Corpora de aprendiz de língua estrangeira: um estudo contrastivo de n-gramas

Tania M G. Shepherd (UERJ – Faperj /CNPq)

RESUMO: O presente trabalho está na interface da Linguistica de Corpus e os textos produzidos por aprendizes. Primeiramente, problematiza-se a pesquisa lingüística sobre corpus de aprendiz, relacionando-a a uma discussão sobre a natureza do erro em escrita em língua estrangeira. Em seguida, o trabalho dá um exemplo prático de abordagem indutiva a dois corpora, compilados em condições semelhantes, os quais contêm textos em inglês como línguas materna e estrangeira, respectivamente. É mostrado um passo a passo de uma abordagem indutiva, como modo de lidar com a análise de n-gramas, ou agrupamentos lexicais. Por fim, o trabalho foca na produção de materiais pedagógicos a partir de corpora de aprendizes no Brasil.

Palavras-chave: Linguística de Corpus; abordagem dirigida pelo corpus; corpora de aprendiz; n-gramas

Introdução

O presente estudo contrasta e compara as escolhas lexicais em inglês feitas coletivamente em dois corpora eletrônicos. Utilizando o ferramental e a conceituação do léxico que embasa a Linguística de Corpus, o trabalho tem como objetivo contrastar um corpus escrito de aprendiz de língua estrangeira que consiste de redações de alunos universitários, aprendizes de inglês de nível avançado com um segundo corpus, que consiste de *essays* de universitários e vestibulandos cuja língua materna é o inglês. Ambos os corpora se encaixam na definição de Scott e Tribble (2006: 133) para corpora de aprendiz, ou seja, “uma compilação de textos escritos, não publicados, produzidos num ambiente de ensino ou treinamento, geralmente para serem avaliados”.

Para atingir este objetivo, o artigo apresenta primeiramente um breve histórico da pesquisa de textos eletrônicos produzidos por aprendizes de língua estrangeira, área da Linguística de Corpus também chamada de pesquisa em ‘corpora de aprendiz’. Logo em seguida, discute a relação entre texto de aprendiz e ´erro´ dentro da pesquisa sobre corpus de aprendiz de língua estrangeira, para depois discutir as visões de linguagem que perpassam a Linguística de

VEREDAS ON-LINE – LINGUÍSTICA DE CORPUS E COMPUTACIONAL – 2/2009, P. 100-116 – PPG LINGUÍSTICA/UFJF – JUIZ DE FOR A – ISSN 1982-2243

Corpus em geral e a pesquisa sobre corpora de aprendiz em particular. A análise dos dados se fixa em unidades que são maiores do que a palavra, visto que parte da premissa de Sinclair (2004: 148) de que "a palavra não é o melhor ponto de partida para a descrição de significado, porque o significado é o resultado de combinações de palavras" [1]. Na parte final do artigo, apresentam-se os resultados da pesquisa lexical de unidades maiores do que a palavra, bem como uma tentativa de apontar caminhos dentro dessa seara promissora de investigação que é o corpus de aprendiz de língua estrangeira.

1. Justificativa

Datam da década de sessenta, o uso de computadores na pesquisa linguística, a compilação do primeiro corpus eletrônico e a avaliação de parâmetros estatísticos para a comparação de dados linguísticos. Com a recente publicação do relatório sobre a pesquisa lexical realizada entre 1967 e 1969 para o *Office for Scientific and Technical Information* da Grã-Bretanha, ou relatório OSTI, (KRISHNAMURTHY, 2004), jamais divulgado anteriormente, sabe-se também que data dessa época a formalização pioneira de critérios para a investigação de relações sintagmáticas entre itens lexicais. Essa investigação concretizaria o que, na década anterior, Firth (1957) já dizia sobre a co-ocorrência de dois itens lexicais, ou seja, que essa co-ocorrência poderia ser prevista por critérios outros que não suas classes gramaticais ou regras sintáticas.

Desde então, muitos têm sido os trabalhos sobre o léxico, a partir do que se convencionou chamar de Linguística de Corpus – o estudo de textos eletrônicos com o auxilio de computador. Inúmeros são também os corpora formados de textos eletrônicos disponíveis tanto comercialmente quanto gratuitamente[2].

Enquanto os estudos descritivos sobre o léxico das mais diversas línguas já têm uma história e pesquisa de vulto, os corpora contendo textos digitalizados produzidos por aprendizes de língua estrangeira começaram a ser explorados somente por volta da década de noventa (HYLAND, 2002: 176). Esses corpora nos fornecem evidência dos recursos léxico-gramaticais e discursivos utilizados por grupos de falantes nativos de uma língua A, ao se expressarem numa determinada língua B. Dada a facilidade e rapidez de poder vasculhar enormes quantidades de dados linguísticos autênticos, pode-se examinar com segurança e confiabilidade, o que é usado de forma coletiva por esses grupos de usuários, o que é usado em demasia e o que é pouco ou nada usado.

Essa fonte de investigação para os linguistas aplicados tem uma interface com a *Análise de Erros* (AE). Entretanto, nem os dados da AE seguiam princípios de compilação e desenho rigoroso, sendo invariavelmente de tamanho reduzido, nem a pesquisa da AE incluía outros itens além dos "erros", sendo a *Análise de Erros* uma área altamente prescritiva. Além dessa prescrição inerente à *Análise de Erros*, outra diferença entre essa área e as pesquisas em corpora computadorizados de aprendiz é que uma vez identificados os erros´ e analisados fora de seus contextos originais (cf. GRANGER, 1998a: 6), os textos da AE eram descartados.

O que acontece hoje é que o mesmo corpus, compilado e armazenado sob critérios pré-estabelecidos, pode ser fonte inesgotável para análise do perfil coletivo da linguagem utilizada

[1] Esta e as demais traduções de citações originariamente feitas em língua inglesa são de minha responsabilidade.
[2] Como exemplo podem ser citadas três iniciativas de Berber Sardinha: o Banco de Inglês, com 193 milhões de palavras, o Banco de Português, com 750 milhões de palavras, e o Corpus Brasileiro (em construção) com um bilhão de palavras (vide http://www2.lael.pucsp.br/corpora para maiores informações).

por determinado grupo de usuários-aprendizes, durante um determinado período de sua aprendizagem da língua estrangeira.
Além desse benefício analítico imediato, os corpora de aprendiz oferecem outro ponto positivo: o potencial de integração entre a pesquisa e a prática pedagógica. Como diz Leech (1998:xiv) acerca do inglês como língua estrangeira:

> Suponhamos que uma professora X, num país que não tenha inglês como primeira língua, ensine inglês a seus alunos todas as semanas, e de vez em quando lhes peça para escrever composições ou outros trabalhos naquela língua. Ora, ao invés de devolver os trabalhos aos alunos com comentários e um suspiro de alívio, ela guarde as composições em seu computador e construa, gradualmente, semana após semana, uma coletânea maior e mais representativa dos trabalhos desses alunos. Ajudada por ferramentas computacionais como um concordanciador, ela poderá extrair dados e informações sobre as frequências lexicais desse ´corpus´ e poderá analisar o progresso de seus alunos enquanto grupo, com alguma profundidade. As questões de pesquisa que se abrem são mais significativas quando se compila um corpus. (minha tradução)

Trabalhar com corpora de textos eletrônicos produzidos por aprendizes permite ao pesquisador, portanto, partir de dados reais para identificar áreas possivelmente nevrálgicas no uso da língua estrangeira. Permite também ao professor fazer um raio-x da 'performance' de seus aprendizes.

2. O corpus eletrônico de aprendiz e novas respostas para velhas perguntas

Trabalhar com um corpus de estudo em formato eletrônico, tanto de natureza oral quanto escrita, e quantificar as ocorrências e co-ocorrências lexicais, pressupõe uma comparação com algum outro corpus. Estudar corpus de aprendiz pressupõe a utilização de um modelo ´ideal´ de comparação. Por outro lado, essas comparações com modelos ideais de uso da língua estrangeira partem do pressuposto de que a linguagem usada pelo aprendiz fica aquém do desejado, ou seja contem erros ou infelicidades[3]. Na realidade, Barlow (2005: 335) admite que uma das motivações que levou à compilação e estudo de corpora de aprendiz adveio da tradição da área de Análise de Erros em identificar, descrever e explicar erros. Entretanto, uma vez identificados os erros, a área de AE pouco fazia, além de proscrevê-los.
A história de correção de erros sempre oscilou entre duas posições diametralmente opostas – corrigir ou não corrigir - tudo ou seletivamente. Tais posições eram ditadas ou pela abordagem metodológica da época (KNOBLAUCH AND BRANNON, 1984:118) ou mais recentemente como resultado do que veio a ser chamado de direitos humanos linguísticos, ou seja o direito que os falantes não nativos têm às suas 'peculiaridades' linguísticas (AMMON, 1998: 278-282).
O estudo de erros tinha, via de regra, os erros individuais como foco (FERRIS, 2004:3). Além disso, os erros eram estudados fora do contexto e do cotexto originais, sendo que não se prestava atenção às formas certas. Quanto à avaliação da seriedade de um erro, parecia não haver consenso sobre o que seria um erro importante em língua estrangeira. Os resultados de várias pesquisas empíricas (HUGHES E LASCARATOU, 1982; MCCRETTON E RIDER, 1993; HYLAND E ANAN, 2006, entre outros) sugerem que mesmo a percepção do que seja

[3] O termo ´infelicity' é empregado por Granger e colaboradores ao categorizar erros de base lexical.

um erro sério dependeria da língua materna daquele que avalia o erro. Por exemplo, erros em inglês como língua estrangeira são corrigidos com maior tolerância se aquele que corrige tem o inglês como língua materna.
A ênfase dos estudos de corpora eletrônicos no léxico, sua frequência e sua co-ocorrência mudou o foco do estudo sobre a produção de aprendizes do individual para o coletivo, do erro atribuído a uma possível ´interferência´ da língua materna, para uso real da língua, uso esse circunscrito a textos e gêneros textuais produzidos em condições semelhantes, além de, em contrapartida, focar nos acertos, também.
E por que avaliar o uso coletivo a partir de corpora de aprendiz? Existem posições consagradas sobre a língua como marcador de identidade, seja ela étnica ou racial. Existem também argumentos em favor do uso da linguagem como marcador de perfil coletivo de um grupo. Fairclough, (2003:15), por exemplo, acredita que, em suas interações sociais, os seres humanos compartilham modos de falar que os identificam como grupo, isto é, o compartilhamento do ´como´ pode identificar um grupo coletivamente. Além dessa posição, há também aqueles que, como Telya et al. (1998) entendem que qualquer análise do coletivo na linguagem implica a análise do léxico empregado em interações sociais: desde unidades lexicais simples ou de unidades compostas por múltiplos itens. Em outras palavras, a descrição de unidades lexicais recorrentes pode levar a identificar os ´modos preferidos' de um grupo de dizer as coisas. Desta forma, entende-se que analisar corpora de aprendiz através da abordagem e preceitos da Linguística de Corpus implica necessariamente a análise de unidades lexicais usadas com menor ou maior frequência por esse grupo.

3. Visões de linguagem na Linguística de Corpus: o repetido e o novo

Segundo Tognini-Bonelli (2001), há dois modos consagrados de abordagem de corpora eletrônicos em geral: a abordagem baseada em corpus (*corpus-based*) e a abordagem dirigida pelo corpus (*corpus-driven*). A abordagem baseada em corpus vem sendo adotada para analisar uma gama de corpora de textos escritos, que vão desde os textos publicados (textos de jornal, revistas acadêmicas, entre outros) até a escrita de aprendizes em diferentes níveis de proficiência. Essa é na realidade uma metodologia que se aproveita do corpus, principalmente para expor ou testar hipóteses e exemplificar teorias e descrições linguísticas pré-existentes. O corpus pode ser anotado automaticamente em termos de classes gramaticais, entre outros tipos de anotação, ou pode ser usado em sua forma ´crua´, sem anotação.
Em um trabalho, cujo objetivo é verificar como o modal *can* em língua inglesa é usado em corpus de aprendiz, por exemplo, pode-se seguir os seguintes passos. Primeiro extraem-se do corpus todas as instâncias de uso do modal em questão. Em seguida rotulam-se as ocorrências de acordo com um sistema previamente escolhido, que pode basear-se, por exemplo, nas funções epistêmica e deôntica, amplamente descritas pelas gramáticas. Executam-se os mesmos passos em um corpus de não aprendizes. Ao fim do processo, podem-se comparar-se as preferências de uso dos dois grupos de sujeitos investigados (ALMEIDA, 2007), tendo sempre presente o gênero textual escolhido para a compilação do corpus e o nível de proficiência na língua estrangeira dos aprendizes.
A produção teórica sobre corpora de aprendiz, principalmente em inglês como língua estrangeira, seguindo essa abordagem 'baseada em corpus' é de vulto (cf. Granger, 1998c e Granger et.al., 2002). Há os estudos de Aijmer (2002) sobre a modalidade em textos em inglês de alunos suecos; os estudos de Ringbom (1998) sobre os advérbios intensificadores; a pesquisa de Altenberg (2002) sobre a forma causativa ´make´ e mais recentemente um estudo sobre substantivos/marcadores discursivos de Flowerdew: (2005). A tônica desses estudos,

segundo Granger (2002: 12) é fazer comparações entre a linguagem de nativos e não nativos, ou entre a 'norma' e a não norma para deixar em evidência tudo aquilo que confere estranheza ao não nativo, incluindo-se aí os erros, o uso em excesso e econômico de palavras, expressões (o léxico) e estruturas (a sintaxe). Se a investigação é detalhada pode-se ainda, segundo a mesma autora, "entender o sistema subjacente à linguagem do aprendiz e ao mesmo tempo, ou em seguida, comparar sua interlíngua com as normas de uso nativo para avaliar a extensão do desvio".

Granger (2002: 12) admite que sofre críticas ao estudar interlíngua comparando os ´desvios´ do aprendiz e as ´normas´ do nativo, porque alguns linguistas entendem que a interlíngua deveria ser estudada em separado, e não como algo deficiente se comparado a ´normas' nativas.

Por outro lado, a abordagem dirigida pelo corpus, a segunda das abordagens rotuladas por Tognini-Bonelli (2001), se deve, segundo Sinclair (2004: xviii) à ausência de uma teoria que desse conta do léxico, na gênese dos trabalhos com corpora eletrônicos. Como explica ainda Sinclair, nos anos sessenta trabalhava-se com o léxico em termos de ***significado***, não havendo teoria que explicasse a ocorrência, frequência e padronização lexicais, além da preferência de certas palavras por outras (e também da rejeição).

A abordagem ditada pelo corpus, portanto, visa à observação de padrões e frequências lexicais. A observação leva à hipótese, que pode levar à generalização. Em outras palavras, os dados obtidos dos corpora podem ser usados para a formulação de descrições léxico-gramaticais.

Os trabalhos que têm o ponto de entrada dirigido pelo corpus propriamente dito se concentram, em sua grande maioria, em corpora de não aprendizes e de usuários da língua materna. Ainda assim, em 2001, Stubbs afirmava que, entretanto, eram poucos os trabalhos que olhavam para o corpus eletrônico em termos de extração e análise de grupos polilexicais ou n-gramas.

A investigação de unidades formadas por vários itens lexicais pode enfocar blocos relativamente fixos ou blocos cujos componentes podem variar. A pesquisa desses dois tipos de bloco lexical em língua inglesa conta com bibliografia extensa a partir de corpora eletrônicos ou não.

Se os blocos são relativamente fixos, a terminologia de referência a essas sequências pode incluir ´formulas´, ´rotinas´, padrões 'pré-fabricados' (*prefabs*, GRANGER, 1998b*)*, 'phrasicon' (DE COCK et al. 1998), 'lexemas frasais' (MOON, 1998), 'enquadramentos colocacionais' (RENOUF & SINCLAIR, 1991), refletindo-se em cada estudo o modo de ver esses aglomerados como blocos composicionais que oferecem pouca ou nenhuma escolha linguística ao falante. (cf. ELLIS, 1994)[4]. Se os blocos contêm elementos que são corpus-dependentes, podem ser chamados de 'n-gramas' (SINCLAI, 2004), 'agrupamentos´(*clusters)*, pacotes ou feixes lexicais (BIBER, 2004 e BIBER, CONRAD & CORTES, 2004). Segundo Scott e Tribble (2006: 131), um agrupamento lexical (ou n-grama ou feixe lexical) nada mais é do que um produto artificial oriundo de programas extratores. Na verdade, segundo esses autores, o agrupamento lexical existe com base em critérios puramente distributivos, ou seja, dada uma combinação de dois, três ou quatro itens lexicais, se essa combinação ocorrer em um número mínimo de vezes dentro de um texto ou coletânea de textos, ela configurará um 'agrupamento' ou 'feixe lexical'.

[4] Há também na literatura menção a amálgamas, 'chunks' automatizados, clichés, construções coordenadas, colocados, lexemas complexos, compósitos, formas convencionalizadas, expressões fixas, expressões idiomáticas, linguagem formuláica, linguagem fossilizada, frases congeladas, *gestalt*, holística, holófrases, frases lexicalizadas, itens multi-palavras, aglomerados lexicais não analisáveis (cf. Wray, 1999 e 2002).

Além de falta de consenso com relação à nomenclatura, os vários estudos citados não se afinam com relação ao número de itens lexicais que devem fazer parte das sequências estudadas, ou com relação aos aspectos que devam ser analisados: forma, função ou ambos.
Apesar dessa discrepância aparente, todos os estudos citados se baseiam na crença (ainda que tácita) de que os usuários de uma língua, em sua forma escrita ou falada, podem recorrer a conjuntos lexicais que contêm de duas ou mais palavras que, por sua vez, podem ter um significado único. Tal crença foi originariamente formulada por Sinclair (1991: 109-110) e verificada em parte, de forma empírica, por Erman e Warren (2000). Em outras palavras, os usuários de uma língua têm à sua disposição dois princípios fundamentais quando constroem seus textos: o princípio idiomático (*the idiom principle)* e o princípio da escolha aberta (*the open choice principle*).
Sinclair afirma textualmente que podemos lançar mão de repertórios de "frases semi-construídas que, na realidade, se constituem em uma única escolha", além de recorrermos a escolhas individuais. Qualquer texto, na opinião do teórico, é o resultado do entrelaçamento desses dois princípios: ora recorremos a unidades compostas por dois ou mais itens, já ouvidos/lidos e internalizados ou fazemos escolhas complexas de natureza léxico gramatical.
Alguns autores, como Hunston (2002: 143), argumentam que é impossível provar ou refutar a existência dos princípios postulados por Sinclair. Entretanto, a verdade é que ao vasculharmos qualquer corpus eletrônico com software apropriado, podemos extrair agrupamentos com mais de um item lexical, os chamados n-gramas, que tendem a aparecer com regularidade em determinados corpora mais do que em outros. Esses padrões, frequentes em corpora eletrônicos, podem fornecer evidência do principio 'idiomático', ou das unidades ouvidas/lidas e internalizadas pelos sujeitos que deram origem aos textos.
Scott e Tribble (2006: 132) vão mais além, afirmando que um exame cuidadoso de uma lista de agrupamentos lexicais pode ajudar a entender como os textos de usuários experientes são formados e até que ponto os textos de aprendizes coincidem ou se diferenciam dos textos de usuários experientes. Esse é o assunto abordado a seguir.

4. Exemplo prático de estudo de n-grama em corpora de aprendiz

O estudo que reportamos abaixo sobre corpora de aprendiz utiliza dois corpora. O corpus de estudo, chamado Br-ICLE (*Brazilian International Corpus of Learner English*)[5] é formado de 127 composições argumentativas escritas por universitários brasileiros, aprendizes de língua inglesa em nível avançado, cursando o quinto período de graduação em língua inglesa em diante. Cada uma das composições coletadas está identificada em termos de sexo, idade, há quanto tempo o universitário estuda inglês, se foi feita sob condições de teste ou não, com tempo limitado ou não. Há também possibilidade de se saber se o sujeito da pesquisa usou ou não material de consulta, como dicionário, gramática ou qualquer outro material. Nesse corpus são controlados também os tópicos de discussão: o aprendiz escolhe o seu tópico a partir de uma lista contendo 13 assuntos.
Com 65.304 palavras, o corpus é considerado pequeno segundo os parâmetros postulados por Berber Sardinha (2004: 26). Entretanto, mesmo com o auxílio de inúmeras universidades brasileiras a coleta de composições que atendam a esses critérios é muito difícil e, portanto, vagaroso.

[5] O Corpus Br-ICLE não atingiu a meta de 250 mil palavras, portanto está ainda em processo de coleta. As composições coletadas são digitadas exatamente da forma original em que foram submetidas. Erros de ortografia são preservados. Entendemos que o aumento do Br-ICLE pode vir a modificar algumas das posições em que determinadas escolhas de léxico aparecem na listagem final. Entretanto, não invalida o fato de que estamos estudando a linguagem do aprendiz enquanto probabilidade combinatória de determinados itens.

O corpus comparável, ou seja, o corpus que serve de comparação para o corpus de estudo é o LOCNESS (*Louvain Corpus of Native Speaker Essays*), que consiste de 324 194 palavras escritas por população semelhante à população do corpus de estudo. Esse corpus de tamanho médio, segundo os mesmos critérios acima, pode ser adquirido comercialmente. O corpus, que é necessariamente pelo menos três vezes maior do que o corpus de estudo, contem a seguinte distribuição: 60 221 palavras oriundas de textos argumentativos de vestibulandos ingleses, 95 447 palavras de textos argumentativos e comentários literários de universitários ingleses, 149 833 palavras de textos argumentativos de universitários americanos e 18 633 palavras de textos variados produzidos por universitários americanos.

Ainda que se possa levantar críticas à utilização de um corpus comparável contendo as variedades americana e britânica do inglês como língua materna, o Projeto ICLE, que preparou o corpus comparável, alega que a diferença lexical nos *essays* argumentativos contidos no LOCNESS não chega a prejudicar os resultados finais – o LOCNESS vem sendo usado com sucesso em múltiplos estudos constrastivos, iluminando a natureza ´não nativa´ dos corpora de estudo.

O presente trabalho adota a abordagem dirigida pelo corpus, isto é, não lança mão de categorias linguísticas pré-estabelecidas para confirmação de hipóteses. Aliás, no início da atual pesquisa, pouco ou nada se sabia em relação à população de estudo e seus hábitos de escrita, em termos de preferências lexicais.

O estudo segue os preceitos de Scott e Tribble (2006) para a análise de corpora de aprendiz: mais do que fornecer indicativos da interlíngua dos aprendizes, a análise procura desenvolver meios para descrever as estratégias usadas ou não usadas pelos aprendizes com a finalidade de ajudá-los e de, no futuro, informar a prática pedagógica (meu grifo).

Para lidar com os dados, é usado o programa *Wordsmith Tools* v.3. (Scott, 1999) e duas de suas ferramentas mais básicas: um listador de palavras e um concordanciador, ilustrado abaixo no Quadro 2. Nenhum dos dois corpora foi anotado, já que seria difícil uma anotação automatizada confiável em corpus contendo possíveis erros.

Como modo de entrada nos dados, e seguindo a abordagem proposta por Scott e Tribble (2006) são extraídas sucessivamente listas de palavras mais frequentes, bigramas mais frequentes e por fim trigramas e quadrigramas mais frequentes, assim como de palavras-chave. Os autores alegam que um exame detalhado dessas listas ajuda a iluminar não só a preferência por determinados itens lexicais por determinados grupos de escritores ou falantes , mas também a fraseologia inerente a determinados tipos de registros.

Br-ICLE	Item	Freq.	%	**LOCNESS**	Item	Freq.	%
1	THE	3.965	6,07	1	THE	21.118	6,51
2	TO	2.285	3,5	2	TO	10.758	3,32
3	OF	2.172	3,33	3	OF	10.730	3,31
4	AND	1.801	2,76	4	AND	8.327	2,57
5	IN	1.543	2,36	5	A	6.854	2,11
6	A	1.394	2,13	6	IN	6.370	1,96
7	IS	1.318	2,02	7	IS	6.313	1,95
8	THAT	1.062	1,63	8	THAT	4.924	1,52
9	IT	800	1,23	9	IT	3.221	0,99
10	ARE	726	1,11	10	BE	3.197	0,99
11	BE	701	1,07	11	FOR	3.145	0,97
12	NOT	672	1,03	12	AS	2.837	0,88
13	FOR	630	0,96	13	THIS	2.807	0,87

14	**PEOPLE**	619	**0,95**	14	ARE	2.557	0,79
15	AS	530	0,81	15	NOT	2.407	0,74
16	THEY	524	0,8	16	HE	2.186	0,67
17	THIS	512	0,78	17	THEY	2.080	0,64
18	HAVE	507	0,78	18	HAVE	2.048	0,63
19	THEIR	430	0,66	19	WITH	1.909	0,59
20	WE	361	0,55	20	ON	1.796	0,55
21	ALL	322	0,49	21	BY	1.704	0,53
22	WITH	317	0,49	22	**PEOPLE**	1.569	**0,48**

Quadro 1: 22 itens mais frequentes extraídos dos corpora Br-ICLE e LOCNESS

Uma breve análise das 22[6] palavras mais frequentes dos dois corpora evidencia uma coincidência de itens como artigos definidos, pronomes pessoais demonstrativos e preposições, todas essas, formas frequentes na língua inglesa em geral. Entretanto, chama a atenção o item *people*, usado no BrICLE quase duas vezes mais usado do que no corpus LOCNESS. O próximo passo é uma investigação mais detalhada dessa palavra e as opções combinatórias feitas pelos dois grupos, através das linhas de concordância com a palavra de busca ´people´, obtidas com o auxílio do programa Wordsmith Tools. O exame cuidadoso busca padrões frequentes tanto à direita, quanto à esquerda em ambos os corpora e os compara.

N	Concordance
123	more bloodhounds, and upwards of 10 **people** on horseback with rifles. In you
124	arms are used to murder nearly 12,000 **people** annually; another 1,750 persons
125	een 1971 and 1990, more than 14,000 **people** nationwide have become ill fro
126	university as a whole. It only adds 15 **people** to the enrollment and creates
130	ountry of 5000 voters. Supposing 2000 **people** vote for party X and 1500 vote f
131	most votes. However there were 3000 **people** who did not want them to be in
132	zing that it is possible to speak with 4 **people** at once, especially when one p
133	arthquake struck Lisbon killing 40.000 **people** or more and this severely shoo
134	ce known to man." More than 400,000 **people** (in the US), are arrested each
139	er? No one. Who lost? The American **people** who lost their jobs. I feel the
140	ssional football players? The American **people** need to think about what is mo
141	er life. What right do we, as American **people** have to say, “she should not ha
142	ho lost their jobs. I feel the American **people** have been unfairly made to pay
143	rpricing stop? It is up to the American **people** to decide.
144	this whole ordeal. I feel the American **people** elect representatives in the gov
145	ion that therefore effects the American **people** who are not supported by the g
146	re, how many stories will the American **people** miss? The concept of the overr
150	ored, since time began allmost British **people** have been farming and central t
151	n's own identity. Is it why many British **people** are slow to educate themselve
152	mence of my defence of "other" British **people** who were nervous about the wh
153	g beef. Another reason for the British **people** to stop eating beef is the push

Quadro 2. Exemplo de linhas de concordância de *people* extraídas do corpus LOCNESS

Evidencia-se através das concordâncias que os sujeitos do Br-ICLE usam o item com sentido indeterminado e que os horizontes mais frequentes do item são *number of people, people do not, people who are, people have to*, e *people in general*. Se estendermos a lista dos elementos à direita de *people,* verificamos que em sua maioria eles são verbos lexicais (*people believe, people do not/have*). Quando há elementos modificadores para *people,* estes consistem de adjetivos quantificadores, mas marcados como vagos (*many people, a large number of people*). Em contrapartida, no corpus LOCNESS transparecem padrões com as seguintes opções à esquerda: numeral + people, adjetivos gentílicos + people (*American, British, French people*); adjetivos que expressam ocupação (*business people*), faixa etária (*old, young*),

[6] O número de 22 itens é aleatório. Escolhi trabalhar com poucos itens devido a problemas de espaço.

em todos os casos havendo uma tentativa de colocar ´*people* ´em um compartimento. Diferentemente das opções feitas no BrICLE, que não usa adjetivação para *people*, os sujeitos do LOCNESS tendem a usar a palavra, mas caracterizando quem são as ´pessoas´ a que se referem os *essays.*

Quando se fala na quantidade de pessoas, o leque de opções feitas no LOCNESS é bem específico em termos de coligação[7] e escolhas à direita: *more and more people* aparece sempre seguido de construções com *are *ing*. Se a opção é por *many people/millions of people* , a expressão é invariavelmente seguida de processos verbais ou mentais[8], como em *admit, announce, argue, assert, assume, believe, claim, say, think,* entre outros, opções que marcam a introdução de outras ´vozes´ no discurso. Essas múltiplas opções, enquanto padrões, não aparecem no Br-ICLE.

A conclusão que se tira dessa pequena amostra é que, apesar de lançarem mão da palavra com frequência duas vezes maior, quando a usam, os sujeitos brasileiros investigados têm um repertório restrito de combinações. A não utilização de uma gama processos mentais e verbais à direita de *people,* (sua única escolha é *people think*) tira-lhes a opção de trazerem opiniões outras, além de suas próprias.

A análise de itens lexicais individuais começou com uma palavra cujo uso poderia ser considerado excessivo. Entretanto, a investigação de listas de palavras individuais pode também se concentrar nas semelhanças percentuais, como por exemplo, os itens *this* e *that,* que apresentam percentuais próximos. Apesar de os dois grupos investigados usarem uma quantidade semelhante desses itens, as opções combinatórias são muito diferentes.

This é usado no LOCNESS, com frequência, como demonstrativo, acompanhado de um substantivo anafórico, cuja função é expressar a opinião autoral, visto que rotula o que foi dito anteriormente no texto. Os substantivos anáforicos escolhidos pela população americana e britânica estudada são os mais variados, como por exemplo: *this segregation, this system of education, this process, this policy, this argument, this approach*. No Br-ICLE os substantivos abstratos se reduzem a *this situation* e *this problem. This* também aparece no LOCNESS sem o substantivo anafórico e dentro da coligação *this would then* mais verbo lexical (*create, lead to, cause*), estabelecendo relação de causa-consequência no discurso – um padrão que não aparece no Br-ICLE.

Com relação a *that*, ambos os grupos o usam primordialmente como pronome relativo ou conjunção. Entretanto, mais uma vez se olharmos os padrões, desta vez aqueles que se formam à esquerda da conjunção, vemos que no LOCNESS, são outra vez os processos verbais e mentais mencionados acima. Somente o uso de *believe* forma padrões no corpus brasileiro; *claim, conclude e consider* são usados individualmente por um ou outro universitário.

Passando aos bigramas, ou seja as formações de duas palavras (ver anexo), fica clara a ausência no corpus BR-ICLE dos seguintes itens na lista dos bigramas mais frequentes: *can be, would be, should be* e o bigrama *this is*. Presentes no corpus LOCNESS, c*an be, would be* caracterizam atenuação no discurso e *should be* caracteriza modalidade deôntica. Esses recursos que expressam dois pólos da expressão de atitude no discurso e que estão presentes como bigramas frequentes no corpus LOCNESS, já foram objeto de discussão por vários autores (cf. AIJMER, 2002). Portanto, por causa do espaço, não vamos discuti-los aqui. Entretanto, a ausência do bigrama *this is* merece algum comentário, mesmo que breve.

[7] A coligação significa os padrões gramaticais em que um item lexical aparece, ou sua frequente co-ocorrência com determinados itens gramaticais.

[8] O termo 'processo' é usado aqui no sentido da gramática sistêmico-funcional.

This is é usado no LOCNESS, com padrões recorrentes à direita como *This is why/ where/how/ because,* um recurso para elaboração de tópico. Com esse padrões explicam-se causas e consequências, lugares e meios através dos quais algo anteriormente mencionado no discurso aconteceu. Além desse recurso de elaboração, os sujeitos do LOCNESS usam *this is* com a seguinte coligação This is + (artigo) + adjetivo + substantivo, como em *this is a positive aspect, this is a welcome solution,* um padrão usado como recurso avaliativo, como expressão da voz autoral. A ausência no corpus brasileiro desse padrão seja compensada, talvez, pelo uso de *I think* , um bigrama frequente nesse corpus, mas também o mais frequente no corpus Camcode, corpus de inglês oral estudado em O`Keefe et al.(2007).
O próximo passo da proposta de análise se concentra em trigramas, ou agrupamentos de três itens obtidos pelo programa extrator. Como dizem Scott e Tribble (2006: 132), o estudo de agrupamentos lexicais ou n-gramas em coletâneas relevantes de textos nos fornece *insights* da fraseologia desses mesmos textos. No caso de textos de autores publicados e de aprendizes, o estudo tem o potencial de aumentar o nosso entendimento (e dos aprendizes) sobre a fraseologia que é usada e aquela que deveria ser preterida nos mesmos textos.

BR	Word	Freq.	%
1	IN ORDER TO	79	0,12
2	THE FACT THAT	35	0,05
3	IT IS NOT	34	0,05
4	AS WELL AS	33	0,05
5	ON THE OTHER	32	0,05
6	ONE OF THE	28	0,04
7	THE OTHER HAND	28	0,04
8	THERE IS NO	25	0,04
9	*THEY DO NOT*	*25*	*0,04*
10	THE END OF	22	0,03
11	IT IS A	21	0,03
12	*MORE AND MORE*	*19*	*0,03*
13	THE NUMBER OF	19	0,03
14	*THE ONES WHO*	*19*	*0,03*
15	BE ABLE TO	18	0,03
16	*IN OTHER WORDS*	*18*	*0,03*
17	THERE IS A	18	0,03
18	*AT THE SAME*	*17*	*0,03*
19	*IT IS POSSIBLE*	*17*	*0,03*
20	*OF THE WORLD*	*17*	*0,03*

Locness	Word	Freq.	%
1	THE FACT THAT	162	0,05
2	IN ORDER TO	130	0,04
3	ONE OF THE	123	0,04
4	THAT IT IS	105	0,03
5	BE ABLE TO	94	0,03
6	THERE IS NO	94	0,03
7	THE RIGHT TO	85	0,03
8	IT IS NOT	84	0,03
9	DUE TO THE	82	0,03
10	THE END OF	82	0,03
11	BECAUSE OF THE	80	0,02
12	THERE IS A	78	0,02
13	THE IDEA OF	77	0,02
14	AS WELL AS	76	0,02
15	END OF THE	70	0,02
16	IT IS A	70	0,02
17	THE USE OF	69	0,02
18	THIS IS A	68	0,02
19	SHOULD NOT BE	66	0,02
20	THE NUMBER OF	65	0,02

Quadro 3: Lista dos 20 trigramas mais frequentes nos corpora Br-ICLE e LOCNESS

Há vários modos de lidar com trigramas, que incluem todos os trigramas dos corpora e/ou somente os mais frequentes, como no quadro acima. O primeiro seria extrair os trigramas-chave que caracterizam o corpus Br-ICLE. Estes são calculados pelo programa extrator e se apresentam nesta ordem de importância, ou seja, estes são usados com mais frequência no corpus de estudo do que se espera, ao contrastar o corpus de estudo com o corpus comparável: *in order to, the ones who, to sum up, in other words, is necessary to, a great number, point of view*. Uma vez extraídos, faz-se uma análise manual dos mesmos, estendendo a busca tanto para a direita quanto para a esquerda nos dois corpora para averiguar as diferentes

preferências colocacionais e coligacionais, que é a abordagem praticada por Scott e Tribble (2006).

Uma outra abordagem seria simplesmente contrastar os dois quadros contendo trigramas mais freqüentes, tendo como ponto de partida aquilo que é compartilhado, e notar o percentual de uso. Fica evidente, por exemplo, que a expressão ´in order to´ é usada três vezes mais no corpus de aprendizes, o que poderia ser indicativo de ausência de formas alternativas para expressar meio/fim por parte dos aprendizes. Se ao contrário, o foco é aquilo que está ausente, o que fica evidente é que na lista do LOCNESS há dois meios de explicar causa-consequência (*because of the* e *due to the*), uma relação que não transparece nos trigramas mais frequentes do Br-ICLE.

Um último caminho de análise para os trigramas seria etiquetá-los com categorias desenvolvidas em outros trabalhos de análise de n-gramas, como por exemplo Biber et al. (2004) ou Hyland (2008) para averiguar se os trigramas apontam prioritariamente para a organização do texto (*as well as, on the other, in other words, at the same*), para a organização das posições do escritor (*be able to, should not be, it is possible*) ou para meios de enquadrar o tópico que está sendo desenvolvido (*the fact that, the number of, the use of*).

Como as análises dos bigramas e trigramas acima, a análise de quadrigramas envolve igualmente "o entrelaçamento de listas de palavras (e às vezes de palavras-chave) com um estudo cuidadoso dos textos de onde elas foram extraídas" (Scott e Tribble, 2006: 134).

Mesmo que essa verificação abranja um mínimo número de quadrigramas, como na lista abaixo, que cobre tão somente os dezesseis mais frequentes dos dois corpora deste trabalho, há evidências de fatos interessantes.

Em termos de quadrigramas-chave do corpus Br-ICLE há somente *at the end of* e *a great number of*. Enquanto que o primeiro quadrigrama expressa ênfase em ancorar o texto numa linha de tempo (*at the end of* é seguido de um século), *a great number of* não existe na língua inglesa, podendo configurar não internalização do quadrigrama *a large number of* (or *a great deal of*).

Br-ICLE		Freq	%		LOCNESS		Freq.	%	
1	ON THE OTHER HAND	28		0,04	1	THE END OF THE	67		0,02
2	IT IS POSSIBLE TO	16		0,02	2	ON THE OTHER HAND	50		0,02
3	AT THE SAME TIME	15		0,02	3	AT THE END OF	42		0,01
4	THE END OF THE	13		0,02	4	ONE OF THE MOST	31		
5	ALL OVER THE WORLD	12		0,02	5	AS A RESULT OF	30		
6	OF THE #TH CENTURY	12		0,02	6	IS ONE OF THE	30		
7	IT IS IMPORTANT TO	11		0,02	7	IN THE CASE OF	28		
8	IT IS NECESSARY TO	10		0,02	8	THE FACT THAT THE	28		
9	ONE OF THE MOST	10		0,02	9	AT THE SAME TIME	25		
10	IN OUR MODERN WORLD	9		0,01	10	THE BEGINNING OF THE	24		
11	AS WELL AS THE	8		0,01	11	TO THE FACT THAT	24		
12	A great number of	7		0,01	12	AT THE BEGINNING OF	22		
13	THAT THERE IS NO	7		0,01	13	DUE TO THE FACT	21		
14	THERE WILL ALWAYS BE	7		0,01	14	THE ONLY WAY TO	21		
15	TO THE FACT THAT	7		0,01	15	THE REST OF THE	21		
16	WE LIVE IN A	7		0,01	16	A GREAT DEAL OF	20		

Quadro 4: Lista dos 16 quadrigramas mais frequentes nos corpora Br-ICLE e LOCNESS

Um outro fato interessante que pode ser depreendido da pequena lista acima, é que há evidência de escolha do enquadramento colocacional *it is + possible/ important/ necessary + to*, no corpus Br-ICLE, como forma preferida para expressar uma atitude autoral. Este enquadramento, não escolhido como alternativa frequente pelos universitários americanos ou britânicos do corpus LOCNESS, é encontrado, em contrapartida, como forma preferida em textos de áreas acadêmicas[9] em língua inglesa (ver Anexo 2). Tal fato, que necessita ser explorado com mais profundidade, pode configurar ou a adoção de escolhas mais informais para os *essays* escritos pelos sujeitos do LOCNESS, ou escolhas mais formais pelos sujeitos do corpus Br-ICLE.

Conclusão

O presente trabalho apresenta uma análise baseadas em n-gramas, extraídos de corpora de aprendiz. Alinhando-se a Scott e Tribble (2006), o estudo entende que o foco em corpora pequenos, mas representativos de populações de aprendizes, pode fornecer recursos práticos para melhor observar as opções léxico-gramaticais desses grupos.

A despeito da simplicidade do presente estudo de natureza indutiva, é possível concluir a partir de um exame das freqüências de que o componente lexical do ensino de escrita argumentativa em língua inglesa para universitários brasileiros poderia incluir alguma conscientização sobre:

a) palavras ditas vagas como *people*,
b) modos de inclusão de outras vozes na argumentação,
c) sinonímia para processos verbais e mentais,
d) as possibilidades anafóricas do pronome *this*;
e) substantivos abstratos anafóricos, entre outros tópicos.

Entretanto, há uma grande distância entre se conscientizar e internalizar.

Nesta parte final do artigo, portanto, retoma-se a discussão proposta por Scott e Tribble (2006) sobre a necessidade de a análise de corpora de aprendiz dar algum retorno à pratica pedagógica, retorno esse que, no momento, parece remoto.

Sabemos hoje em dia que o plano inicial para que os resultados das pesquisas em Linguística de Corpus levassem automaticamente à criação de atividades de ensino e aprendizagem baseados em corpora não se concretizou (BRAUN et al., 2006: 1). Um breve levantamento desse possível retorno pedagógico de dados oriundos de corpora se mostra incipiente (ver Mukkerjee, 2006, para o estado da arte na prática pedagógica baseada em corpus). A produção materiais para a sala de aula de língua inglesa, por exemplo, se restringe às atividades em DDL (*data-driven learning*) criadas e defendidas por Johns (1986) para o ensino de inglês com fins específicos. Essas atividades, que exploram formas e significados de itens lexicais através de linhas de concordância fornecidas aos aprendizes, foram adaptadas em alguns livros didáticos, como por exemplo, aqueles produzidos no projeto Cobuild para o ensino de inglês geral. Outras tentativas de aproveitamento consistem em séries de livros didáticos que apregoam uma base em corpus, mas que se restringem a utilizar o corpus para, baseando-se em frequências extraídas de corpora, selecionar itens lexicais para o ensino de inglês geral.

[9] Este enquadramento faz parte dos quadrigramas mais frequentes no sub-corpus de linguagem acadêmica do *British National Corpus* (ver http://site.ebrary.com/pub/benjamins/docDetail.action?docID=10126062&p00=scott%20tribble

No Brasil, o Grupo GELC (Grupo de Estudos em Linguística de Corpus) vem, desde 2002, produzindo trabalhos acadêmicos voltados para a interface Linguística de Corpus e Prática Pedagógica[10], com especial destaque a Veirano (2008), Boscariol-Bertolino (2008) e Moreira Filho (2008). Em 2008 ainda, Moreira Filho também criou software dirigido ao professor de língua inglesa e espanhola para preparação semi-automática de atividades de leitura com a ajuda de corpora[11]. Na área de corpora para a sala de aula de inglês com fins específicos, há Perroti-Garcia e Rebechi (2007) e Perroti-Garcia (no prelo). Há ainda, em desenvolvimento, software para correção automática de erros em inglês baseados em etiquetação de erros frequentes ocorridos no corpus Br-ICLE (BERBER-SARDINHA e SHEPHERD, 2008).
A despeito desses esforços, os trabalhos com corpus de aprendiz e corpora em geral permanecem na área descritiva e fazem pouco investimento no pedagógico.

Abstract: The present work lies at the interface between Corpus Linguistics and texts produced by learners. An attempt is made to problematise linguistic research using learner corpora and relate it to a discussion on the nature of errors in foreign language writing. The work thus provides a practical example of an inductive approach to two corpora, compiled in comparable settings, containing written texts in English as a mother tongue and as a foreign language. A step by step procedure is suggested as a means of coping with both the analysis of units which incorporate n-grams, i.e., lexical bundles. Finally, the work focuses on the production of pedagogical materials in Brazil, which stem from the analyses of learner corpora in Brazil.

Key-words: Corpus Linguistics; corpus-driven analysis; learner corpora; n-grams

Referencias bibliográficas

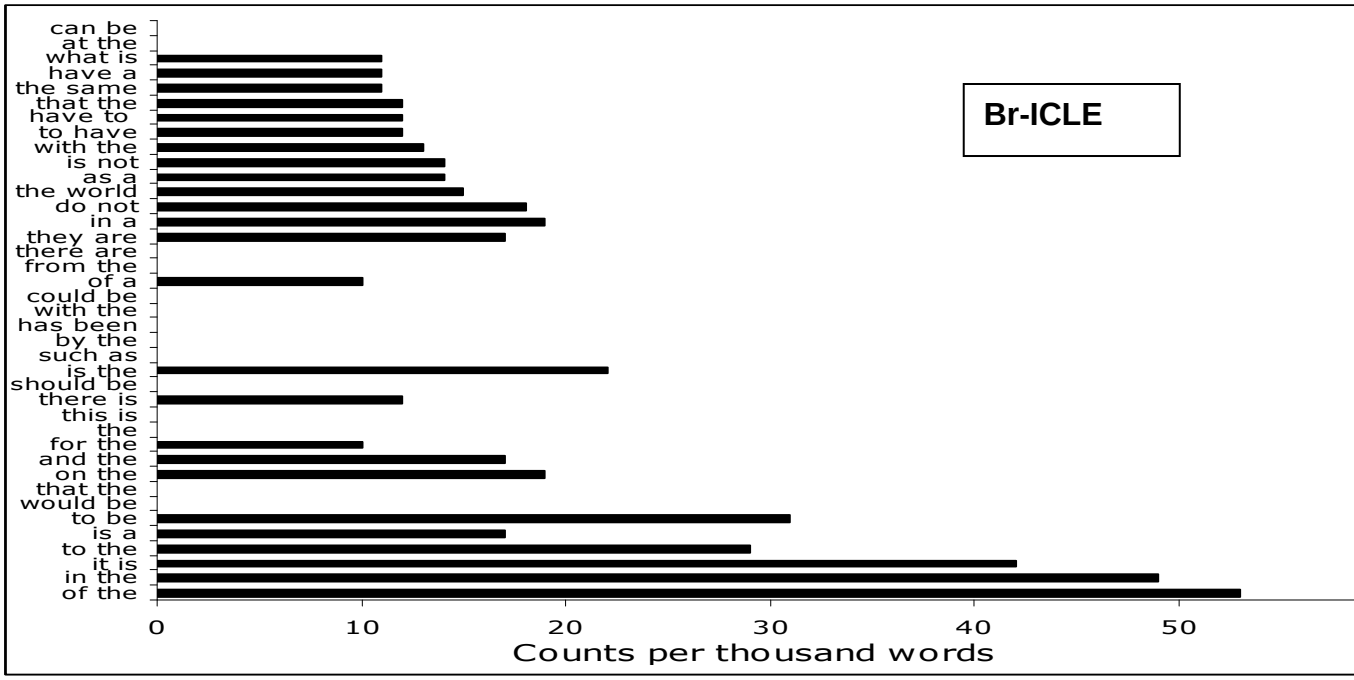
AIJMER, K. *English Discourse Particles: evidence from a corpus*. Amsterdam: John Benjamins, 2002. 299p.
ALTENBERG, B. Using bilingual corpus evidence in learner corpus research. In.: GRANGER, S.; HUNG, J.; PETCH-TYSON, S. (eds.) Computer learner corpora, second language acquisition and foreign language teaching. Amsterdam: John Benjamins, 2002. p37-54.
ALMEIDA, M.I.A. Trabalhando com o computador na pesquisa linguística: o uso do modal *can* por brasileiros e ingleses. In.: VASCONCELLOS, Z.; AUGUSTO, M.; SHEPHERD, T.M.G.. (Orgs.). *Linguagem, Teoria, Análise e Aplicações (3)*. Rio de Janeiro: Editora Letra Capital, 2007.
BARLOW, M. Computer-based analyses of learner language. In.: Ellis, R.; Barkhuizen, G. (eds.). *Analysing Learner Language*. Oxford: Oxford University Press, 2005. p. 335-357,.
BERBER SARDINHA, T. A. *Linguística de Corpus*. São Paulo: Manole, 2004. 410p.
BERBER SARDINHA, T. A.; XXXXXXX. An online system of error identification in Brazilian learner English. *Proceedings of the 8th Teaching and Language Corpora Conference*. Lisboa: Associação de Estudos e de Investigação Científica do ISLA, 2008. p.257-263.
BIBER, D.; CONRAD, S.; CORTES, V. If you look at…: lexical bundles in university teaching and textbooks. *Applied Linguistics*, v. 25, n. 3, p. 371-405,. 2004.

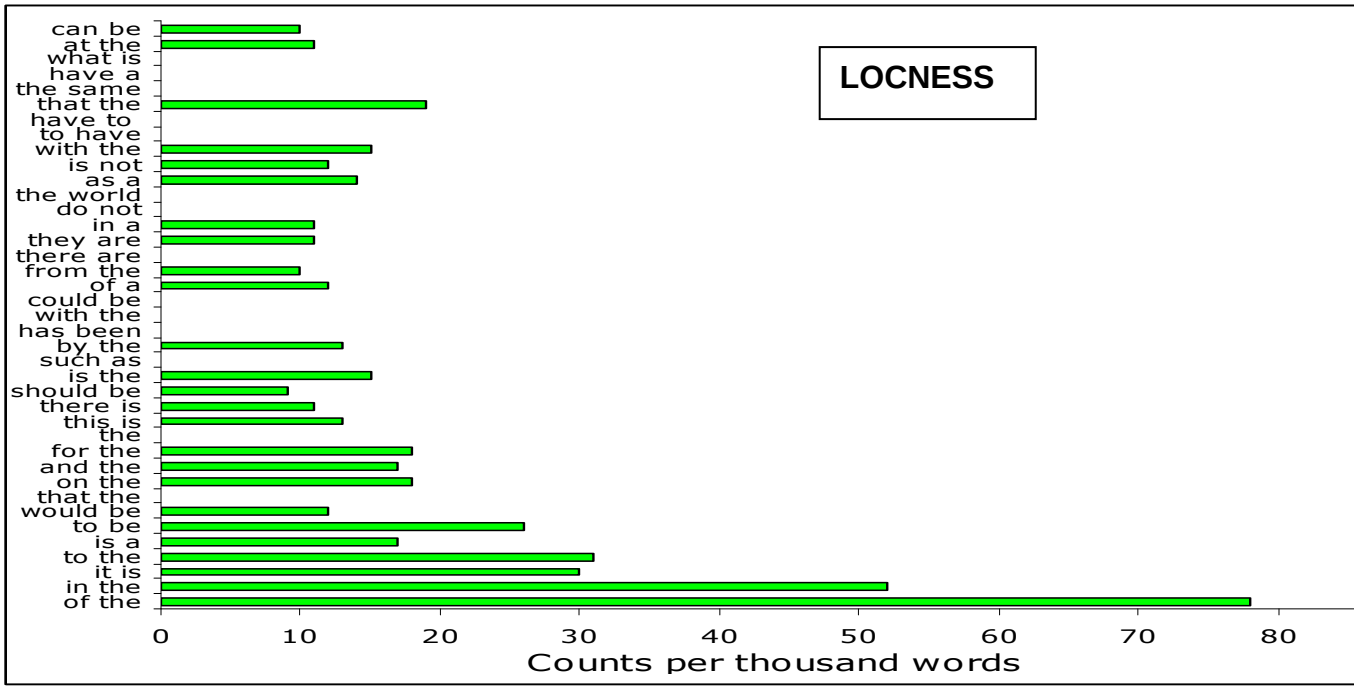
[10] Consultar http://www.pucsp.br/pos/lael/lael-inf/def_teses.html
[11] Disponível em http://www.maxwell.lambda.ele.puc-rio.br/cgi-bin/db2www/PRG_1440.D2W/REPORT1?CdLinPrg=pt&NrSeqFas=58

BIBER, D. Lexical bundles in academic speech and writing. In.: LEWANDOWSKA-TOMASZCZYK, B. (ed.). *Practical applications in language and computers*. Frankfurt: Peter Lang, 2004. p.165-178.
BIBER, D. et al. *Longman grammar of spoken and written English*. London: Longman, 1999.
BOSCARIOL-BERTOLINO, M. *A linguagem dos RPGs eletrônicos e o ensino de inglês.* 2008. Dissertação (Mestrado em Linguística Aplicada e Estudos da Linguagem) – Pontifícia Universidade Católica de São Paulo. São Paulo, 2008.
DE COCK, S. et al. An automated approach to the phrasicon of EFL learners. In.: GRANGER, S. (Ed.). *Learner English on computer*. London: Longman, 1998c. p. 67-79.
GRANGER, S.; HUNG, J.; PETCH-TYSON, S. (eds.) Computer learner corpora, second language acquisition and foreign language teaching. Amsterdam: John Benjamins, 2002. 257p.
ELLIS, R. *The study of second language acquisition*. Oxford: Oxford University Press. 1994. 824p.
ERMAN, B.; WARREN, B.. The idiom principle and the open choice principle. *Text* 20.1, 29-62, 2000.
FAIRCLOUGH, N (2003) *Analysing discourse: textual analysis for social research*. London; New York: Routledge, 2003
FERRIS, D. *Treatment of error in second language student writing.* Ann Arbor: The University of Michigan Press, 2004.
FIRTH, J R. (1957) Modes of meaning. *Essays and Studies. The English Association, 118-149.*
GRANGER, S. The computer learner corpus: a versatile new source of data for SLA research. In.: _____ (ed.). *Learner English on computer*. London: Longman, 1998a. p. 3-18.
GRANGER, S. Prefabricated patterns in advanced EFL writing: collocations and formulae. In.: COWIE, A. P. (Ed.). *Phraseology: theory, analysis and applications*. Oxford: Oxford University Press, 1998b. p. 145-160.
GRANGER, S. (ed.). *Learner English on computer*. London: Longman, 1998c. 228p
GRANGER, S.; HUNG, J. PETCH-TYSON, S. (eds.) Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching. Amsterdam: John Benjamins, 2002. 246p.
HUGHES, A.; LASCARATOU, C. Competing criteria for error gravity. *ELT Journal,* 36/3, p.175-182, 1983.
HYLAND, K. As can be seen: lexical bundles and disciplinary variation. *English for Specific Purposes.* 27, 4-21, 2008.
HYLAND, K. *Teaching and researching: Writing*. Harlow: Longman, 2002. 248p.
HYLAND, K; ANAN, E. Teachers' perceptions of error: the effects of first language and experience. *Syste,* v4-34, p509-520, 2006.
HUNSTON, S. *Corpora in applied linguistics.* Cambridge: Cambridge University Press, 2002. 241p.
KNOBLAUCH, C.H; BRANNON, L. *Rhetorical Traditions and the Teaching of Writing*. Upper Montclair, NJ, Boynton/Cook, 1984.184p.
KRISHNAMURTHY, R. (ed.) *English Collocation Studies: the OSTI Report*. London: Continuum, 2004. 208p.
JOHNS, T. From printout to handout: grammar and vocabulary teaching in the context of data-driven Learning." In.: ODLIN, T (ed.) *Perspectives on Pedagogical Grammar*. New York: Cambridge University Press, 1994. p.293-313.

LEECH, G. Teaching and Language Corpora: a convergence. In.: WICHMAN, A.; FLIGELSTONE, S; MCENERY, T.; KNOWLES, G.. (eds.) *Teaching and Language Corpora*. Harlow: Addison Wesley, 1997. 343p.
McCRETTON, E.; RIDER, N. Error gravity and error hierarchies. *International Review of Applied Linguistics.* 12-2, p. 180-196, 1993.
MOON, R. Frequencies and forms of phrasal lexemes in English. In: COWIE, A. P. (Ed.). *Phraseology: theory, analysis and applications*. Oxford: Oxford University Press, 1998. p. 79-100.
MOREIRA FILHO, J. *Desenvolvimento de um software para preparação de aulas de inglês com corpora*. 2007. Dissertação (Mestrado em Linguística Aplicada e Estudos da Linguagem) – Pontifícia Universidade Católica de São Paulo. São Paulo, 2007. Disponível em http://www.pucsp.br/pos/lael/lael-inf/teses/jose_lopes_moreira_filho.pdf
MUKHERJEE, J. Corpus Linguistics and language pedagogy: the state of the art – and beyond. In.: BRAUN, S.; KOHN, K.; MUKHERJEE, J. (eds.) *Corpus Technology and Language Pedagogy*. Frankfurt am Main: Peter Lang, 2006. p.5-24.
O´ KEEFFE, A.; MCCARTHY, M.; CARTER, R. From Corpus to Classroom. Cambridge: Cambridge University Press, 2007.
PERROTI-GARCIA, A. J. *Curso de Inglês Odontológico (Dental English*) . São Paulo: Editora Galpão, no prelo.
PERROTI-GARCIA, A. J.; REBECHI, R. *Vocabulário para Química. Série 1001 Termos.* São Paulo: SBS, 2007.
RENOUF, A.; SINCLAIR, J. Collocational frameworks in English. In: AIJMER, K.; ALTENBERG, B. (Ed.). *English corpus linguistics*. London: Longman, 1991. p. 128-143.
RINGBOM, H. Vocabulary frequencies in advanced learner English: a cross-linguistic approach. In.: GRANGER, S. (ed.) *Learner English on computer*. London: Longman, 1998. 228p.
SCOTT, M. *Wordsmith Tools*. Oxford: OUP, 1999.
SCOTT, M.; TRIBBLE, C. (2006). *Textual Patterns: keywords and corpus analysis in language education.* Amsterdam: John Benjamins, 2006. 214p.
SINCLAIR, J. *Corpus, concordance, collocation*. Oxford: Oxford University Press, 1991. 197p.
SINCLAIR, J. Preface. In: LEWANDOWSKA-TOMASZCZYK, B. (Ed.). *Practical applications in language and computers*. Frankfurt: Peter Lang, 2004. p. 7-11.
TELIYA, V. et al. Phraseology as a language of culture: its role in the representation of a collective mentality. In.: COWIE, A. P. (ed.). *Phraseology: theory, analysis and applications*. Oxford: Oxford University Press, 1998. p. 55-75.
TOGNINI BONELLI, E. *Corpus Linguistics at Work.* Amsterdam: John Benjamins, 2001. 223p.
VEIRANO, M. *O uso de things, thing, anything, something e everything em corpora de aprendiz*. 2008. Dissertação (Mestrado em Linguística Aplicada e Estudos da Linguagem) – Pontifícia Universidade Católica de São Paulo. São Paulo, 2008. Disponível em http://www.pucsp.br/pos/lael/lael-inf/teses/tese_marcia.pdf
WRAY, A. *Formulaic language and the lexicon*. Cambridge: Cambridge University Press, 2002, 332p.
WRAY, A. Formulaic language in learners and native speakers. *Applied Linguistics* (32), 213-231, 1999.

Anexo 1. Demonstrativo das diferenças de padrão de bigramas entre os corpora





VEREDAS ON-LINE – LINGUÍSTICA DE CORPUS E COMPUTACIONAL – 2/2009, P. 100-116 – PPG LINGUÍSTICA/UFJF – JUIZ DE FOR A – ISSN 1982-2243

Anexo 2 Distribuição do enquadramento colocacional it is + adjetivo + to nos vários registros do *British National Corpus*.