Resenha/Review

Berber Sardinha, Tony. 2004. *Lingüística de Corpus.* Barueri, SP: Editora Manole.

Resenhado por/by: Izabella dos Santos Martins
(LAEL/PUC-SP-CNPq)

A descoberta da pólvora. Uma revolução. Um admirável mundo novo, cheio de possibilidades e caminhos nunca d´antes navegados. Logo no início da leitura do livro de Tony Berber Sardinha – que define a Lingüística de Corpus como área de estudos que trata da "Coleta e da exploração de corpora, ou conjunto de dados lingüísticos textuais coletados criteriosamente, com o propósito de servirem para a pesquisa de uma língua ou variedade lingüística" (p. 3) –, o leitor não familiarizado com o tema abordado deve, ao se deparar com a Lingüística de Corpus, sentir-se impactado com essas várias impressões passando por sua cabeça. Os leitores já familiarizados com certeza sentir-se-ão presenteados com a publicação do livro, que é um verdadeiro tratado teórico e manual prático, de grande utilidade para os que lidam com corpora.

O autor do livro, Tony Berber Sardinha é professor associado do Departamento de Lingüística e do Programa de Estudos Pós-Graduados em Lingüística Aplicada e Estudos da Linguagem da Pontifícia Universidade Católica de São Paulo (PUC-SP), e não é exagero dizer que é o maior expoente da área de Lingüística de Corpus no Brasil. O autor vem apresentando uma produção profícua, que culminou no lançamento do livro em questão, cuja leitura permite ao leitor intuir – mesmo que de forma comedida, como convém ao discurso acadêmico – seu entusiasmo e sua paixão pelo trabalho com corpora.

São várias as qualidades do livro. Chama a atenção o seu didatismo: o autor passa para o leitor, de forma clara, concisa e sem o rebuscamento e o maneirismo típicos de alguns textos acadêmicos, os conceitos, as teorias e os procedimentos básicos necessários para o trabalho com corpora. Outro

ponto notável do livro é a riqueza de referências bibliográficas: por ser obra que se propõe a tratar do mais variado leque de temas relacionados à Lingüística de Corpus, trata-se de qualidade muito bem vinda, que demonstra a maturidade do autor para executar tarefa de tal porte, bem como a expertise necessária para fazer comparações entre obras e autores e para criticá-las. Em suma, o livro reúne o melhor dos dois mundos: consegue agradar a leigos e aos membros da comunidade acadêmica.

À parte o seu pioneirismo – não há outros livros de autores brasileiros sobre Lingüística de Corpus –, o livro aborda o tema em todos os seus aspectos principais, ilustrando o estado da arte da matéria no Brasil e no mundo. O autor discute conceitos – não os apresenta como dados nem como inquestionáveis –, informa sobre os corpora disponíveis para pesquisa atualmente, apresenta obras de referência – bem como críticas a elas –, traça o histórico e levanta perspectivas de pesquisa na área. Como bônus, o livro apresenta ainda um dicionário de freqüências inédito do português brasileiro. Outro destaque da obra é seu caráter de manual, principalmente no que concerne aos procedimentos de utilização do WordSmith Tools, o programa de análise lingüística mais completo usado pelos lingüistas do corpus.

Uma questão abordada no livro – que talvez seja a mais relevante e seu ponto de maior destaque – é a explicação sobre a mudança de paradigma que o advento do trabalho com corpora proporcionou para a Lingüística. Ela pode ser resumida na frase de Fillmore (1992, 35): "Todo corpus me ensinou coisas sobre a linguagem que eu não teria descoberto de nenhum outro modo". Dessa maneira, o racionalismo é colocado em xeque e o empirismo é apontado como novo modelo e forma de pensar na Lingüística. Muitos achados, e uma quantidade surpreendente de evidências lingüísticas, só são possíveis de obter pela observação e o trabalho com a linguagem em uso, autêntica, pressupostos para a pesquisa em Lingüística de Corpus (LC). O advento do computador, devido à sua grande memória e capacidade de armazenamento, teve e tem um papel central nessa mudança de olhar sobre a linguagem, já que proporciona sistematização aos fatos e os evidencia.

É recomendável, antes de comentar individualmente os capítulos do livro, expor alguns conceitos que são caros à LC, e que constituem os seus fundamentos. A LC é uma **abordagem empirista** da **linguagem**, que é

vista como um **sistema probabilístico**. Quando se fala em empírico, o que se quer significar é que aos dados obtidos, oriundos da observação da linguagem, é dada primazia, e a teorização é feita *a posteriori* É uma visão que se choca com a dos modelos mentalistas, que apregoam que a linguagem deve ser estudada por meio da introspecção, e que o conhecimento provém de princípios estabelecidos *a priori*. O processamento cognitivo da linguagem, bem como seus modelos estruturais de funcionamento, são o foco da pesquisa de vertentes mentalistas. Quando se diz que a linguagem é vista como sistema probabilístico, o que significa é que a língua é vista mais como uma questão de probabilidade que de possibilidade: ou seja, embora os traços lingüísticos sejam possíveis teoricamente, eles não ocorrem com a mesma freqüência, essas diferenças não serem aleatórias é o fato mais importante. Para esta abordagem da linguagem, há uma correlação entre as características lingüísticas e contextuais. Para seus seguidores, a linguagem é vista como padronizada. Os padrões (colocações, coligações ou estruturas) podem ser lexicais e léxico-gramaticais e apresentam regularidade e variação sistemática. A **padronização** é evidenciada pela recorrência, pela repetição sistemática: pelo conhecimento da freqüência atestada pode-se estimar a probabilidade teórica. Citando Berber Sardinha (:40), a padronização é uma "regularidade expressa na recorrência sistemática de unidades coocorrentes de várias ordens (lexical, gramática, sintática etc.)". Segundo o autor, os padrões podem ser de três tipos: a colocação – tipo de padrão mais enfocado nos estudos de corpora – vem a ser a "associação entre itens lexicais, ou entre léxico e campos semânticos". A coligação é a "associação entre itens lexicais e gramaticais" e a prosódia semântica, por sua vez, é a "associação entre itens lexicais e conotação (negativa, positiva ou neutra) ou instância avaliativa".

Nos modelos mentalistas, o que interessa aos pesquisadores é a determinação de quais agrupamentos sintáticos são permissíveis, dado o conhecimento que o falante possui de sua língua: o que está em foco é a competência lingüística, ao contrário do foco dos empiristas, que está no desempenho, no uso lingüístico. Os lingüistas de corpus aplicam à linguagem um **princípio** que chamam de **idiomático**. Para eles, a língua apresenta uma grande quantidade de frases pré ou semi construídas, que são escolhidas de forma única (a escolha de um elemento implica na escolha de outro), apesar de parecerem escolhas segmentadas. Em outras palavras, a linguagem é vista como formada por porções lexicais. Segundo Sinclair,

que para Berber Sardinha é "o maior lingüista de corpus da história" (:13), o espaço formado pelo léxico e pela sintaxe é uno: a escolha de um item lexical ou de uma categoria gramatical reduz as possibilidades de escolha tanto lexicais quanto gramaticais. Na lingüística sistêmico-funcional de Michael Halliday, que apresenta uma visão da linguagem muito parecida com a de Sinclair e seus seguidores, esse espaço é chamado de léxico-gramática.

Sobre o estatuto da Lingüística de Corpus, Berber Sardinha expõe as três visões de diferentes pesquisadores sobre o tema. Para um grupo, a LC é vista como uma metodologia, já que seu instrumental pode ser aplicado livremente em várias disciplinas, sem mudar a orientação teórica destas. Entendendo metodologia como um "modo típico de aplicar um conjunto de pressupostos de caráter teórico, então a Lingüística de Corpus pode ser vista como uma metodologia" (:.36). Um segundo grupo defende que a LC é mais que uma metodologia, uma vez que seus praticantes produzem conhecimento novo, não adquirível com o uso de outras ferramentas e outros pressupostos teóricos. Finalmente, o terceiro grupo vê a LC como uma abordagem, uma perspectiva, uma maneira de enxergar a linguagem.

Segundo Kennedy (citado por Berber Sardinha: por pg), as pesquisas em LC concentram-se em quatro áreas principais: a compilação de corpus; o desenvolvimento de ferramentas; a descrição da linguagem e a aplicação de corpora.

No capítulo 1, o livro apresenta uma visão geral sobre a Lingüística de Corpus, a começar por seu histórico. História fascinante, por sinal. O autor nos lembra que o primeiro corpus lingüístico eletrônico, o Brown University Standard Corpus of Present-Day American English, foi lançado em 1964, e era composto por 1 milhão de palavras. Ainda naoNa época, o paradigma dominante na Lingüística era o mentalista, propalado por Noam Chomsky, que apenas sete anos antes havia lançado seu *Syntactic structures*, livro que continha as bases de uma visão da linguagem que acredita que os dados necessários para o lingüista estão em sua própria mente, e podem ser acessados por meio da introspecção.já disse acima Uma vez que a Lingüística de Corpus explora a linguagem por meio de evidências empíricas, extraídas por computador, vê-se o tamanho da revolução e do estranhamento causado pelo lançamento do corpus Brown e com tudo que isso trouxe em seu ensejo: uma verdadeira revolução no modo de pensar sobre, de ver e de

lidar com a linguagem, baseado nesta que pode ser considerada a palavra-chave do trabalho com corpus: empirismo.nao causou nenhum impacto

O autor esclarece que havia corpora (cujo sentido original é corpo, conjunto de documentos) antes do advento do computador, desde a Antigüidade, principalmente de citações da Bíblia. Ao longo do século XX – ao contrário de hoje, em que predomina o trabalho de descrição de linguagem – a ênfase dos trabalhos com corpora era o ensino de línguas. Como lembra o autor, foi um corpus não computadorizado que deu feição aos corpora atuais, o Survey of English Usage, compilado em Londres a partir de 1959. Atualmente, é tendência, como observa o autor, a parceria entre universidades que desenvolvem pesquisas em LC e editoras ou outras empresas, principalmente com vistas à criação de dicionários, como é o caso do pioneiro dicionário Cobuild, primeiro a ser compilado a partir de um corpus computadorizado, parceria entre a editora Collins e a Universidade de Birmingham. No Brasil, a LC ainda está em estágio inicial.

Berber Sardinha expõe alguns marcos da LC, destacando-se Sinclair (1966) – que, segundo ele, foi o trabalho pioneiro na área de pesquisa lexical, tendo traçado os rumos da maior parte dos trabalhos em LC que são desenvolvidos atualmente – e Synclair (1991), em que o autor disserta sobre várias idéias centrais na área de LC, em especial sobre a colocação. Além das obras importantes, o autor lista ainda outros veículos importantes de divulgação de pesquisas e trabalhos na área e isso nao é importante?. O Autor apresenta muitas outras referencias

A seguir, é analisada a definição de corpus, destacando alguns critérios necessários para que um conjunto de dados lingüísticos possa ser considerado um corpus: a **origem** (os dados devem ser autênticos e escritos por falantes nativos – a menos que recebam o título de corpora de aprendizes, que são aqueles cujos dados são textos de falantes não nativos; o **propósito** (os dados devem ser objeto de estudo lingüístico); a **composição** (os dados devem ser escolhidos e colhidos com critério); a **formatação** (os dados devem ser legíveis por computadores); a **representatividade** (os dados devem ser representativos de uma língua ou de uma variedade lingüística, o que na prática significa dizer que o corpus deve ser o maior possível) e a **extensão** (o material deve ser vasto para ser representativo).

No capítulo 2, o autor trata de questões práticas que surgem quando se escolhe trabalhar com a LC, como coleta, limpeza e organização do cor-

pus no computador. A www como fonte de pesquisa de corpus mereceu atenção especial, sobretudo o papel dos *offline browsers*, ferramentas que podem ser usadas para coleta em massa de textos para criação de corpus. O passo a passo da utilização da ferramenta é descrito pelo autor, bem como o caminho a ser percorrido quando da utilização do Perl e do emulador de Unix Cygwin, ambos usados para limpeza dos textos. A seguir, Berber Sardinha ilustra a limpeza de textos em html e de códigos SGML. Sobre a busca em massa, Berber Sardinha salienta que um utilitário muito usado para esses procedimentos é o Grep, que é usado na linha de comando, uma vez que não possui interface gráfica. Sobre a substituição em massa, o autor observa que há opções além do Microsoft Word e do Perl – como é o caso do utilitário Sed, que pode ser usado com eficiência para fazer substituições em massa de textos do corpus, procedimento que não pode ser feito a contento usando-se o Word.

O Text Converter, que faz parte do WordSmith Tools, é apontado como boa opção para o procedimento de busca e substituição em massa, podendo ser usado para fazer uma ou várias substituições ao mesmo tempo. Sobre a organização do corpus, o autor afirma que não há regras gerais a serem seguidas, mas salienta que os textos devem estar de preferência em uma pasta principal em que só existam textos do corpus. O item "Criação de Cabeçalhos" – que vêm a ser uma parte do arquivo de cada texto do corpus que contém informações sobre ele – é abordado a seguir. O autor ensina que essas informações podem ser codificadas de várias maneiras, sendo uma delas as etiquetas de SGML, e as de tipo Cocoa. Uma outra maneira bastante usada por pesquisadores é a criação de arquivos separados de cabeçalho.

O capítulo 3 é inteiramente dedicado a um programa extremamente útil para a análise do corpus, o "WordSmith Tools" versão 3, de autoria de Mike Scott. O programa, obtido pela internet, pode ser baixado gratuitamente na versão demo e, se houver interesse, pode ser comprado. No ato da compra da licença, o usuário recebe um código que transforma a versão demo em versão completa. Logo na introdução do capítulo, o autor enumera algumas vantagens de um maior emprego dos computadores na investigação da linguagem: o fato de serem consistentes, podendo realizar tarefas tediosas de modo confiável e eficiente; o fato de permitirem ao pesquisador a possibilidade de lidar com uma maior quantidade de dados, devido à sua grande memória e, finalmente, o fato de o uso dos compu-

tadores permitir a descoberta de dados novos, que muitas vezes podem levar à contestação de crenças estabelecidas.

A seguir, o autor discorre sobre as ferramentas do WordSmith Tools. São elas: WordList, que possibilita a criação de listas de palavras por ordem alfabética ou de freqüência; KeyWords, que possibilita, através da comparação entre 2 listas de palavras-chave, a criação de listas de palavras-chave que, por sua vez, são palavras cujas freqüências são estatisticamente diferentes no corpus de estudo e no corpus de referência, conhecido também como corpus de controle – cujo tamanho deve ser 2, 3 ou 5 vezes maior que o corpus de estudo; e Concord, que possibilita a criação de concordâncias – ou listas das ocorrências no corpus de um item determinado pelo usuário, acompanhado do texto ao seu redor (seu cotexto). Segundo Berber Sardinha, três princípios básicos norteiam o funcionamento do WordSmith Tools. São eles: 1) Ocorrência: os itens devem necessariamente ocorrer num corpus, ser observáveis; 2) Recorrência: os itens devem ocorrer pelo menos duas vezes no corpus; 3) Coocorrência: os itens devem estar na presença de outros, mas não é necessário que estejam aparecendo seqüencialmente.

Em relação à lista de palavras-chave, são explicitadas os critérios de escolha dos corpora de estudo e de referência, bem como todos os procedimentos para obtenção das listas de palavras-chave – incluindo aí as operações que são usadas pelo programa, explicitando o tipo de raciocínio utilizado. Isso, aliás, o autor faz ao longo de todo o livro: lançar luzes em um terreno um tanto sombrio. Ao colocar todos os porquês e os comos para o leitor, este acaba se convencendo de que não se está lidando com nenhum dogma matemático, o que dá um estímulo principalmente para o leitor que tem ressalvas com procedimentos quantitativos e estatísticos. Sobre as concordâncias – instrumentos indispensáveis no estudo da padronização lexical e da colocação –, o autor apresenta os instrumentos que a ferramenta disponibiliza para o usuário e mostra o passo a passo de sua utilização. De grande utilidade para a pesquisa em LC são as listas de colocados, que elencam as palavras que ocorrem ao redor da palavra de busca (ou nódulo), em posições determinadas.

Muitas vezes, de acordo com o tipo de pesquisa que se quer realizar, será necessário marcar no corpus a classe gramatical de cada palavra. Para isso, o autor apresenta uma ferramenta importante: o etiquetador mor-

fossintático, que insere automaticamente no corpus códigos que indicam a classe gramatical de cada palavra, e permite o tratamento de grandes quantidades de texto rapidamente. A etiquetagem pode ser morfossintática, sintática, semântica ou discursiva. Há duas opções de etiquetadores: os que rodam on-line e os que podem ser instalados no computador. Uma vez que, para funcionarem de maneira correta, alguns etiquetadores exigem que o corpus esteja itemizado, o autor dedica uma seção do livro para tratar da itemização – separação das unidades ortográficas. Berber Sardinha instrui sobre o uso de alguns programas de itemização, como o Text Converter do WordSmith Tools, o Perl, o Java, sobre a etiquetagem por e-mail, fornecida pela Universidade de Leeds, e sobre a etiquetagem via web, oferecida pela Universidade do Sul da Dinamarca. Esta última opção, aliás, é realçada pelo autor como sendo uma ótima opção para os pesquisadores cujos corpora de pesquisa são escritos em língua portuguesa, uma vez que há no etiquetador via web uma interface para a etiquetagem do português.

O capítulo 5 do livro de Berber Sardinha trata de uma ilustração prática do processo de desenho, planejamento e pré-processamento de um corpus. Usando a sua própria experiência da execução do banco de dados do projeto DIRECT (Em Direção à Linguagem de Negócios)– que reúne textos em português e inglês de linguagem profissional –, o autor oferece dicas valiosas, que podem ser aproveitadas por pesquisadores dos mais diversos temas que se aventuram no trabalho com LC.

O capítulo 6 oferece uma análise das freqüências do corpus Banco de Português. Essa análise, como ressalta o autor, é útil para o entendimento da língua como um todo, já que se trata de um corpus extenso e de linguagem geral. O autor lembra que uma vantagem evidente de se analisar as freqüências de um corpus grande – como é o caso do corpus em questão – é a possibilidade de se conhecer quais palavras são freqüentes e quais são raras. O estudo da freqüência das palavras é importante até mesmo para quem nunca pensou na questão no viés aqui proposto – leia-se os pesquisadores de tradição chomskyana: a freqüência das palavras chega até a implicar nas- ou mesmo determinar – as normas gramaticais. Segundo Biderman (1998, citado por Berber Sardinha, p. 163), "a norma lingüística nada mais é do que a média dos usos freqüentes das palavras que são aceitas pelas comunidades dos falantes".

O capítulo 7 trata das concordâncias, desenvolvidas para a observação dos padrões de uso das palavras de um corpus. Nas palavras do autor, a

concordância é "uma listagem de ocorrências de um item específico, dispostas de tal modo que a palavra de busca (aquela que se tem interesse em investigar) aparece centralizada na página (ou na tela do computador) (p. 187). Essas palavras são acompanhadas de seu cotexto – as palavras que aparecem próximas a elas num corpus. O autor difere essas palavras dos colocados – que são as palavras que ocorrem ao redor do nódulo, em posições relativas a ele. A palavra de busca é definida pelo usuário, é uma escolha pessoal, enquanto que os colocados são todas as palavras que ocorrem ao redor do nódulo. O capítulo evidencia elementos práticos da análise da linguagem por meio de concordâncias, e apresenta um glossário com a definição dos termos mais freqüentes no trabalho com elas.

O autor cita vários programas concordanciadores, recomendando o MicroConcord e o WordSmith Tools Concord. São apresentadas algumas limitações, como em relação à quantidade de linhas de concordância (16000 é o limite para o WordSmith Tools e 1600 para o MicroConcord). Em relação a esse problema, Berber Sardinha explica o procedimento que deve ser tomado pelo usuário quando o corpus apresenta mais do que esse número de ocorrências para determinada palavra de busca – e aí está outro trunfo do livro: os problemas são expostos com honestidade, e são sempre acompanhados de sugestões para que o usuário faça sua pesquisa o melhor possível.

As listas de colocados são também abordadas pelo autor, no que concerne a procedimentos de obtenção e classificação. Item especial é dedicado às estatísticas de associação, e são apresentadas três maneiras (razão observado/esperado; informação mútua e escore T) estatísticas de calcular se a associação entre itens lexicais pode ser considerada não aleatória a ponto de poder ser considerada uma colocação; em outras palavras, são demonstradas maneiras de se saber se a associação é mais comum do que o esperado. O autor ensina a maneira de se obter as estatísticas de associação por meio de cálculo em planilhas eletrônicas, no programa WordSmith Tools e via internet.

O capítulo seguinte trata da padronização da língua portuguesa segundo a Lingüística de Corpus. Usando o já citado Banco de Português, Berber Sardinha estuda a padronização da partícula **só**, e chega a alguns achados importantes. A importância maior da pesquisa está em demonstrar como a léxico-gramática do português pode ser compreendida melhor

quando é estudada com base na abordagem de corpus. O capítulo seguinte trata da comparação entre as prosódias semânticas de duas línguas, usando para tanto as ferramentas de concordâncias e as listas de colocados. O capítulo é especialmente relevante para os estudiosos de Estudos da Tradução, que podem ver o quão útil o trabalho com corpus pode ser para obterem resultados consistentes em suas pesquisas. São mostrados os resultados das comparações entre itens lexicais em português e em inglês, tais como causar/cause e acontecer/happen. O autor conclui que, devido à relevância dos achados, as informações obtidas por meio do procedimento descrito deveriam constar de dicionários e glossários.

O capítulo 10 trata da Lingüística de Corpus e sua conexão com a Lingüística Aplicada, especialmente no tocante ao ensino de língua estrangeira. O autor faz considerações sobre a descrição da linguagem nativa e sobre a linguagem de aprendizes, atendo-se ao uso das concordâncias no ensino (bem como nas avaliações), que servem sobretudo para exemplificar o uso de traços lingüísticos e as situações nas quais ele ocorre. Berber Sardinha (p.279) elenca alguns benefícios advindos do uso das concordâncias no ensino, quais sejam: a) Obtenção de respostas a perguntas sem resposta nas obras de referência; b) Desenvolvimento do espírito pesquisador; c) Independência em relação ao professor, ao curso, ao livro didático e aos materiais de referência; d) Incentivo à postura ativa do aluno; e) Centramento no aluno e individualização do aprendizado. O autor apresenta ao leitor três abordagens de uso da Lingüística de Corpus no ensino de línguas: a) o Currículo Lexical, que se centra no léxico e se norteia pela máxima de que os sentidos mais comuns da linguagem são expressos pelo vocabulário mais freqüente; dessa maneira, o conteúdo de cada nível se pautou na freqüência do vocabulário, atestada no corpus do projeto Cobuild. Nessa abordagem, as atividades são centradas em tarefas; b) a Abordagem Lexical, que também tem no léxico seu elemento central. De acordo com Berber Sardinha, difere-se do Currículo Lexical sobretudo pela ênfase na colocabilidade do léxico, uma vez que este é descrito por meio de porções léxico-gramaticais, que são itens prefabricados, realizadas como colocações ou polipalavras e ensinadas através de textos. Nesta abordagem, as atividades são centradas na forma, e não em tarefas; c) Aprendizado Movido por Dados, que visa a tornar o aluno um pesquisador. Aqui, ao contrário do que acontece nas abordagens anteriores, nas quais a abordagem da LC é vista na produção dos materiais, o uso do computador e das concordâncias é conduzido às últimas conseqüências, com os próprios alunos pro-

duzindo as concordâncias e induzindo regras, que não são dadas *a priori*. Os alunos seguem três etapas: identificação, classificação e generalização.

O último capítulo do livro trata do estudo da variação com Lingüística de Corpus, através da Análise Multidimensional que, nas palavras do autor (p. 299) é "uma abordagem para análise de corpus que usa procedimentos estatísticos (principalmente análise fatorial), visando ao mapeamento das associações entre um conjunto variado de características lingüísticas dentro de um corpus de estudo". Uma dimensão é definida como "conjunto de traços que subjazem a um corpus". Essa abordagem tem um caráter comparativo, já que promove contraste entre textos ou registros, e combina análises macro com análises micro. Neste capítulo, são descritos termos e conceitos empregados neste tipo de análise, bem como as etapas na realização de uma análise multidimensional. A seguir, o autor tece considerações sobre a descrição multidimensional da língua inglesa, além de comentar sobre esse tipo de análise de outras línguas.

Com tudo isso, conclui-se que a empreitada de escrever um livro de tal alcance foi realizada a (muito) contento. Pudera: falando-se de Lingüística de Corpus, o "top of mind" da comunidade acadêmica brasileira é sem dúvida Berber Sardinha. Não é à toa.

Recebido em fevereiro de 2007
Aprovado em junho de 2007
E-mail: izabella.martins@gmail.com

REFERÊNCIA

FILLMORE, C. Corpus linguistics or computer corpus linguistics. In: *Directions in corpus linguistics. Proceedings of nobel symposium 82, Stockholm, Ed. Jan Svartvik, 35-60.* Berlim/Nova York, De Gruyter, 1992.