TRABALHANDO COM O COMPUTADOR NA PESQUISA LINGUÍSTICA: O USO DO MODAL *CAN* POR BRASILEIROS E INGLESES.

Maria Izabel de Andrade Almeida[1]

RESUMO: Este estudo compara a utilização do modal can em inglês como língua materna (L1) e como língua estrangeira (L2). O trabalho utiliza os princípios e métodos da Lingüística de Corpus ao contrastar e comparar o uso do modal em dois corpora digitalizados contendo composições argumentativas escritas por vestibulandos ingleses e aprendizes brasileiros de inglês de nível avançado. O objetivo especifico do estudo é verificar como é usado o modal can na segunda língua e na língua materna em termos de freqüência, ou se seja, se é subusado ou se usado em excesso. Visto que o modal em tela pode expressar modalidade deôntica (habilidade) ou epistêmica, (probabilidade), o estudo se propõe também a verificar escolha semântica de cada um dos dois grupos de usuários, dentro do discurso argumentativo. Por fim, o estudo aponta para algumas possíveis conclusões pedagógicas sobre o ensino de modais em língua inglesa..

1) Introdução

O estudo aqui desenvolvido se alinha a outros estudos sobre o inglês de aprendizes brasileiros, sob o prisma da Lingüística de Corpus (doravante LC), realizados na UERJ e na PUCSP, como parte do Projeto Br-ICLE (*Brazilian International Corpus of Learner English*). Esses estudos, por sua vez, se somam ao Projeto ICLE, que congrega pesquisas sobre o inglês de aprendizes cujas línguas maternas são as línguas da União Européia (GRANGER, 1998 e 2002; RINGBOM, 1998).

A pesquisa faz parte de um projeto de Iniciação Cientifica, cujo foco é a utilização de corpora eletrônicos contendo composições de aprendizes brasileiros para estudar o recurso da modalidade em língua inglesa. A ênfase do presente artigo é tão somente no modal *can.* O estudo é de base contrastiva, isto é a utilização do modal *can,* é comparada em dois bancos de textos digitais contendo composições argumentativas escritas por universitários brasileiros, estudantes de língua inglesa e vestibulandos ingleses. As perguntas a que pretendemos responder ao longo do trabalho têm a ver com a freqüência com que os aprendizes brasileiros de inglês como língua estrangeira (L2) e os usuários de inglês como língua materna (L1) usam o modal *can* em composições argumentativas. Visto que o modal em tela pode expressar modalidade deôntica (habilidade) ou epistêmica, (probabilidade), uma outra pergunta a ser respondida tem a ver com as escolhas de significado feitas pelos usuários de inglês como L1 e de L2. Por fim, o estudo pretende apontar para algumas conclusões pedagógicas.

---

1 Bolsista do CNPq; orientadora Tania Shepherd**.**

O presente trabalho se situa dentro da área da pesquisa lingüística chamada LC, a qual privilegia a análise de dados empíricos para a descrição da linguagem em uso. A LC pode também ser definida como a área de investigação que se ocupa "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" (BERBER SARDINHA, 2004: 3). Em outras palavras, o lingüista de corpus objetiva a descrição de uma determinada língua ou variante lingüística, tal como ela é utilizada por um determinado grupo de enunciadores em um evento comunicativo específico, a partir da análise de um *corpus* especialmente compilado para o objetivo do pesquisador.

A popularização dos estudos na área de LC está intrinsecamente ligada ao desenvolvimento de novas tecnologias, conforme aponta Tognini Bonelli (2001: 5). O microscópio eletrônico permitiu novas abordagens à biologia e à genética. Segundo a autora, da mesma forma, o computador alterou a natureza da investigação lingüística, uma vez que deu acesso à análise, seleção e sistematização de *corpora* contendo milhões de palavras, em curto espaço de tempo, tarefa que era praticamente inviável atrás quando os computadores pessoais e os programas garimpadores de texto não existiam. Na verdade, a análise lingüística a partir de corpus não é invenção do século XX. Entretanto a confiabilidade dessas análises, que eram manuais, era justamente uma das críticas feitas a esses estudos, de acordo com Berber Sardinha (2004: 4).

A LC tem como ponto de partida para a pesquisa científica a *performance*, ou o uso real de uma língua e não a competência, ou o uso idealizado da mesma. Desta forma, o foco adotado em trabalhos que privilegiam a LC como metodologia é a freqüência relativa de itens lingüísticos que tenham ocorrido em contexto natural e não escassos exemplos inventados (KNOWLES, 1997) . Berber Sardinha (2004: xvii) vai mais além, afirmando que "ao revelar uma quantidade surpreendente de evidências lingüísticas provindas de corpora eletrônicos, a Lingüística de Corpus questiona os paradigmas estabelecidos dos estudos lingüísticos" até o início dos anos 80.

Entre os inúmeros corpora já compilados e investigados com auxílio do computador, aqueles contendo trabalhos escritos por aprendizes vêm sendo também objeto de pesquisa. Desde a década de noventa, o foco tem sido especialmente a escrita de aprendizes de línguas estrangeiras, com um maior numero de trabalhos sobre a língua inglesa como língua estrangeira.

Uma das maiores contribuições dos estudos de corpora de aprendizes é que possibili-

tam a investigação de fatos lingüísticos que antes eram difíceis ou impossíveis de investigar (LEECH, 1998 e GRANGER, 1998). Via de regra, as investigações sobre a escrita de aprendizes de língua estrangeira, especialmente os de inglesa, eram baseadas na comparação entre um número restrito de exemplos da produção de alguns poucos aprendizes e exemplos de usuários nativos da língua estrangeira (geralmente exemplos fornecidos pelo próprio analista/pesquisador). Com o advento dos computadores e a possibilidade da compilação de grandes *corpora* eletrônicos, tais investigações aumentaram em número, extensão e escopo e passaram a constituir o que Granger (op.cit.) denominou de ECI (Estudos Contrastivos de Interlíngua), uma área que conjuga a grande área da Aquisição de Linguagem e a Lingüística de Corpus.

Uma outra contribuição dos estudos de *corpora* de aprendizes reside no fato de que esses estudos podem mostrar precisamente de que forma e com que freqüência grupos específicos de aprendizes expressam certos significados numa língua estrangeira, em termos de léxico-gramática. Ou nas palavras de Hyland (2002: 176), como, coletivamente, os *corpora* de aprendizes podem apontar áreas da *performance* que são de caráter nativo e não-nativo.

Os estudos sobre a produção escrita em língua inglesa por aprendizes oriundos de países da União Européia estão em franco desenvolvimento através do Projeto ICLE (International Corpora of Learner English) da Universidade de Louvain na Bélgica. O ICLE é um grande banco eletrônico, composto de composições argumentativas de aprendizes de inglês, em sua maioria universitários, coletadas sob as mesmas circunstâncias e girando em torno das mesmas temáticas. Para cada país da comunidade (com exceção de Portugal, que não participa dp projeto) adota-se uma sigla. PICLE, por exemplo, é o corpus composto de inglês de aprendizes poloneses; SWICLE, de inglês de aprendizes suecos, e daí por diante.

Em contrapartida, a investigação sobre 'inglês brasileiro' através de *corpora* eletrônicos está somente começando. Atualmente os estudos se concentram no Projeto Br-ICLE, subordinado ao ICLE e coordenado pela PUC de São Paulo. Além de seguir os padrões e fundamentos de compilação de *corpora* eletrônicos, temáticas e gênero para a produção escrita de aprendizes de língua inglesa estabelecidos pelo ICLE, o Br-ICLE também faz uso do mesmo *corpus* de referência, o corpus LOCNESS, composto de *essays* digitalizados escritos por universitários americanos e ingleses compilados pela Universidade de Louvain.

Tendo estabelecido o contexto da pesquisa, passamos agora a discorrer sobre a identificação do problema, a descrição do método de coleta e do tratamento dos dados para então enfocar a análise de dados algumas conclusões preliminares pertinentes ao modal *can*.

2) Problema

De acordo com Ringbom (1998), que estuda advérbios como marcadores de atitude no inglês de aprendizes dos países da União Européia, a idéia que se tem desses aprendizes é que eles são obtusos, repetitivos, sem imaginação e freqüentemente verborrágicos. O vocabulário limitado é a principal razão dessa impressão geral.

Uma outra impressão geral é de os aprendizes não sabem como modalizar seu discurso de maneira eficiente, ou seja não parece haver uma voz nos textos de aprendizes que saiba expressar atitude e avaliação de forma variada. Quando expressam atitude e avaliação, os aprendizes deixam a impressão que o fazem de maneira repetida, usando, via de regra, um mesmo repertório de recursos.

De acordo com Downing e Locke (1992: 382), modalidade é uma categoria semântica que cobre noções como possibilidade, probabilidade, necessidade , vontade, obrigação e permissão, além de desejo, dúvida, e noções temporais como usualidade. Modalidade, segundo os mesmos autores, expressa uma relação do enunciador com a realidade. O uso de advérbios é uma das muitas maneiras de o enunciador inscrever essa relação no discurso. O uso de modais é uma outra.

O presente trabalho enfoca tão somente um modal da língua inglesa, o modal *can*, visto que este é o modal usado com mais freqüência pelos aprendizes brasileiros (ver anexo 1) *Can* pode expressar dois significados distintos. O primeiro tem a ver com certeza, possibilidade ou probabilidade e é chamado modalidade epistêmica. Downing e Locke (op.cit: 383) dizem que a modalidade epistêmica expressa conhecimento, ou melhor a falta de conhecimento. A modalidade epistêmica se contrasta com a modalidade não epistêmica, ou deôntica, através da qual podem ser expressas obrigação e permissão. Essas duas modalidades ajudam na execução de duas importantes funções comunicativas que são, respectivamente, comentar e avaliar uma interpretação da realidade e intervir/operar mudanças em acontecimentos.

Assim como o fazem com outros recursos de léxico-gramática, os aprendizes de uma língua (L2) tendem a usar alguns modais com excessiva freqüência *(overuse[2])*, se comparados com a freqüência dos usuários da língua materna. Podem também usar outros modais com parcimônia, *(underuse)*. Podem simplesmente não usar certos modais em determinadas circunstâncias que demandaria o seu uso ou usá-los de forma errada *(misuse)*

A investigação a que nos propomos, com foco exclusivo no modal *can*,tem o objetivo

---

2 Os rótulos *overuse, underuse e misuse* foram cunhados por Granger (1998) para expressar uso excessivo, subuso e uso errôneo (ou infeliz) respectivamente. Não usar é tido por Granger como *undersuse*. Entretanto, parece-nos que há lugar aqui para uma categoria de 'uso zero' (ou zero use).

de verificar , portanto,

1) quais as semelhanças e diferenças no uso do modal *can* por aprendizes brasileiros, em contraste com os usuários de inglês como língua materna?

2) que aspecto semântico (epistêmico ou deôntico) é privilegiado por cada um dos grupos estudados.

4) Metodologia

A metodologia de investigação utilizada neste trabalho é a da LC , cujas origens já delineamos na seção de introdução do presente artigo. Cabe agora falarmos sobre seus princípios e processos.

A escolha da LC como metodologia tem a ver com a natureza do estudo: trata-se de um estudo empírico, cujo ponto de partida é a coleta de dados lingüísticos autênticos que refletem as preferências de uso de determinado grupo. A outra razão se relaciona à utilização de um programa de computador para a análise lingüística (TOGNINI BONELLI, 2001: 2)

A LC pode ser considerada uma metodologia de extração de dados de corpora digitais ou um ramo de investigação lingüística com seus próprios preceitos e conceituação. O ponto de partida para ambos é a freqüência relativa do item lexical (ou de amarrados lexicais) em um corpus como medida da preferência por seu uso. Um outro processo que é ditado meramente pelo emprego de programa de computador na extração de dados é KWIC – *key word in context* (palavra-chave em contexto). Podemos dar um comando para o programa para que liste um determinada palavra de vários textos linha por linha. Cada uma das linhas é chamada de concordância, porque é extraída através de um programa concordanciador. Através das concordâncias é possível ver os ´colocados´ das palavras, isto é os itens ou grupos de itens que tendem a ocorrer com freqüência à direita ou à esquerda da palavra de busca. Assim o adjetivo *fresh* em inglês, que normalmente associamos a *water* e *air* (*water* e *air* seriam colocados de *fresh* à sua direita), pode também ocorrer com as palavras *invigoration, injection of capital, complexion e finance,* como mostram as linhas de concordância abaixo, extraídas aleatoriamente do *British National Corpus* em http://corpus.byu.edu/bnc

for Smith and shows encouraging signs of **fresh** invigoration. But for all
T he company desperately needs a **fresh** injection of capital but its financial
slimly built, with a **fresh** complexion and ginger hair, and should
d that it would need up to £1.6bn in **fresh** finance on top of the £6bn already

Além dos conceitos de colocados, que são preferências semânticas, a LC lida também

com o conceito de coligados, que são as preferências sintáticas, ou seja, quais os lugares sintáticos preferidos de determinado item lexical.

Como a LC lida com freqüências relativas, os lingüistas de corpus costumam trabalhar com um corpus de referência, ou um corpus que serve de parâmetro para a relativização das freqüências de uso. No caso do presente estudo, o corpus de referência é o corpus LOCNESS, um corpus comercial, que consiste de *essays* argumentativos, descritivos e literários escritos por vestibulandos e universitários americanos e ingleses.

Somente uma seção do corpus LOCNESS foi utilizada na presente pesquisa, ou seja, a seção de composições argumentativas escritas por alunos britânicos à época do *A-level* – exame semelhante ao ENEM, Exame Nacional do Ensino Médio – constituindo um total de 60.209 palavras. Tal decisão foi tomada a fim de haver compatibilidade entre o corpus de referência e corpus compilado, totalmente argumentativo.

4.1. Passo a passo

O primeiro passo foi a extração de lista com as 50 palavras mais freqüentes em termos de ocorrência , a partir da qual se verificou que, dentro dos auxiliares modais, o modal *can* é mais usado pelos aprendizes brasileiros

O segundo passo consistiu em extrair tanto no corpus LOCNESS (corpus de inglês como língua materna) quanto no corpus Br-ICLE (corpus de inglês produzido por brasileiros). todas as ocorrências do modal *can*. Entretanto, conforme sugerido por Scott e Tribble (2006), não basta somente olhar o item lexical isoladamente. Por vezes o percentual de uso individual de uma palavra pode ser semelhante em dois corpora, mas ao examinarmos as combinações freqüentes dessa palavra (ou seus n-gramas), vemos que existem diferenças significativas de uso.

Portanto, foram extraídos em seguida bigramas e trigramas do modal *can* em ambos os corpora, ou seja modal *can* acompanhado de uma palavra e modal *can* acompanhado de duas palavras). Todos os dados obtidos foram analisados contrastivamente, após as freqüências terem sido normalizadas, ou seja, equiparadas, estatisticamente, de acordo com o tamanho de cada um dos corpora, após o que procedeu-se à análise.

6) Análise dos dados

As porcentagens abaixo expressam a freqüência do uso do modal *can,* isoladamente, por brasileiros e ingleses. Aparentemente a distribuição do *can* não apresenta diferenças im-

portantes. Em cada cem vezes em que os aprendizes investigados escolhem usar um modal, os alunos brasileiros escolhem usar o modal *can* 37 vezes, e os ingleses o fazem 41 das vezes.

Tabela 1 – Freqüência relativa de *Can* como palavra isolada

BRASILEIROS	INGLESES
37%	41%

Porque essa diferença não parece espelhar verdadeiramente o uso do modal em tela nos textos, foi necessário investigar o uso em termos de semântica. O modal *can* pode ser usado de forma **deôntica**, com o sentido de habilidade ou de forma **epistêmica**, ou seja, expressando a atitude do falante em relação à verdade da proposição através da probabilidade. Ao fazermos a análise e a classificação de cada uma das ocorrências de *can* em ambos os corpora, verificamos que os brasileiros e os ingleses empregam *can* em composições argumentativas de forma distinta, como veremos nos exemplos que se seguem extraídos do corpus Br-ICLE:

(1) What *can* we do?

(O que podemos fazer?)

(2) One *can* easily notice the problem.

(Pode-se facilmente perceber o problema)

Nos casos acima o modal *can* é claramente usado de forma deôntica, para expressar **habilidade**.

Por outro lado, como no exemplo a seguir, extraído do corpus LOCNESS, o modal *can* é um instrumento de expressão epistêmica, ou seja, mostra **probabilidade**.

(3) It affects the nervous system very badly and *can* lead, in some cases, to paralysis.

(Afeta muito o sistema nervoso e pode levar à paralisia)

O que não configurava uma diferença marcante em termos numéricos, quando contabilizado em termos semânticos, mostra um perfil marcadamente distinto. Na tabela abaixo encontram-se as escolhas dos dois grupos estudados. No corpus LOCNESS, para cada três vezes em que o modal *can* é escolhido, duas delas são para a expressão epistêmica. Em contrapartida, no corpus Br-ICLE as escolhas são numericamente semelhantes: ou o modal *can* expressa possibilidade ou habilidade, conforme ilustrado na tabela abaixo.

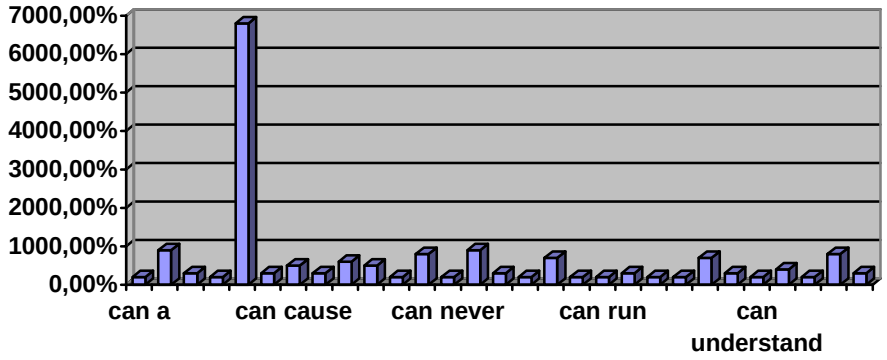
Tabela 2 – Freqüência absoluta e relativa do modal *Can* (epistêmico ou deôntico)

can	*LOCNESS*	*%*	*Br-ICLE*	*%*
epistêmico	*158*	*63*	*43*	*47*
deôntico	*96*	*37*	*50*	*53*
total	*245*		*93*	

Em termos de bigramas, vemos uma preferência nos dois corpora pelo feixe *can be*. Os demais bigramas têm a coligação de *modal + verbo lexical* em ambos os corpora; e modal + advérbio (*can easily*, *can never*). Pode-se ainda elencar as coligações *pronome + can.*

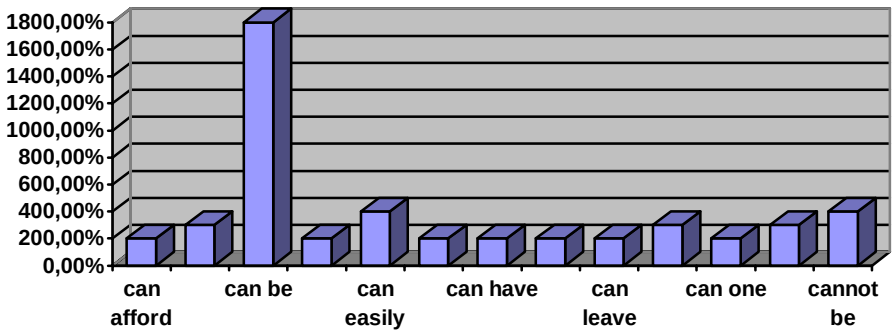
O gráfico 1 mostra as combinações, em termos de bigramas, extraídas do LOCNESS:

Gráfico 1 – Bigramas do modal *Can* (LOCNESS)



O gráfico 2 mostra as combinações feitas no Br-ICLE:com exceção do bigrama *can be*, que é igualmente usado pelos dois grupos (e representado pela barra de maior altura nos gráficos), os demais bigramas do Br-ICLE são em número mais reduzido.

Gráfico 2 – Bigramas do modal *Can* (Br-ICLE)

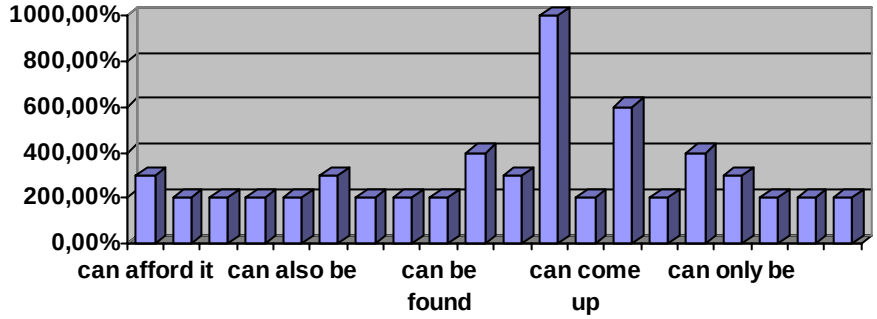


Finalmente passemos aos trigramas. Os trigramas são conjuntos de três palavras que

ocorrem juntas na língua e que mostram a fraseologia do discurso, ou seja, suas expressões típicas (Scott e Tribble, 2006). No caso da presente trabalho, os quadros de trigramas espelham a fraseologia possível do discurso argumentativo, da forma que é moldado por escritores experientes ou aprendizes.

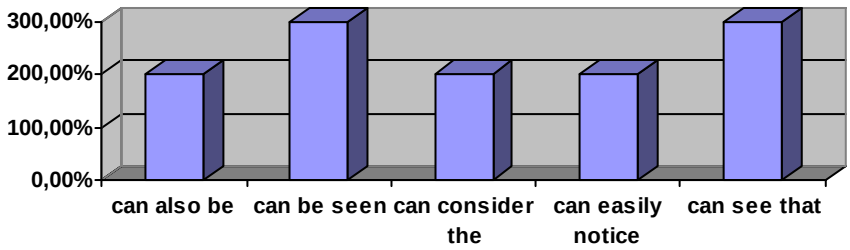
O gráfico 3 mostra os trigramas formados com o modal *can* no corpus LOCNESS; nota-se aí uma maior variedade de combinações

Gráfico 3 – Trigramas do modal *Can* (LOCNESS)



A seguir , no gráfico 4, vemos as combinações feitas por brasileiros, que configuram um leque de .escolhas consideravelmente menos do que no Gráfico 3 acima.

Gráfico 4 – Trigramas do modal *Can* (Br-ICLE)



Além disso, observamos que a forma negativa de *can*, extraída como trigrama no LOCNESS, não aparece como trigrama (ou seja, como três colocados freqüentes no corpus Br-ICLE) - se a fraseologia é um indicativo da natureza do discurso, então os trigramas extraídos dos textos dos brasileiros nos indicam que a fraseologia argumentativa não está sendo

usada da forma e com a freqüência desejada.

7) Conclusões Preliminares

As conclusões preliminares com relação ao modal *can* a partir dos dois corpora estudados são as seguintes:

1. O *can* é escolhido tanto com sentido deôntico como epistêmico no Br-ICLE - em contraste com o corpus LOCNESS, no qual os ingleses mostram preferência pelo *can* epistêmico (probabilidade).
2. Os colocados de *can* são em número reduzido no Br-ICLE - possivelmente por causa das poucas alternativas que o aprendiz conhece no que se refere a sua "competência colocacional"
3. A fraseologia do modal *can*, típica do LOCNESS, não faz parte do discurso argumentativo dos brasileiros estudados. Só há dois trigramas do *can* que são coincidentes (*can* also be/*can* be seen).

É importante ressaltar que a LC está em seu início, principalmente no que tange ao estudo de corpora de aprendizes brasileiros. Conclui-se que outros modais devam ser estudados e que o corpus do Br-ICLE seja aumentado. Em termos pedagógicos, a importância deste estudo está relacionada com o contato do aprendiz com o inglês em produções daqueles que utilizam o inglês como língua materna, possibilitando assim, que o aprendiz tenha uma visão da língua inglesa em seu uso real, diferente da forma tradicional como ela costuma ser abordada nas aulas de inglês como língua estrangeira. O acesso dos aprendizes a bancos de dados como os expostos nesse trabalho poderiam tornar-se uma importante ferramenta no processo de aprendizagem desses alunos.

Referências bibliográficas

BERBER SARDINHA, T. A. *Lingüística de Corpus*. São Paulo: Manole, 2004.

DOWNING, A e LOCKE, P. *A University Course in English Grammar.* Hemmel Hempstead: Prentice Hall, 1992.

GRANGER, S**. "**A Bird's-eye View of Computer Learner Corpus Research". In Granger S., Hung J. and Petch-Tyson S. (eds) *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Language Learning and Language Teaching 6. Amsterdam & Philadelphia: Benjamins, 2002, pp. 3-33.

GRANGER, S. "The computer learner corpus: a versatile new source of data for SLA research". In Granger, S. (org.) *Learner English on Computer*. London: Longman, 1998.

HYLAND, K. *Teaching and researching: Writing*. Harlow: Longman, 2002.

LEECH, G. “Teaching and Language Corpora: a convergence”. In Wichmann, A et.al. (orgs.) *Teaching and Language Corpora*. Harlow: Addison Wesley, 1997.

LEECH, G. Learner corpora: what they are and what can be done with them. In Granger S. (ed.) *Learner English on Computer*. London: Longman, 1998, xiv-xx.

RINGBOM, H**.** (1998) “Vocabulary frequencies in advanced learner English: A cross-linguistic approach”. In Granger S. (ed.) *Learner English on Computer*. London Longman, 41-52.

SCOTT, M. *Wordsmith Tools*. Oxford: OUP, 1999.

SCOTT, M. e TRIBBLE, C. *Textual Patterns.* Amsterdam : John Benjamins, 2006.

SINCLAIR, J. *Reading Concordances*. London: Longman, 2003.

TOGNINI BONELLI, E. *Corpus Linguistics at Work*. Amsterdam: John Benjamins, 2001.