonçalves, 9500,Bairro Agronomia, 91501-970 Porto Alegre (RS) / Brasil; Tel: 00-55-51-33086695; E-mail: felixv@uol.com.br

O capítulo I apresenta ainda vários parágrafos dedicados aos âmbitos de aplicação da linguística de *corpus*, tais como a linguística variacionista, a lexicografia, o ensino de línguas estrangeiras, a tradução e a linguística computacional. Especialmente interessantes são as considerações feitas sobre o uso de *corpora* na análise de erros e dificuldades como "feedback" no planejamento do ensino de línguas estrangeiras.

O capítulo II ("Arten von Korpora") está reservado, como seu nome indica, à descrição dos diferentes tipos de *corpora*. CS elenca dois macroparâmetros de classificação. O primeiro é um critério quantitativo, que classifica os *corpora* segundo o seu tamanho. Assim, há *corpora* pequenos (com até 1.000.000 de ocorrências), corpora de tamanho médio (vários milhões de ocorrências) e *corpora* grandes (com mais de 100.000.000 de ocorrências). Muito mais interessante, no entanto, é a classificação baseada em critérios qualitativos. Com um total de nove traços em oposição binária, é possível classificar todas as opções de *corpora* possíveis.

A primeira distinção é entre *corpora* processados digitalmente ("computerlesbar") e *corpora* não processados digitalmente ("nicht computerlesbar"). A autora lembra que a noção de *corpus* não está necessariamente atrelada ao meio digital, embora, nos últimos anos, um dos critérios empregados para definir *corpus* seja justamente o armazenamento e processamento digital. Na opinião de CS, *corpus* é um "Oberbegriff" que inclui ambas as acepções.

A segunda distinção diz respeito ao caráter dia- ou sinssitémico dos dados arrolados.

A terceira oposição é entre *corpora* de textos completos e *corpora* de excertos de textos ("Probenkorpora").

A quarta distinção é entre *corpora* estáticos ("statische Korpora") e *corpora* de monitoreamento. Os primeiros correspondem àqueles em que os critérios de seleção de textos permanecem inalterados. Nos *corpora* de monitoramento, esses critérios são modificados ao longo do tempo.

A quinta oposição é entre *corpora* de linguagem escrita e *corpora* de linguagem oral.

A sexta distinção é entre *corpora* de linguagem contemporânea e *corpora* históricos.

Uma oposição muito interessante pela qualidade dos dados que pode oferecer é feita entre *corpora* de referência e *corpora* especiais. Um *corpus* de referência objetiva representar uma língua na sua totalidade. Em termos coserianos, trata-se de um corpus diassitêmico, tal como o British National Corpus (BNC). Um *corpus* especial, por outro lado, preocupa-se com um eixo específico do diassitema. Nesse contexto, CS salienta o valor desse tipo de *corpus* para o ensino-aprendizagem de uma língua estrangeira, lembrando o caso específico do "Fehler-annotierter Lernkorpus des Deutschen als Fremdsprache" (FALKO).

A última distinção proposta pela autora está relacionada com o número de línguas. Os *corpora* são classificados como mono- ou multilingües. Entre os *corpora* multilíngues, por sua vez, é preciso distinguir entre *corpora* paralelos e *corpora* comparados [vergleichbare Korpora].

O capitulo III ("Analyse von Korpusdaten"), por outro lado, está dedicado à descrição dos procedimentos metodológicos no trabalho com *corpora*, apresentando conceitos-chave tais como "ocorrência" [Textwort], *type* / *token*, *hapax legomena*, "análise quantitativa", "análise qualitativa", "concordância", etc. Destacam-se não

somente a exposição clara e com abundante exemplificação, mas também a aplicação dos conceitos apresentados na análise de *corpora* já existentes. Esses exemplos de aplicação permitem compreender ainda melhor como avaliar os dados obtidos. CS recomenda que a comparação de dados seja feita utilizando-se, como base quantitativa da comparação,1.000.000 ocorrências em cada *corpus*. A autora, empregando os dados arrolados na comparação de três *corpora* do alemão (p.38), lembra que se deve observar uma relação entre as ocorrências do fenômeno sob análise e o tamanho do *corpus*.

Comentário à parte merecem os parágrafos dedicados à análise colocacional (p. 46-48), pela complexidade que esse fenômeno apresenta. Para a linguística de *corpus*, "colocação" é o padrão regular de palavras que acompanham uma determinada unidade léxica. A caracterização apresentada pela autora é muito parecida com a de outras teorias: "(...) palavras que aparecem tipicamente em combinação com uma unidade determinada" [Wörter die typischerweise in Verbindung mit einem Zielwort auftreten] (p. 46). Os exemplos arrolados pela autora, no entanto, permitem comprovar que não se trata exatamente do conceito de colocação adotado por teorias lexicológicas e/ou lexicográficas. CS oferece a seguinte listagem de "Kollokationspartnern" para *Hund*: *Leine, bellen, Herrschen, Rassen, beißen, Schwanz, wedelt, Gassi, Haustiere, Zucht, streicheln* (p.47). Os exemplos se aproximam mais do que Coseriu chama de "solidariedades léxicas". A autora parece reconhecer a dificuldade de compreender esse fenômeno linguístico, já que considera que essas unidades léxicas fazem parte do "conhecimento estereotipado de determinados conceitos" [stereotypisches Wissen über bestimmte Begriffe] (p.47). A sua análise projeta-se, assim, muito mais em direção a uma semântica de frames, embora não se aluda a ela ao longo da exposição.

A parte final do capítulo III está dedicada a considerações referentes aos índices de frequência, salientando-se as possíveis aplicações que esses dados podem ter.

O capítulo IV ("Arbeiten mit einem eigenen Korpus") apresenta os procedimentos para montar um *corpus*. Segundo CS, há quatro passos que devem ser seguidos sequencialmente. Em primeiro lugar, formular de maneira precisa qual o objetivo da pesquisa; em segundo lugar, decidir entre compilar um corpus próprio ou procurar um corpus já existente; em terceiro lugar, levantar os dados a partir do *corpus*, e, finalmente, interpretá-los à luz do objetivo da pesquisa.

No que diz respeito à formulação de um objetivo de pesquisa, CS propõe as coordenadas que ajudam nessa tarefa: 1) o(s) ponto(s) do diassistema a ser(em) considerado(s), 2) o nível de organização da língua que será pesquisado e 3) o viés qualitativo ou quantitativo com que os dados serão analisados (p. 53).

Em relação ao emprego de *corpora* já existentes ou à necessidade de compilar um corpus próprio, a autora lembra que nem sempre essa última alternativa é a melhor opção. De fato, o capitulo V está reservado integralmente a apresentar a variada gama de recursos que *corpora* já existentes oferecem. Se se conclui, no entanto, que um *corpus* próprio é necessário, Carmen Scherer oferece uma detalhada explanação das questões que devem ser levadas em conta para essa tarefa. Curiosamente, começa salientando que nem sempre a digitalização dos dados é o melhor caminho, como acontece, por exemplo, quando se trabalha com materiais muito antigos redigidos em caligrafias, tais como a letra gótica [Fraktur], que, ao serem escaneados, apresentam um alto índice de erros de leitura.

O terceiro passo no processo de compilação corresponde ao etiquetamento. A autora fornece informações muito detalhadas sobre como proceder com *tagging* e *parsing* (p. 58-59).

Finalmente, o último passo, a interpretação dos dados, é a consequente aplicação dos três passos anteriores.

O quinto e último capítulo ("Arbeiten mit bestehenden Korpora") é, talvez, um dos mais interessantes do livro para o germanista, pois apresenta, com grande luxo de detalhes e ótimos exemplos, todos os recursos disponíveis em três *corpora* do alemão, que, aliás, estão disponíveis livremente na internet. Carmen Scherer, no entanto, introduz o capítulo avaliando a validade da *web* como *corpus*. É bom lembrar que no nosso meio há uma aberta rejeição ao uso da rede como *corpus*. CS, em consonância com o espírito analítico dos alemães, examina a questão adotando como premissa a definição de *corpus* como conjunto de textos, e conclui que, sob essa perspectiva, a *web* é sim um corpus. No entanto, se *corpus* é definido como um conjunto de textos compilados segundo critérios linguísticos, então a resposta é negativa (p. 74). A autora considera que um grande empecilho ao se considerar a *web* como um *corpus* é o fato de que se desconhece o seu tamanho e, como já tinha sido explicado nos capítulos anteriores, esse parâmetro é essencial; posto que os resultados elencados de um *corpus* devem ser avaliados em relação ao seu tamanho. Sendo assim, não há como fazer uma análise quantitativa. Análises qualitativas, no entanto, são possíveis (p. 75). Por outro lado, Carmen Scherer lembra também que a *web* permite o acesso a uma enorme quantidade de dados linguísticos autênticos, sendo possível documentar a existência de unidades tais como expressões idiomáticas e palavras de baixa frequência, que, muitas vezes, constituem dados marginais em *corpora* existentes. Da mesma forma, os filtros de algumas *search engines* permitem buscar por domínios determinados. Julgamos que a análise feita pela autora é um "acerto de contas" equilibrado em relação à *web*. Concordamos plenamente com as suas palavras. O que a *web* pode efetivamente fazer é mostrar tendências da língua geral. No entanto, se o objetivo de um *corpus* é tirar conclusões quantitativas, conforme o exposto por CS, a *web* é, de fato, um *corpus* deficitário.

A parte final do capítulo V está dedicada à descrição detalhada de três *corpora* do alemão: o DWDS-Kernkorpus (Digitales Wörterbuch der deutschen Sprache), o conjunto de corpora do IDS, composto por cinquenta *corpora*, e o TIGER-Korpus, que tem como traço diferencial ser o único *corpus* etiquetado apto para pesquisas no nível da frase e não da palavra, como é o caso dos dois primeiros. A autora oferece abundante exemplificação de como esses *corpora* podem ser empregados para pesquisa, salientando que eles estão à disposição gratuitamente na internet.

Em síntese, Carmen Scherer nos oferece um panorama abrangente e rico em informações e exemplificações do interessante mundo da linguística de *corpus*. Uma leitura obrigatória para o linguista, seja ele germanista ou não.