LINGUÍSTICA DE CORPUS: HISTÓRICO, METODOLOGIA, CAMPOS DE APLICAÇÃO

Simone Sarmento[*]

RESUMO: *A Linguística de Corpus (LdC) é uma área que tem recebido muita atenção de linguistas nos últimos anos por ter revelado uma nova forma de enxergar, estudar e entender as línguas. Dessa forma, o objetivo deste artigo é oferecer um panorama geral da LdC. Começo com um breve histórico e discorro sobre as principais características da LdC, os diferentes tipos de corpora, e as variadas formas de analisar corpora. Serão também definidos termos específicos da área como colocação, fraseologia e prosódia semântica. Finalmente, mostrarei algumas das aplicações da LdC, as limitações das pesquisas baseadas em corpora e as principais vantagens em utilizar a LdC.*

PALAVRAS-CHAVE: *Linguística de Corpus; Linguística Empírica; Colocações.*

ABSTRACT: *Corpus Linguistics (CL) is an area which has been receiving growing attention from linguists in the last few years as it has revealed a new way to view, study, and understand languages. This, this paper aims at offering an overview of CL. I start by presenting a brief historical background of the area and then listing its main characteristics, the different kinds of corpora and various ways to analyze corpora. Also, specific terms such as collocation, phraseology and semantic prosody are defined. Finally, some of CL's applications are shown as well as the limitations of corpus based research and, conversely, its main advantages.*

KEYWORDS: *Corpus Linguistics; Empirical Linguistics; Collocation.*

PANORAMA HISTÓRICO

Desde os anos 60, os corpora eletrônicos têm sido considerados um valioso recurso para o estudo linguístico. Apesar de o seu uso ainda ser motivo de controvérsia (HUNSTON, 2002; McENERY e GABRIELATOS, 2006), sua contribuição ao ensino de línguas assim como à linguística é amplamente reconhecida. Essa contribuição foi primeiramente sentida na linguística inglesa devido ao trabalho pioneiro de corpora de língua inglesa, como o Brown Corpus (FRANCIS e KUCERA, 1964) que deu origem à grande parte dos trabalhos produzidos utilizando corpora desde então.

O Brown Corpus foi o primeiro corpus computadorizado compilado para fins de pesquisa linguística. Entretanto, sua importância torna-se ainda maior se levarmos em consideração o fato de essa compilação ter acontecido em um momento em que o paradigma cada vez mais dominante, liderado por Noam Chomsky, era totalmente contrário ao registro e à pesquisa do desempenho linguístico. Para os linguistas gerativo-transformacionais, o

[*] Doutora em Estudos da Linguagem/UFRGS; Professora adjunta do Instituto de Letras da UFRGS.

estudo da língua deve descrever não o que os falantes fazem com a língua (desempenho), mas o que os falantes sabem sobre uma língua (competência). Ou seja, fundamenta-se no estudo da linguagem através da introspecção para a verificação dos modelos de funcionamento da linguagem. Os dados estão na mente do linguista, e acessíveis através da introspecção. As teorias são verificadas através de frases inventadas, muitas vezes pelo próprio pesquisador. Esse paradigma racionalista perdurou como predominante por um longo período, distanciando grande parte das pesquisas linguísticas dos estudos descritivos de desempenho. Chomsky, entre outros, discordava do uso de corpora e dos modelos de competência probabilísticos baseados em estatística, derivados do estudo do desempenho linguístico. Foi nesse ambiente acadêmico adverso ao uso de corpora, que Nelson Francis e Henry Kucera iniciaram o que a enorme tarefa de compilar um corpus sincrônico de aproximadamente um milhão de palavras representativas do inglês escrito publicado nos Estados Unidos em 1961. O trabalho foi finalizado em 1964 com velocidade surpreendente tendo em vista os recursos da época. O Brown Corpus foi então disponibilizado em fita de computador acompanhado do respectivo manual para o usuário.

Nascia mais uma dicotomia linguística: de um lado os linguistas neo-firthianos, britânicos na sua maioria, que tratam os corpora como repositórios de instâncias do uso real da língua, nos quais os exemplos que se encaixam ou não em uma teoria, ou que sustentam ou não um ponto de vista em uma discussão devem ser selecionados querendo o pesquisador ou não (McENERY e GABRIELATOS, 2006). Do outro lado, temos os chomskianos americanos, ou seja, linguistas que buscam na intuição e nos exemplos introspectivos (em grande parte inventados pelos próprios pesquisadores) suas fontes de dados. Entretanto, Fillmore acredita que os linguistas deveriam fazer uso dos dois paradigmas, apesar das dificuldades, uma vez que ambas as correntes têm muito a contribuir para a área.

CARACTERÍSTICAS DA LINGUÍSTICA DE CORPUS

O termo "linguística de corpus" (LdC) é entendido (MCENERY e WILSON, 1996) como o estudo da linguagem baseado em exemplos da vida real. A LdC não é um ramo da linguística como a sintaxe, a semântica ou a pragmática, que concentram-se na descrição ou explicação de algum aspecto da língua em uso (RAYSON, 2002). A LdC é uma metodologia que pode ser aplicada a uma grande variedade de estudos linguísticos, ou ainda ao ensino de línguas, ou seja, é uma das várias maneiras de fazer linguística.

Biber, Conrad e Reppen (1998. p. 4) listam as características essenciais da linguística baseada em corpus:

· É empírica, ou seja, analisa os padrões reais de uso em textos

naturais;

· Utiliza uma grande coletânea de textos [1] (um corpus, com princípios de coleta pré-estabelecidos) como base para análise;

· Faz um extenso uso de computadores para análise, podendo também utilizar técnicas automáticas e interativas;

· Depende de técnicas analíticas quantitativas e qualitativas.

Os corpora são usados para gerar conhecimento empírico sobre uma língua, que pode suplementar, ou muitas vezes suplantar, informações provenientes de fontes de referência e introspecção (LEECH, 1991, 1992). Stubbs (2001) acrescenta ainda que a LdC vê a linguagem como sendo um sistema probabilístico, ou seja, embora muitas combinações e características linguísticas sejam possíveis, nem todas são prováveis de ocorrer. Dessa forma, por ser uma técnica adequada à análise estatística, os corpora podem fornecer informações sobre a frequência relativa de muitos aspectos da língua. Berber Sardinha afirma que "o mais importante da diferença de frequências entre traços é não serem aleatórios" (2004, p. 31). Se essas diferenças fossem aleatórias, o fator frequência não seria significativo e não adicionaria informações a respeito da estrutura da língua. Contudo, grupos de características linguísticas apresentam uma variação sistemática em textos específicos, variações oriundas de situações comunicativas específicas. A variação sistemática, ou seja, a recorrência de traços linguísticos (colocação, coligação, padrão sintático, entre outros) indica que a linguagem é padronizada (*patterned*) e motivada por diversos fatores além das necessidades comunicativas. Por exemplo, ao escolher o determinante *the* as escolhas das palavras subsequentes são automaticamente limitadas, isto é, adjetivo, advérbio ou substantivo. Outros fatores influenciam a seleção de palavras, tais como, a proficiência linguística do autor, colocações, tópico, tipo de texto, e, no caso deste trabalho, normas de redação para os manuais de aviação. Os padrões apresentam regularidades e variações sistemáticas em variedades textuais, dialetais, etc. A verificação dessas regularidades não pode ser alcançada através da intuição de um falante nativo. Somente a observação empírica de dados reais, em diferentes contextos de uso pode fornecer essa informação. Cabe assim dizer, que a frequência de ocorrência de traços linguísticos, não constitui uma constatação trivial, como havia afirmado Chomsky.

TIPOS DE CORPORA

Um corpus pode ser definido como uma "coletânea de exemplos

[1] Berber Sardinha (2004, p. 17) salienta que no lugar de "textos" a expressão "porções de linguagem" parece mais adequada devido aos problemas relacionados à delimitação do conceito de texto. Para o autor, na LdC, pode-se considerar um artigo científico, seu resumo inicial ou ainda um trecho de um diálogo como textos.

naturais de linguagem, que consistem desde algumas frases até conjuntos de textos escritos ou gravações orais que foram coletados para serem usados como base para pesquisa linguística" (HUNSTON, 2002, p. 2). Mais recentemente, a palavra "corpus" (cujo termo mais comumente aceito para o plural é corpora) tem sido usada para referir-se a coletâneas de textos (ou partes de textos) que são armazenadas e que podem ser acessadas por meio de computadores. Textos escritos retirados de jornais ou revistas podem ser escaneados, retirados de um CD ou da internet. Textos orais, como conversas, são gravados e posteriormente transcritos; ou seja, são copiados palavra por palavra de forma que os textos dessas conversas possam ser alimentados em um computador. Torna-se assim possível analisar a língua contida no corpus através de softwares específicos para o estudo linguístico, como por exemplo, o *Wordsmith Tools* (SCOTT, 1996).

Entende-se por "exemplos naturais de linguagem" (conforme citação anterior) aqueles exemplos que não tenham sido produzidos, ou criados, para serem utilizados em um corpus. Berber Sardinha (2004) acrescenta que a idéia de natural inclui também o fato de a linguagem ser produzida por humanos, excluindo, assim, programas de geração de textos. Entretanto, apesar de os textos serem naturais, um corpus é um objeto artificial, pois foi criado com a finalidade específica da pesquisa.

Não há uma especificação do tipo de conteúdo que um corpus deveria conter. Um corpus pode conter desde a obra completa de Shakespeare, até instruções expressas nas caixas de sabão em pó, ou textos jornalísticos sobre o Grêmio Football Porto-Alegrense no ano em que (quase) foi campeão brasileiro. Com relação à dimensão, não há um consenso quanto ao tamanho mínimo ou máximo aceito para um corpus. Segundo Hoffmann (1998[2007]), as interpretações sobre o tamanho mínimo necessário para os corpora nas pesquisas linguísticas divergem amplamente. Na pesquisa de linguagem especializada, já foram obtidos resultados úteis com amostras de 35.000 palavras, mas sugere a dimensão de 200.000 palavras como mínima. O autor defende que o tamanho dependerá dos objetivos da pesquisa e do tipo de corpus. Desta forma, "não há nenhuma fórmula matemática amplamente aceita que informe a quantidade ou distribuição de palavras ou textos que um corpus deva ter para ser representativo" (BERBER SARDINHA, 2000, p. 104). Entretanto, a maior parte das palavras tem frequência de ocorrência muito baixa e para que elas apareçam em um corpus é necessário que ele possua um grande número de palavras. O mesmo pode ser dito com relação aos diferentes sentidos ou significados de uma mesma palavra: há os mais e os menos frequentes. Os sentidos mais raros terão uma maior probabilidade de aparecer em um corpus maior.

A questão da representatividade envolve ainda conhecer o "todo" que, no caso da linguagem, não é conhecido. Deve-se tentar dividir esse todo estimado em partes. Por exemplo, um corpus de "linguagem

jornalística" deve incluir diferentes tipos de jornais, os populares e os mais tradicionais, por exemplo. Deve também incluir textos das diferentes seções, como variedades, esportes, editoriais, negócios, entre outras. Para ser considerado representativo e equilibrado, um corpus desse tipo deveria incluir um número aproximado de palavras em cada categoria: negócios nos populares, negócios nos tradicionais, esporte nos populares, esporte nos tradicionais, etc.

Conforme Hunston (2002), outro detalhe relacionado ao tamanho de um corpus é a velocidade e a eficiência do software de acesso a esse corpus, assim como a capacidade do computador de acessá-lo. Se, por exemplo, obter a listagem das formas do presente e passado do verbo *to be* levar mais de alguns minutos, o pesquisador pode preferir utilizar um corpus menor, cujo resultado pode ser considerado tão confiável quanto o de um corpus maior, mas para o qual o software trabalhará muito mais rapidamente.

O objetivo da pesquisa também influencia o tamanho que um corpus necessita ter. Carter e McCarthy (1995, p. 143) afirmam que para estudar gramática na linguagem falada, um corpus relativamente pequeno pode ser suficiente, pois as palavras gramaticais tendem a ser muito frequentes. Por outro lado, itens de baixa frequência necessitam de um corpus bem maior.

Os corpora são geralmente coletados com base em um projeto de pesquisa linguística específico em mente, tal como fornecer informações sobre frequências para verbetes de dicionários, ou produzir material didático para o ensino de língua estrangeira, um dos propósitos deste trabalho. Algumas vezes, contudo, os corpora são coletados sem um propósito específico e são disponibilizados como um recurso da língua geral para linguistas, professores de línguas, lexicógrafos, entre outros.

Há vários tipos de corpora dependendo do tamanho, propósito e forma como foram compilados. Sinclair (1995) sugere a seguinte lista de tipos de corpora:

-Corpus Geral- Um corpus contendo muitos tipos de texto. Pode incluir linguagem escrita, falada ou ambas; textos produzidos em um país ou vários. Por ser de cunho geral, muito provavelmente esse tipo de corpus não será representativo de nenhum "todo" (como por exemplo, um corpus que contenha todos as bulas de remédio de um laboratótio), mas incluirá o maior tipo de textos possível. Um corpus de língua geral precisa ser muito maior do que um corpus específico. É muitas vezes utilizado como contraste em relação aos corpora mais especializados. Por essa razão, são por vezes denominados de Corpora de Referência. Um dos corpora mais famosos da língua inglesa é o *British National Corpus* (BNC).

-Corpus Monitor- Corpus projetado para verificar mudanças atuais em uma língua. Esse tipo de corpus é alimentado anualmente,

mensalmente, ou até mesmo diariamente, aumentando de tamanho rapidamente. Entretanto, a proporção de tipos de texto mantém-se constante, de forma que cada período de tempo possa ser comparado com o anterior. Um exemplo é o *Bank of English*, que atualmente conta com cerca de 400 milhões de palavras.

-Corpus Comparável- Dois (ou mais) corpora em línguas diferentes (inglês e português, por exemplo) ou em diferentes variedades de uma língua (português do Brasil e de Portugal, por exemplo). São compilados seguindo as mesmas diretrizes, isto é, conterão a mesma proporção de gêneros: textos jornalísticos, romances, conversas informais, etc. Podem ser usados por tradutores ou por aprendizes para identificar diferenças e equivalências em cada língua. O exemplo mais citado desse tipo de corpus é o ICE (*International Corpus of English*), que contém mais de um milhão de palavras de diversas variedades da língua inglesa.

-Corpus Paralelo- Dois (ou mais) corpora em línguas diferentes contendo textos que foram traduzidos de uma língua para outra (por exemplo, um romance traduzido do inglês para o português), ou textos que foram produzidos simultaneamente em duas ou mais línguas (por exemplo, normas da União Européia).

Além dos tipos de corpora citados acima, Hunston (2002, p.14) adiciona ainda os seguintes:

-Corpus de Aprendiz- Uma coletânea de textos-redações produzidos por aprendizes de uma língua. O propósito desse tipo de corpus é identificar em que aspectos os aprendizes diferem entre si e em relação a falantes nativos (em comparação a um corpus de falantes nativos). Provavelmente, o mais conhecido seja o *International Corpus of Learner English* (ICLE), que consiste em redações escritas em inglês por falantes de várias línguas nativas (Português, Francês, Alemão, etc.).

-Corpus Pedagógico- Corpus que consiste na linguagem a qual um aprendiz é exposto. Pode consistir de livros didáticos e gravações. Esse tipo de corpus pode, por exemplo, ser comparado a um corpus de linguagem autêntica (produzida sem propósitos pedagógicos) para verificar se o aprendiz está sendo exposto à linguagem útil e natural.

-Corpus Histórico ou Diacrônico- Um corpus de textos de diferentes períodos de tempo. É utilizado para averiguar o desenvolvimento de certos aspectos de uma língua através dos tempos.

-Corpus Especializado- Um corpus contendo um tipo específico de texto (ou gênero), tal como resumos (*abstracts*), artigos acadêmicos sobre um assunto específico, conversas telefônicas, etc. Esse tipo de corpus tem por objetivo ser representativo de certo tipo de texto, ou linguagem. É comumente compilado pelo próprio pesquisador para refletir o tipo de linguagem que quer investigar. Não há limite para o grau de especialização envolvido, mas há parâmetros para limitar o tipo de texto incluído. Um

exemplo é o Corpus de Aviação (*Aviation Corpus*), compilado pela autora (SARMENTO , 2008), contendo textos de três diferentes manuais da aeronave 737 da *Boeing*.

ANÁLISE DE CORPORA

Um corpus, como já dito anteriormente, é um repositório de textos digitais. Para que seu conteúdo seja acessado, é necessário que haja recursos, ou ferramentas para tal. Os corpora maiores, como o BNC, geralmente possuem seus próprios recursos ou ferramentas de acesso. Outros corpora, por exemplo, o Corpus de Aviação (SARMENTO, 2008), necessitam ser armazenados e acessados através de programas específicos para a descrição linguística, como é o caso do *WordSmith Tools* ou do Corpógrafo [2]. Em qualquer uma das formas acesso, os recursos mais utilizados nas investigações linguísticas são:

- Concordâncias;
- Lista de frequência de palavras;
- Lista de colocados [3]

Concordâncias

O concordanciador é provavelmente a ferramenta computacional mais utilizada para processar informações em um corpus. Um concordanciador é um programa que busca, em um corpus, uma palavra selecionada ou um sintagma, apresentando todas as ocorrências daquela palavra ou sintagma no centro da tela do computador com as palavras que as antecedem ou seguem à esquerda e à direita, isto é, o co-texto. A palavra selecionada que aparece no centro da tela é chamada de nódulo ou palavra nódulo (*node* ou *node-word*). O material é disposto de forma a facilitar a visualização dos padrões da palavra-nódulo. Assim, as observações de padrões como colocação [4], coligação [5] e prosódia semântica [6] são otimizados. No exemplo abaixo foi utilizado o corpus Brown, a palavra nódulo escolhida foi o VM *must*, e a busca [7] foi aleatória:

---

[2] O Corpógrafo é um software de acesso livre disponível em http://poloclup.linguateca.pt/ferramentas/gc/

[3] Apesar de estarem sendo citadas nesta seção, a listas de colocados serão discutidas na seção seguinte.

[4] Colocação refere-se à forma na qual duas ou mais palavras são tipicamente usadas juntas. Por exemplo, usa-se *heavy rain*, mas não *heavy sun*. Nesse contexto, *heavy* é colocado de *rain*, ou *heavy* e *rain* são colocados.

[5] Coligação refere-se à associação entre itens lexicais e gramaticais. Por exemplo, *start* é mais comum com sintagmas nominais e orações-*ing*, enquanto *begin* é mais usado com um complemento *to*" (BERBER SARDINHA, 2004, p. 40).

[6] Prosódia semântica é o termo usado para referir a palavra ou expressão usadas em um contexto específico de tal forma que a palavra/expressão adquira certa conotação daquele contexto. Um exemplo seria a expressão *sit through* (HUNSTON, 2002, p. 141), que é geralmente usada com itens

[7] Esse tipo de busca é também conhecido como KWIC –*Key Word in Context*

1 ce, one of two alternative courses must be taken: _1._ Five
2 cent of the voters in each county must sign petitions requestin
3 llot, or _2._ The Republicans must hold a primary under the
4 sertion, and ADC dependency". #"MUST SOLVE PROB-
LEM"# The mont
5 Co&, committee chairman. "We must solve the problems which
6 negative side of the balance sheet must be set some disappointme

A mesma busca pode ser realizada com as palavras ordenadas alfabeticamente à direita da palavra nódulo (*sort right*):

1 inue in his chosen profession, he must abandon his own code and
2 and to which even law enforcement must accommodate itself.
One
3 union members under contract, it must accomplish its payroll b
4 of time, a method of preservation must accomplish the destructi
5 ating the antennae and receivers, must account for much of the d
6 ades across Berlin the free world must acquiesce in dismemberme

Ou, ainda, com as palavras ordenadas à esquerda (*sort left*):

1 es cleaning and drying equipment a must for modern gin operat
2 ry style in fine fashion and is a must for those who want to col
3 education (read "reading") was "a must". He moved in a "highly
4 ques. The platform accelerometers must be slightly modified
for
5 erials and library accommodations must be planned. In the secon
6 was modern, large, on five acres. Must have cost plenty. The St

Lista de frequência de palavras

Uma lista de frequência é simplesmente uma lista de todas as formas, ou vocábulos (*types)* em um corpus juntamente com o número de ocorrências de cada forma/vocábulo (*tokens*). A lista pode ser classificada por ordem de frequência, com as formas mais frequentes em primeiro lugar, ou, ainda, alfabeticamente. Essa listagem pode também ser lematizada ou não. A comparação de listas de frequência pode fornecer informações interessantes sobre os diferentes tipos de textos, uma vez que para a LdC, textos são formatados por textos anteriores, através de repetições ou através de rotinas e convenções. "Os textos são historicamente herdados". (STUBBS, 1996, p. 34).

Essa comparação é especialmente importante entre corpora especializados. Kennedy (1998, p. 102) salienta que "quanto mais especializado" for um corpus (inglês acadêmico comparado com inglês geral, ou inglês para economia, comparado com inglês acadêmico geral) maior será o número de palavras lexicais (ou com mais conteúdo) entre as mais frequentes. Nesse sentido, o autor menciona que em um corpus de economia, 18, entre as 50 palavras mais frequentes são lexicais; enquanto

em um corpus de inglês acadêmico geral, somente três, entre as 50 mais frequentes são lexicais; no corpus Birmingham, por exemplo, somente *said* (considerado lexical) está entre as 50 mais frequentes, as outras 49 são consideradas palavras gramaticais.

Na comparação entre dois corpora, somente o resultado normalizado não é prova suficiente de que o resultado é significativo, ou seja, que não é aleatório. A aplicação de testes estatísticos pode fornecer resultados até 99% confiáveis de que as diferenças são motivadas, ou seja, não são aleatórias.

Conforme Rayson (2002), o teste estatístico com melhores resultados para a comparação da frequência de palavras ou expressões entre dois corpora é o *Log-Likelihood* (ou LL). Se o resultado obtido após a aplicação do LL foi de 6,63 ou mais, a probabilidade de a diferença entre os dois corpora ter acontecido aleatoriamente é de menos de 1%. Dessa forma, o pesquisador pode estar 99% confiante de que o resultado é significativo. Esse resultado é geralmente expresso como $p<0.01$[8].

Palavras que são significativamente (com base estatística) mais frequentes em um corpus que em outro são também conhecidas como "palavras chave" (*keywords*). O *WordSmith Tools* (SCOTT, 1996) inclui um recurso que compara dois corpora (geralmente um maior, mais geral, e outro menor, mais especializado) automaticamente.

COLOCAÇÃO, PADRONIZAÇÃO E FRASEOLOGIA

Firth, com sua célebre frase "*You shall know a word by the company it keeps*" mostrou ao mundo da linguística a importância dos estudos descritivos da linguagem, especialmente, a importância do co-texto de uma palavra de forma a conhecê-la. Nesse sentido, as colocações habituais das palavras são simplesmente os acompanhantes desta palavra. Ligado ao conceito de colocação, há os conceitos de fraseologia e padronização. Apesar de esses conceitos não serem necessariamente dependentes da LdC, os apresento em detalhes nesta seção, pois, ao referir-me à padronização e à fraseologia, estarei referindo aos conceitos relacionados à LdC. Da mesma forma, por serem termos polissêmicos, considero importante estabelecer de forma clara e precisa o que se entende por padronização e fraseologia.

Como já mencionado anteriormente, colocação refere-se à tendência com que as palavras co-ocorrem com outras. A palavra *toy* (brinquedo), por exemplo, co-ocorre com *children* (crianças) mais frequentemente do que com *men* (homens) e *women* (mulheres) (HUNSTON, 2002, p. 68). Hunston considera essa colocação como motivada, pois há uma explicação lógica para tal. Outras colocações, no entanto, como *strong tea* (chá forte) e *powerful car* (carro potente), não possuem uma motivação aparente.

[8] O cálculo do LL pode ser realizado automaticamente no site http://ucrel.lancs.ac.uk/llwizard.html.

As colocações podem ser observadas com números absolutos, constituindo uma "associação entre itens lexicais" (Berber Sardinha, 2004, p. 200), mas tornam-se mais confiáveis se medidas estatisticamente. Dessa forma, é possível verificar até que ponto a relação palavra-nódulo e colocado não é aleatória, ou seja, uma "associação não aleatória entre itens lexicais" (ibid). Para Berber Sardinha (2004, p. 201): "Uma associação não-aleatória é aquela que é mais comum do que o esperado. Para saber se uma associação entre palavras não é aleatória, precisamos de apoio estatístico, na forma de medidas estatísticas de associação".

Programas que calculam a colocação em números absolutos, contam as ocorrências de todas as palavras que ocorrem em um certo horizonte, por exemplo, quatro palavras para à esquerda da palavra-nódulo e quatro para à direita. É importante ter acesso a uma grande quantidade de dados para calcular a colocação, principalmente da língua geral, pois assim, haverá mais confiabilidade de que as colocações mais significativas são realmente mostradas.

São vários os testes de significância utilizados. Os mais comuns são o qui-quadrado (*chi-square*), o informação-mútua (*mutual information*), o *z-score* e a razão do LL. Rayson (2002) e Gómez (2002) reportam que o qui-quadrado torna-se não confiável quando a frequência esperada é muito pequena (menos do que cinco), possivelmente superestimando a significância de palavras muito frequentes e/ou ao comparar um corpus relativamente pequeno com outro muito maior. Para Rayson (2002), o LL é preferido sobre o qui-quadrado. Além disso, Gómez (2002) menciona que a Informação Mútua tende a superestimar o grau de associação quando os eventos são raros.

Um dos importantes usos das informações provindas das listas de colocações é ressaltar os diferentes significados de uma palavra. Hunston (2002, p. 76) lista os colocados do verbo *leak*. Algumas das associações são relacionadas ao significado físico de *leak*: *oil*, *water*, *gas*, *roof*; enquanto outras são associadas ao sentido metafórico: *information*, *report*, *memo*, *confidential*. Em outras palavras, a lista de colocados fornece uma espécie de perfil semântico das palavras envolvidas.

Sinclair (1991) sugere que quando duas palavras de frequências diferentes são consideradas como colocados, a colocação tem um valor diferente na descrição de cada uma dessas palavras. Se uma palavra "a" é duas vezes mais frequente do que uma palavra "b", cada vez que elas ocorrerem juntas é duas vezes mais importante para "b" do que para "a", pois o evento é responsável por duas vezes a proporção da ocorrência de "b" com relação à "a". Ou seja, quando todas as ocorrências de "a" com "b" são calculadas, um resultado é registrado no perfil de "a" e outro resultado é registrado no perfil de "b". Dessa forma, Sinclair considera separadamente esses dois tipos de colocação, e utiliza o termo "nódulo"

para a palavra que está sendo estudada e "colocado" para qualquer palavra que ocorre no ambiente específico de um nódulo. Cada palavra consecutiva em um texto é assim, nódulo e colocado, embora nunca ao mesmo tempo. Quando uma palavra mais frequente "a" se coloca com uma palavra menos frequente "b", Sinclair denomina de "colocação descendente", e o contrário de "colocação ascendente". Segundo o autor, a colocação ascendente não possui um valor estatístico significativo, e a maioria das palavras tendem a ser elementos de estruturas gramaticais ou hiperônimos. A colocação descendente, por sua vez, demonstra uma análise semântica da palavra.

Conforme Berber Sardinha (2004), "De um modo geral, a padronização é a regularidade expressa na recorrência sistemática de unidades co-ocorrentes de várias ordens (lexical, gramatical, sintática, etc.)" (ibid, p. 47). Isto é, para que sejam definidos os padrões de uma palavra, faz-se necessário averiguar as palavras e as estruturas frequentemente associadas a ela que de alguma forma refletem no seu significado. Para Berber Sardinha (2004), "padrão" e "fraseologia" são muitas vezes utilizados como sinônimos. Dessa forma, há por vezes o emprego de expressões como "a fraseologia da palavra X" referindo-se aos padrões observáveis da palavra em questão. Seja qual for o termo utilizado para essa descrição, ela é considerada de extrema relevância para o ensino de língua estrangeira, pois aspectos como naturalidade e fluência são demonstrados por meio de padrões.

Sinclair (1991) considera a fraseologia [9] como a base da descrição linguística desafiando outras visões sobre a linguagem. Para o autor:

-Não há distinção entre padrão sintático e significado;

-A língua possui dois princípios de organização, o princípio idiomático (*idiom principle*) e o princípio da livre escolha (*open-choice principle*);

-Não há distinção entre léxico e gramática.

Se uma palavra possui vários sentidos, cada sentido tende a ser associado com um conjunto diferente de padrões. Por exemplo, quando o adjetivo *mobile* é usado para qualificar coisas, significando "que pode ser carregado" (*can be moved*), geralmente precede o substantivo, por exemplo, *mobile unit*, *mobile library*. Quando o mesmo adjetivo é usado para qualificar pessoas, significando "não impedido de mover-se por doença ou falta de recursos" (*not prevented from moving by disability or lack of resources*), geralmente sucede o verbo de ligação, por exemplo *I'm still very mobile* (HUNSTON, 2002, p. 139). Sinclair (1991) acredita que essa noção de fraseologia possa substituir a palavra isolada como unidade para o ensino de vocabulário, simplificando, dessa forma, a tarefa do aprendiz, uma vez que cada item

[9] Fraseologia é uma palavra polissêmica que se refere a diferentes fenômenos. Para outras definições ver Altenberg (1998), Bevilacqua (2001) e Robertson (1988).

lexical conteria mais informações sobre o seu uso.

Ainda nesse sentido, palavras com o mesmo padrão tendem a compartilhar aspectos comuns de significado. Hunston (2002, p. 140) cita o caso da sequência: "verbo seguido por substantivo seguido por *as* seguido por substantivo". Nesses casos, a associação entre padrão e significado é tão forte que o significado parece pertencer à frase inteira, e não a cada palavra individual. Verbos com esse padrão parecem significar "fazer com que alguém ou alguma coisa seja ou pareça ser algo". Por exemplo:

-he described it as a legalised theft;
-he revealed himself as a man of deep culture;
-I would like to appoint you as managing director.

> Da mesma forma, a prosódia semântica somente pode ser conhecida através da observação de um grande número de ocorrências de uma palavra/expressão, pois baseia-se no uso típico da palavra/expressão. Como a prosódia semântica nem sempre faz parte do conhecimento consciente de um falante (seja ele nativo ou não), pode não ser ensinada, mas muitas vezes pode consistir emum aspecto importante da linguagem. O ensino de vocabulário deveria levá-la em consideração. Entretanto, isso só poderá acontecer se a abordagem for fraseológica e não baseada na palavra.

Relacionado ao conceito de fraseologia, Sinclair (1991, p. 115) estabelece dois princípios organizadores da língua, simultaneamente alternativos e complementares, a partir dos quais é possível interpretar o significado das palavras: (i) o princípio da livre escolha (*open-choice principle*), em que o falante tem como única restrição a gramaticalidade do enunciado; (ii) o princípio idiomático (*idiom principle*), em que o falante tem à sua disposição um grande número de grupos de palavras pré-construídos (ainda que possam apresentar alguma variação, nomeadamente no plano lexical, flexional ou de ordem das palavras). Quando uma elocução não pode ser interpretada sob o princípio idiomático, o usuário da língua recorreria ao princípio da livre escolha. Hunston (2002) menciona a dificuldade em provar, ou não, a existência do principio idiomático, mas argumenta que algumas sequências de palavras notoriamente constituem fraseologias, ou combinatórias, como atestado em corpora maiores. Nesse sentido, para a autora, não é irracional supor que tais combinatórias serão codificadas e decodificadas como entidades únicas, e não como uma composição de significados de palavras individuais.

Sinclair (1991) sugere que quaisquer grupos ou sequências de palavras são construídos e entendidos a luz de um dos dois princípios, mas nunca de ambos simultaneamente. Ou seja, o significado pode ser construído pela frase como um todo, operando de acordo com a fraseologia convencional, ou pelas palavras individualmente, operando de acordo com

as regras gramaticais. A escolha entre os princípios idiomáticos ou da livre escolha tornam a ambiguidade teoricamente possível; o fato de que apenas um ou outro princípio seja empregado por um usuário da língua em um momento, explica por que a ambiguidade é raramente um problema para falantes ou ouvintes. Hunston (2002) oferece *grasp the point* como exemplo. Para a autora, a frase é ambígua. De acordo com o princípio idiomático, ela significaria "entender a idéia principal de algo" (*understand the main idea of something*); interpretado de acordo com o princípio da livre escolha, *grasp* combinado com qualquer objeto sólido, significaria "segurar a ponta de algo" (*take hold of the sharp end of something*). Entretanto, falante e ouvinte utilizariam apenas um dos princípios, (possivelmente) eliminando a ambiguidade.

Sinclair (1991) salienta que entender linguagem como fraseologia (uma visão obtida observando-se uma grande porção de linguagem, e não palavra por palavra), necessita a rejeição de léxico e gramática como entidades separadas. Para o autor não há uma diferença essencial entre "palavras lexicais" e "palavras gramaticais". Além disso, os padrões observáveis de itens lexicais são observações sobre o léxico e a gramática.

A visão mais tradicional sobre a linguagem é de que palavras lexicais são facilmente distinguíveis das palavras gramaticais, e que fatos lexicais, tais como colocação, são separados de fatos gramaticais tais como transitividade. A distinção entre palavras gramaticais e lexicais é baseada em algumas noções: de que palavras gramaticais são mais frequentes que palavras lexicais; que palavras gramaticais são mais facilmente relacionadas paradigmaticamente, ao passo que palavras lexicais sintagmaticamente; e que palavras gramaticais não possuem significado próprio, mas que palavras lexicais possuem. Algumas palavras gramaticais possuem meios formais de identificação. Os verbos modais [10] em inglês, por exemplo, não flexionam, mas os verbos auxiliares (de uma forma geral), sim. Ainda sobre as palavras gramaticais na língua inglesa, essas são as únicas com menos de três letras no inglês escrito [11].

Entretanto, as diferenças listadas no parágrafo acima são desafiadas por evidências provenientes das pesquisas com corpora. Embora as palavras gramaticais sejam geralmente as mais frequentes em listas de palavras, e as palavras lexicais sejam menos frequentes, nem todas as palavras gramaticais são mais frequentes do que as palavras lexicais. No BNC, por exemplo, *and*, *it*, *is*, *was*, *I*, *that*, *you*, *be*, *he* e *are* são as mais frequentes. Contudo, o verbo lexical *said*, é mais frequente do que *up*, *in*, *did*, entre outras palavras gramaticais. Sinclair (1999) também ressalta que algumas palavras gramaticais muito frequentes tais como *a*, participam do principio

[10] Os verbos modais parecem ocupar uma posição intermediária no continuum palavra gramatical e palavra lexical.
[11] Com exceção da palavra "*ax*" do inglês americano que quebra essa regra.

idiomático da mesma forma que palavras lexicais. Um exemplo de *a* nessa situação seria a expressão *come to a head*, em que o artigo indefinido *a* não estaria em contraste com o artigo definido *the*. Ou seja, o comportamento paradigmático é anulado, pois se comporta da mesma forma que a palavra lexical *head* na mesma expressão. Hunston 2000 [12] apud HUNSTON 2002) salienta que a fraseologia pode distinguir entre os significados dos VMs *may* e *must* da mesma forma que entre palavras lexicais.

A terceira distinção tradicional é que palavras gramaticais não possuem significados, ao contrário de palavras lexicais. Hunston (2002, p. 150) lembra que a palavra *lexical point*, em expressões como *from your point of view*, *the point is that* e *from that point on* tem um significado isolado bastante fraco. Ao passo que a palavra gramatical *would* deveria ser ensinada como tendo um significado próprio e não como parte de uma abstração gramatical chamada de "condicional". Sinclair resume o princípio idiomático afirmando que "o usuário de uma língua tem disponível um grande número de sintagmas semi-pré construídos que constituem escolhas únicas, embora pareça que os sintagmas possam ser analisados em segmentos "(SINCLAIR, 1991, p. 110).

Para o autor, o fenômeno do princípio idiomático parece refletir a recorrência de situações semelhantes na vida e relações humanas; pode assim ilustrar uma tendência natural à economia de esforço; ou pode ser em parte motivado pelas exigências da conversação em tempo real. Independentemente de sua motivação, esse fenômeno foi relegado a uma posição inferior na maior parte dos estudos linguísticos por não se encaixar no modelo do princípio da livre escolha.

CORPORA E APLICAÇÕES

Como já visto, um dos pontos fortes da LdC reside na sua natureza empírica que agrupa um grande número de dados tornando a análise linguística mais objetiva. Nesta seção mostro a aplicação de corpora em várias áreas da linguística.

Estudos do léxico e lexicografia

McEnery et al. (2006) mencionam que os corpora revolucionaram a elaboração de dicionários de tal forma que praticamente todos os dicionários (principalmente os da língua inglesa) publicados a partir de 1990 são baseados em corpora. A maior vantagem do uso de corpora na lexicografia é de natureza automatizada que permite que lexicógrafos consigam extrair exemplos típicos e autênticos do uso de um item lexical de uma grande

[12] HUNSTON, S. Phraseology and the modal verb: a study of pattern and meaning. In: Heffer, C. e Saunston, H. *Words in Context: a tribute to John Sinclair on his retirement*. University of Birmingham CD-ROM, 2000

quantidade de dados em apenas alguns segundos. Outra vantagem é relacionada às informações sobre frequência e quantificação das colocações que um corpus pode fornecer. Hunston (2002, p. 96) resume as mudanças ocasionadas pelo uso de corpora na elaboração de dicionários em cinco ênfases:

1. Frequência
2. Colocação e fraseologia
3. Variação
4. Léxico na gramática
5. Autenticidade

Estudos gramaticais

McEnery et al. (2006) citam duas obras importantes no estudo da gramática da língua inglesa: *A Comprehensive Grammar of the English Language* (Quirk et al. 1985) e mais recentemente a *Longman Grammar of Spoken and Written English* (ou LGSWE de Biber et al. 1999). A LGSWE é baseada em um corpus de 40 milhões de palavras e propõe-se a descrever a gramática inglesa através de exemplos reais focalizando tanto o inglês escrito quanto o inglês oral. Esse último tem sido pouco explorado nas gramáticas. A LGSWE também considera diferenças dialetais e entre registros.

Variação e análise de gênero

Um dos aspectos do estudo linguístico que tem sido assistido pelo desenvolvimento de corpora é o estudo da variação entre linguagens produzidas em diferentes situações. Apesar de haver uma longa tradição na investigação de diferentes registros [13] e de gêneros, os corpora adicionaram uma nova dimensão aos tipos de pesquisa que podem ser conduzidas. O estudo da variação é essencialmente o estudo de comparações entre discursos produzidos em momentos diferentes, com diferentes objetivos ou propósitos, por diferentes grupos de pessoas, ou sob diferentes condições. Enquanto diferenças claras e significativas são facilmente identificadas entre diferentes registros, é também possível identificar diferenças claras e significativas intra-registros. Hunston (2002) menciona diversos trabalhos que encontraram diferenças entre artigos acadêmicos de diferentes disciplinas, (BIBER et al., 1998; GLEDHILL, 1995 [14]). São vários os parâmetros a serem utilizados entre os registros/gêneros. Um

[13] Um registro é um subconjunto de uma linguagem utilizado para um propósito específico ou em um contexto social específico. É importante salientar que, principalmente na literatura sobre a LdC, as noções de "registro" e "gênero" se sobrepõem, não havendo uma distinção clara entre elas.

[14] GLEDHILL, C. Collocation and Genre Analysis. The Phraseology of Grammatical Items in Cancer Research Abstracts and Articles. 1995. In S. Botley, J. Glass, T. McEnery, & A. Wilson (Eds.), *Proceedings of the Teaching and Language Corpora* , UCREL Technical Papers 9: 108-126. 1996.

desses parâmetros é a frequência de palavras. Muitas palavras não são similarmente distribuídas entre diferentes registros, mas ocorrem mais frequentemente em um ou outro registro. Biber et al. (1999, p. 376) [15] mostram que o verbo lexical *get* é o mais frequente em conversações, mas bastante infrequente nos registros escritos (ficção, notícias, prosa acadêmica). Por outro lado, *make* é o verbo lexical mais frequente na prosa acadêmica, e apenas o décimo primeiro na conversação (ibid, p. 375). Características gramaticais são, da mesma forma, distribuídas diferentemente entre registros. Interrogativas, por exemplo, "são 47 vezes mais frequentes em conversação do que em prosa acadêmica ou notícias, mas apenas quatro vezes mais frequentes do que em ficção" (BIBER et al., 1999, p. 211).

Estudos da tradução

Os estudos da tradução envolvem o uso de corpora comparável ou paralelo. Esses estudos são de dois tipos: teóricos e práticos. Os estudos teóricos visam ao estudo dos processos tradutórios explorando como uma idéia em uma língua é transmitida em outra língua e através da comparação das características linguísticas e suas frequências em textos traduzidos e em textos originais. Na abordagem prática, os corpora fornecem um banco de dados para o treinamento de tradutores e uma base para o desenvolvimento de aplicações como tradução por máquina, ou seja, uma interface com a linguística computacional.

Ensino e aprendizagem de línguas

Parece haver um crescente interesse na aplicação de pesquisas baseadas em corpus no ensino de línguas. Essa aplicação pode ser de duas formas: o uso direto de corpora com os aprendizes e o uso indireto. No uso direto de corpora em aula os alunos agem como "detetives linguísticos" (JOHNS, 1997 p. 101), descobrindo fatos sobre a língua que estão estudando através de exemplos autênticos. Johns denominou esse tipo de metodologia de ensino de *Data Driven Learning* (DDL). Corpora podem também ser usados de forma indireta, através da elaboração de materiais baseados em linhas de concordância.

Além desses usos mais diretos, os corpora vêm sendo cada vez mais usados na elaboração de materiais didáticos. A LdC pode oferecer informações relacionadas a vocabulário, gramática, formalidade e informalidade, diferenças entre a linguagem escrita e falada, como as pessoas começam e terminam uma conversa, entre outros aspectos. Desta forma,

---

[15] Biber et al. (1999) consideram quatro abrangentes registros em sua gramática baseada no *Longman Spoken and Writen English Corpus (LSWEC)*: notícias, prosa acadêmica e ficção (todos registros escritos) e conversação (naturalmente, registro oral).

estudos baseados em corpora podem sugerir os itens linguísticos e processos que serão mais provavelmente encontrados por usuários de uma língua e que, portanto, merecem mais investimento em termos de tempo.

Para a utilização de um corpus na elaboração de material didático (livros, polígrafos ou exercícios) é necessário, primeiramente, decidir (no caso da língua inglesa) quanto ao tipo e variedade de inglês que servirá como base para a elaboração do material, uma vez que corpora diferentes apresentarão palavras diferentes e, frequentemente, diferentes usos e funções das palavras a serem ensinadas. A palavra *nice*, por exemplo, é uma das quinze palavras mais frequentes no inglês falado (McCARTEN, 2007). Entretanto, ela é bastante rara no inglês acadêmico escrito, ocorrendo sempre em citações de literatura ou em entrevistas. Portanto, a escolha (ou a compilação) de um corpus pode afetar as palavras a serem incluídas nos materiais didáticos, assim como seus sentidos e usos.

Além dessas áreas, McEnery et al. (2006) mencionam aproximações da LdC com os estudos linguísticos diacrônicos, a pragmática, a semântica, a sociolinguística, a análise do discurso critica, a estilística e os estudos literários, e a linguística forênsica. Como visto, a LdC pode auxiliar na maioria da áreas da linguística.

LIMITAÇÕES

Como a maioria das áreas, os estudos baseados em corpora têm algumas limitações. Primeiramente, um corpus não consegue informar se algum fenômeno linguístico é possível ou não, apenas se é frequente ou não. Por um lado, as descrições linguísticas (especialmente da língua inglesa) estão cada vez mais concentradas no que é típico, distanciando-se das noções de boa formação, ou correção (foco das pesquisas racionalistas) (SINCLAIR, 1991, p. 17). Entretanto, a pergunta "É possível dizer isso?" ainda necessita ser respondida. Para Hunston (2002), a intuição do falante nativo ainda é a melhor maneira de responder essa pergunta.

Um corpus não consegue mostrar nada mais além de seu conteúdo. Por mais representativo que um corpus proponha-se a ser, generalizações feitas a partir de resultados de um corpus são, na verdade, extrapolações. Uma declaração sobre um corpus é uma declaração sobre aquele corpus, e não sobre a linguagem ou registro o qual o corpus representa. Dessa forma, conclusões a respeito da linguagem inferidas a partir de um corpus devem ser tratadas como deduções, não como fatos.

Um corpus pode oferecer evidências, mas não pode fornecer informações. Por exemplo, o que *something of a* significa antes de um substantivo, em expressões do tipo *something of a surprise*? Presume-se ser um "mitigador", *something of a surprise* é uma *small surprise* (pequena surpresa). Ou seja, um corpus apenas fornece uma abundância de exemplos ao

pesquisador, mas apenas o pesquisador pode interpretá-los (HUNSTON, 2002, p. 23).

Finalmente, e, conforme Hunston (2002), a falha mais grave do uso de um corpus é que ele apresenta a língua fora de seu contexto natural. Por exemplo, quando os textos estudados possuem ilustrações, devido às limitações da tecnologia disponível, elas devem ser descartadas. Em outras palavras, transcrições de dados orais não conseguem representar fielmente todas as informações sobre entonação, linguagem corporal e outras características paralinguísticas. Esse fato aponta para a necessidade de um corpus ser apenas uma das ferramentas, entre outras, em um estudo linguístico.

VANTAGENS

Apesar das limitações da LdC, acredito que seja a metodologia disponível mais indicada para a averiguação de dados reais sobre a língua, uma vez que:

-A LdC constitui um método rigoroso para a obtenção de dados atestados da língua "*in vivo*" em que é possível acessar um conjunto de dados reais e ricos no sentido de que, se o corpus for representativo de uma certa porção de linguagem, aparecerão, de forma clara, as unidades de comunicação mais utilizadas e as menos utilizadas. Além disso, pode-se acessar seus padrões semânticos, as associações que as palavras estabelecem entre si, suas colocações, as variações das unidades lexicais, entre outras características.

-Os corpora "simplificaram" a vida dos linguistas. Por exemplo, um linguista que deseje verificar

o uso dos verbos modais, pode facilmente reunir todos esses verbos modais em um só lugar para a observação. O ato de reunir evidências é simplificado, liberando os esforços do pesquisador para o ato interpretativo. (HUNSTON, 2002, p. 214).

-Os corpora mostraram que a língua é padronizada de uma forma muito mais detalhada do que sugerido anteriormente. Regras tidas como gerais, geralmente podem ser aplicadas somente em certos contextos. Como resultado, novas idéias sobre língua emergem e velhas idéias podem necessitar reavaliação.

-Afirmações mais objetivas podem ser feitas tendo em vista observações baseadas em corpora quando comparadas a observações introspectivas. Falantes nativos podem saber uma língua perfeitamente, mas nem sempre sabem o que eles dizem ou como o fazem. Da mesma forma, há uma discrepância entre o sentido intuitivamente priorizado e o mais frequente.

-Os corpora provêm a possibilidade da "prestação de contas total"

(*total accountability*) das características linguísticas e não apenas de traços salientes individuais (*individual salience*).

-O estudo das colocações pode ajudar a organizar o contexto em padrões principais. Pode-se utilizar o conhecimento desses padrões para acessar o comportamento da língua ou os usos de palavras específicas no texto. Isso pode facilitar a diferenciação entre os significados de uma única palavra ou ainda determinar a variação de características sintáticas.

Foram descritos neste artigo vários aspectos relacionados à LdC, tais como seu surgimento, suas principais características, suas principais aplicações, suas limitações e, por fim, suas vantagens. Por ser uma metodologia relativamente nova, a LdC ainda suscita inquietudes, dúvidas e questionamentos. Entretanto, como visto neste artigo, a LdC abre um novo leque de possibilidades para os estudos linguísticos e para a forma como a língua tem sido até hoje entendida.

REFERÊNCIAS

ALTENBERG, B. *On the Phraseology of Spoken English: The Evidence of Recurrent Word-Combinations. Phraselogy*. Ed. A.P.Cowie. Oxford: Clarendon Press. p. 101, 1998.

BERBER SARDINHA, A.P. *Linguística de Corpus: histórico e Problemática*. D.E.L.T.A., Vol.16 N 2 :323-367, 2000.

BERBER SARDINHA, A.P. *Linguística de Corpus*. Barueri: Manole, 2004.

BEVILACQUA, C. R. *Unidades Fraseológicas Especializadas: Novas Perspectivas para sua Identificação e tratamento.* In: KRIEGER, M. G.; BECKER, A. M. M. *Temas de terminologia.* São Paulo: FFCH/USP, 2001.

BIBER, D., CONRAD, S. e REPPEN, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

BIBER, D.; JOHANSSON, S.; LEECH, G.; CONRAD, S.; FINEGAN, E. *Longman Grammar of Spoken and Written English*. London: Longman, 1999.

CARTER, R.; McCARTY, M. *Grammar and the Spoken Language.* Applied Linguistics, 16, 141-158, 1995.

FRANCIS, W.N. e KUCERA, H. *Brown Corpus Manual: MANUAL OF INFORMATION to accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers, 1964*. Internet: <http://khnt.hit.uib.no/icame/manuals/brown/INDEX.HTM > Acessado em 10/2/2008.

GÓMEZ, R. *Variability and Detection of Invariant Structure*. Psychological Sciences, Vol.3. No 5, 431-436, 2002.

HOFFMANN, L. Possibilidades de aplicação e a aplicação atual de métodos estatísticos na pesquisa de linguagens especializadas (Título Original: Anwendungsmöglichkeiten und bisherige Anwengung von Statistischen Methoden in der Fachsprachenforschung, 1998). Disponível em: *Cadernos de Tradução*, Porto Alegre, nº 20, janeiro-junho, p. 61-76,

2007

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

JOHNS, T. Contexts: the background, development and trialling of a concordance-based CALL program in Wichmann, Fligelstone, McEnery and Knowles (eds.), *Teaching and Language Corpora*. London: Longman, 1997. 100-115, 1997.

KENNEDY, G. D. *An introduction to corpus linguistics*. Nova York : Longman, 1998.

LEECH, G. N. *The State of Art in Corpus Linguistics.* London: Longman, 1991.

LEECH, G. N. *Corpora and Theories of Linguistic Performance.* Berlin: Mouton de Gruyter, 1992.

MACIEL, A. M. B. . Novos horizontes para o ensino do léxico. *Revista Língua & Literatura*, Frederico Westphalen, v. 6 e 7, p. 123-130, 2005.

McCARTEN. *Teaching Vocabulary-Lessons from the Corpus, Lessons for the Classroom*, CUP: Cambridge 2007.

McENERY, T.; GABRIELATOS, C. *English corpus linguistics*. In B. Aarts & A. McMahon (eds.), *The Handbook of English Linguistics* (pp. 33-71), Oxford: Blackwell, 2006

McENERY, T. e WILSON, A. *Corpus Linguistics.* Edinburgh: Edinburgh University Press, 1996.

McENERY, T; XIAO, R. e TONO, Y. *Corpus-based Language Studies: an advanced resource book*, Oxon: Routledge, 2006

OTHERO, G. (2006), Linguística Computacional: uma breve introdução. *Letras de Hoje,* Vol 41, N.2. Porto Alegre: EDIPUCRS, 2006.

QUIRK, R.; GREENBAUM, S.; LEECH, G.; SVARTIK, J. *A Comprehensive Grammar of the English Language*. London: Longman, 1985.

RAYSON, P. *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison*. Tese de doutorado. Universidade de Lancaster, 2002.

ROBERTSON, F. A. *Airspeak: Radiotelephony Communication for Pilots*. Oxford: Prentice Hall, 1988.

SARMENTO, S. *O uso dos verbos modais em manuais de aviação em inglês: Um estudo baseado em corpus***.** Tese de doutorado. UFRGS: Porto Alegre, 2008.

SCOTT, M. *WordSmith Tools.* (1996) Oxford: Oxford University Press. Versão 5, 2008.

SINCLAIR, J. *Corpus, Concordance, Collocation.* Oxford: OUP, 1991.

SINCLAIR, J. *Paper Presented at XI Encontro da Associação Portuguesa de Linguística*. Lisboa, 1995.

SINCLAIR, J. *A Way With Words.* In H. Hasselgard and S. Oksefjell (eds.). *Out of Corpora:* Studies in Honor of Stig Johansson. Amsterdan: Rodopi, 1999.

STUBBS, M. *Corpus and Text Analysis.* Oxford: Blackwell, 1996.

STUBBS, M. *Words and Phrases.* Oxford: Blackwell, 2001.

ANEXO

Websites relacionados ao uso de corpora:

http://corpus.byu.edu/bnc/

http://davies-linguistics.byu.edu/personal/

http://devoted.to/corpora

http://www.edict.com.hk/concordance/

http://www.linguateca.pt/

http://www.americancorpus.org/

http://www.scottishcorpus.ac.uk/

http://www.revel.inf.br/ (Ano 2, Número 3)

http://www.hltmag.co.uk/