O ESTATUTO DA LINGUÍSTICA DE CORPUS: METODOLOGIA OU ÁREA DA LINGUÍSTICA?

Tania M.G. Shepherd
UERJ/FAPERJ/CNPq

RESUMO

Este trabalho problematiza a pesquisa linguística a partir de corpora digitais, à luz de questionamentos sobre o papel do que se convencionou chamar Linguística de Corpus (doravante LC). Assim, o artigo inicia com discussão sobre o estatuto da LC: se a LC pode ser considerada um ramo da Linguística ou se é uma mera metodologia que lança mão do computador na investigação de fenômenos linguísticos. Sem favorecer uma ou outra posição, em um segundo estágio o trabalho apresenta exemplos práticos de abordagem indutiva e dedutiva em recentes pesquisas realizadas no âmbito dos estudos de inglês como língua estrangeira em corpora de aprendiz brasileiro. O artigo sugere que o status de ´ramo da linguística´ ou 'metodologia de tratamento de dados' pode ser estabelecido ao se decidir o ponto de entrada em dados oriundos de textos eletrônicos. Qualquer que ele seja, os resultados obtidos fazem importantes contribuições para o modo que são vistos os fenômenos linguísticos.

PALAVRAS-CHAVE: Linguística de corpus; abordagem dirigida pelo corpus; abordagem baseada em corpus.

1. Introdução

O presente artigo é sobre Linguística de Corpus (doravante LC) e por conseguinte, lida com significado. O artigo não é sobre outras linguísticas, como por exemplo a Linguística Cognitiva. Em outras palavras, ele aborda significado, mas não lida com entendimento ou cognição, apesar de, conforme alega Teubert (2004, p. 38) significado e entendimento/ cognição poderem ser entendidos, por vezes, como complementares. Por ser sobre LC, o artigo aborda significados co-construídos: o significado de um item lexical ou expressão é o que é, porque nos foi passado por uma ou várias pessoas e essas devem tê-lo ouvido ou lido de outras. Além disso, porque o trabalho é sobre LC, ele enfoca significados extraídos de compilações de textos em formato digital, ou corpus. Em suma, sendo sobre Linguística de Corpus, o presente trabalho entende que seu ´objeto de estudo´ não é um fenômeno mental, mas um fenômeno social, algo observável e acessível através de evidências que advêm de corpus digitalizado.

Definir o que é um c*orpus* (plural, *corpora*) para a LC parece um ponto de partida interessante para a presente discussão. As definições de corpus, via de regra, ressaltam que um corpus é uma coletânea de textos em linguagem natural, escritos ou falados, geralmente armazenados de forma organizada e informada, além de serem digitalizados a fim de que possam ser lidos por computador. Ainda que a definição de autores diversos ressalte uma ou outra característica, um corpus para a LC espelhará esses fatores principais.

Entretanto, se a definição de 'corpus' não motiva grandes discussões teóricas, o rótulo Linguística de Corpus causa muita comoção[1]. Uma pergunta feita invariavelmente pelos recém-iniciados nos estudos de corpora digitalizados é se a LC seria uma área da Linguística, com objeto de investigação e conceitos próprios, ou uma metodologia de investigação que parte de grandes quantidades de dados textuais armazenados em forma digital.

Pode-se afirmar, sem medo de errar, que essa preocupação em estabelecer qual a identidade da LC caracteriza o início de um grande número de livros sobre o assunto. Isto acontece de tal forma, e com tanta frequência, já há duas décadas, que se poderia caracterizar a seção introdutória de alguns livros sobre LC como um ´movimento retórico´ que consiste em fornecer a opinião do autor sobre o impasse linguística ou metodologia.

Em um dos primeiros volumes dedicados à LC, McEnery e Wilson (1996, p. 2) perguntam e respondem textualmente "A Linguística de Corpus é um ramo da Linguística? A resposta a essa pergunta é tanto 'sim' como 'não'"[2]. Os autores alegam que a LC não tem o mesmo estatuto da Semântica, Sintaxe ou Sociolinguística visto que estas disciplinas têm um objeto de investigação definido. Ao mesmo tempo, os autores alegam que o termo corpus pode ser atrelado a cada uma das áreas da Linguística, gerando portanto a 'Semântica de corpus', a 'Sintaxe de corpus', por exemplo, em oposto à Semântica ou Sintaxe não baseadas em corpora.

O foco de Biber et al. (1998, p. 3) não fica muito distante daquele proposto por McEnery e Wilson. Biber et al. (obra citada) dizem que os estudos da linguagem têm normalmente dois enfoques, ou seja, existem aqueles estudos que priorizam a estrutura e aqueles que priorizam o uso. Acrescentam ainda que examinar o uso significa encontrar ´padrões´ de ocorrência dentro de fatores contextuais. Tal empreitada não pode depender das intuições do analista visto que os seres humanos tendem a reconhecer aquilo que não é típico com muito mais frequência do que aquilo que é padronizado. Daí o papel da LC: fornecer meios de lidar com grande quantidades de dados provenientes do uso, além de, simultaneamente, acompanhar as variáveis contextuais. Biber et al. (1998, p. 11) confirmam o status de metodologia da LC dizendo que uma abordagem que parte do corpus pode ser aplicada a praticamente qualquer área de investigação linguística, algo semelhante ao que McEnery e Wilson alegam quando dizem ser possível atrelar a expressão ´de corpus´ a qualquer ramo da Linguística.

Em 1999, Kennedy (1999, p. 6) retrocede mais ainda e confere à LC um status que é pouco mais do que ser fonte de exemplos. O autor diz que seria errado sugerir que a LC é uma teoria da linguagem que compete com outras teorias da linguagem, como por exemplo a gramática transformacional. Para Kennedy, seria mais errado ainda afirmar que a LC é um ramo novo ou separado da Linguística. Kennedy alega que, assim como durante um período a Linguística recorreu à intuição ou introspecção do linguista para fornecer exemplos, da mesma forma os linguistas de corpus recorrem à LC para fornecer evidência que advém diretamente de textos digitalizados.

Mais recentemente, as afirmativas sobre o estatuto da L.C. ainda pendem para um lado ou para o outro. Scott e Tribble (2006, p. 3-4),

por exemplo, afirmam que os métodos baseados em corpora são somente isto: uma metodologia que utiliza corpora formados de textos escritos ou falados, ou seja de exemplos genuínos de linguagem em uso. Empregando imagem de recentes técnicas cirúrgicas, dizem que a Linguística de Corpus está para a Linguística assim como a vídeolaparoscopia está para a Medicina, isto é, vídeolaparoscopia não é um ramo da Medicina e LC não é Linguística. Entretanto, acrescentam que a LC está abalando as fundações da Linguística assim como a descoberta da estrutura básica do DNA (a molécula de dupla hélice que guarda todas as informações genéticas dos organismos dentro das células) está abalando todas as concepções da biogenética, e assegurando um lugar para novos conceitos como a clonagem, os transgênicos, entre outros.

Alguns autores deixam um pouco menos nas entrelinhas e, assim como Bernardini et al. (2003) aceitam que "... a linguística de corpus está hoje em dia firmemente estabelecida como área de pesquisa *e* como metodologia.[3] (meu grifo). Ou ainda, especialmente como Tognini-Bonelli (2001) apregoam que há dois modos consagrados de abordagem de corpora eletrônicos em geral: a abordagem baseada em corpus (*corpus-based*) e a abordagem direcionada pelo corpus (*corpus-driven*). Essas duas práticas já estabelecidas entre os linguistas de corpus dependem de como se entra no corpus.

A abordagem baseada em corpus é na realidade uma metodologia que se aproveita do corpus essencialmente para testar e exemplificar teorias e descrições linguísticas pré-existentes. Através dessa abordagem, o corpus pode ser usado como fonte de exemplos, que são quantificáveis em sua frequência e extensão. Esses exemplos podem advir de um corpus ´cru´, sem anotação, ou de um corpus anotado automática ou manualmente em termos de unidades gramaticais ou semânticas, entre outros inúmeros tipos de anotação.

Por outro lado, a abordagem direcionada pelo corpus se deve, segundo Sinclair (2004a, p. xviii) à ausência de uma teoria que explicasse as relações lexicais na gênese dos trabalhos com corpora eletrônicos. Como explica ainda Sinclair, nos anos sessenta, não havia teoria que explicasse a ocorrência, frequência e padronização lexicais, além da preferência (e também a rejeição) de certas palavras por outras. A abordagem direcionada pelo corpus, portanto, visa à observação e análise de padrões e frequências lexicais. A observação desses padrões pode levar à hipótese, que pode levar à generalização. Em outras palavras, os

dados obtidos a partir de corpora podem ser usados para a formulação de descrições de natureza léxico-gramatical.

Para entendermos as abordagens direcionadas pelo corpus, precisamos voltar no tempo. 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. A pesquisa lexical realizada entre 1967 e 1969 para o *Office for Scientific and Technical Information* britânico descrita no Relatório OSTI, (KRISHNAMURTHY, 2004) revela que também data dessa época a formalização pioneira de critérios para a investigação de relações sintagmáticas entre itens lexicais. Essa pesquisa lexical concretizaria a natureza contextual do significado, já argumentada por Firth (1957, p. 11)[4], ao postular que "se conhece uma palavra pelas companhias com as quais ela anda".

Dentre as inúmeras propostas feitas através da abordagem direcionada pelo corpus, a partir de observação de corpora, há cinco conceitos principais que descrevem de certa forma a tendência que certas palavras têm de associar a outras. O primeiro conceito, verificável através de corpora, é o de 'colocação', ou associação sintagmática de itens lexicais com outros itens lexicais. Como exemplo, pode-se citar os colocados de ´café´, à sua direita, como ´com pão e manteiga´, ´da manhã´ e ´pingado´ e à sua esquerda como, ´sacas de´, ´exportação de´'. O segundo conceito é o de 'coligação' ou associação sintagmática de itens lexicais em posições estruturais, incluindo-se aí o conceito de coligação textual, isto é, certos itens lexicais aparecem somente em determinados locais dentro dos textos. O terceiro conceito é o de preferência semântica, ou seja determinados itens lexicais parecem preferir determinadas áreas conceituais como por exemplo 'atrás', que prefere primordialmente conceitos temporais (dois anos atrás). O quarto conceito é o de prosódia semântica, ou seja o fato de que certos itens, apesar de sinônimos, têm prosódias negativas ou positivas. Como exemplo, citamos as palavras 'consequência' e 'resultado', estudo de Berber Sardinha (2004). Um exame cuidadoso das ocorrências dos dois itens indica que 'consequência' tem prosódia negativa, como em a 'consequência de seus atos', a 'consequência do terremoto'; 'resultado´, por outro lado é neutro ou positivo, como em 'resultado da pesquisa'. O quinto conceito tem a ver com associações entre itens lexicais e seus contextos: não só os itens lexicais aparecem em determinados contextos

(ou registros, ou gêneros) preferivelmente, mas o fazem com frequências específicas (ver BIBER et al., 1999).

Uma outra importante descoberta é fruto também da abordagem direcionada pelos dados e tem a ver com os dois princípios organizacionais da linguagem postulados por Sinclair (1991, p. 109-110), discutidos mais abaixo - o 'principio da livre escolha' e o 'principio idiomático' - que são vitais para entender por que cadeias lexicais se repetem com frequência dentro de determinados gêneros ou registros. Tais agrupamentos lexicais, também chamados n-gramas, podem ser formados com 2, 3, e por vezes formados de 5 ou mais elementos.

Além das verificações empíricas sobre colocados, coligados, n-gramas, entre outros, a abordagem dirigida pelos dados foi diretamente responsável pela Teoria do Priming Lexical que postula a reversão dos papéis desempenhados pelo léxico e gramática (HOEY, 2005). A teoria afirma que o léxico é complexa e sistematicamente estruturado e que a gramática nada mais é que o resultado dessa estrutura lexical.

Os estudos com corpus digital, portanto, já existem há cerca de 40 anos e têm uma prática de fornecer evidências sólidas para seus achados. Esses estudos podem conter tanto uma metodologia de tratamento de dados, quanto 'um caminho para a Linguística', expressão propositalmente ambígua formulada por Hoey (comunicação pessoal).

A pesquisa em LC, entretanto, nem sempre teve credibilidade, apesar das evidências fornecidas em sua defesa. Sinclair (2004, p. 2) afirma que, no início, "a evidência contida em corpora foi ignorada, rejeitada, e descaracterizada até que a sua importância ficou óbvia demais para ser retirada da mesa de discussões."[5] As argumentações usadas contra os estudos baseados em corpus, resumidas por Sinclair (idem) consistiram por muito tempo em a) um corpus não é um exemplo acurado da língua; b) porque algo está contido num corpus, não quer dizer que esteja ´correto´; c) frequência de ocorrência não quer dizer importância de ocorrência e d) um corpus, ainda que enorme, tem lacunas, e, desta forma, não representa o todo da língua. Apesar de todos os ataques, desde a década de sessenta, muitos têm sido os trabalhos sobre o léxico, a partir das duas abordagens (baseada em ou dirigida por corpus) que constituem a Linguística de Corpus – o estudo de textos eletrônicos com o auxílio de computador. Inúmeros são também os corpora formados de textos eletrônicos disponíveis tanto comercialmente quanto gratuitamente[6].

Tendo discutido alguns pontos iniciais relativos ao possível duplo estatuto da LC, reporto-me agora a dois exemplos práticos de sua aplicação.

2. LC como método de investigação e LC como Linguística?

As inúmeras tarefas analíticas executadas a partir de corpus digitalizado podem ser traduzidas na extensa produção bibliográfica sobre LC nestas últimas duas décadas. Uma busca por "corpus linguistics" em www.amazon.com à época em que este artigo estava sendo escrito originou 1.087 entradas. Além desses volumes, há ainda a produção constante dos periódicos *Corpora* e *International Journal of Corpus Linguistics*.

O uso da LC como metodologia para investigação de textos eletrônicos aparece com destaque. Baker (2006) dá múltiplos exemplos de análises de discurso que partem de unidades previamente identificadas pelo analista. Semino e Short (2004) instauram o que já está conhecido como Estilística de Corpus, através de um volume que investiga a apresentação da fala e do pensamento em obras de ficção, através de dados eletrônicos.

Em termos de produção teórica sobre corpora de textos de aprendiz em todos os estágios de desenvolvimento, principalmente aqueles produzidos em inglês língua estrangeira, o uso do computador como ferramenta para olhar categorias pré-estabelecidas é muito frequente (ver GRANGER, 1998c e GRANGER et.al., 2002). Há os estudos de Aijmer (2002) sobre a modalidade; 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, p. 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 texto produzido pelo 'não nativo', incluindo-se ai os 'erros', o uso em excesso ou econômico de palavras, expressões e estruturas.

No âmbito do Brasil, por exemplo, Zyngier e Shepherd (2003) compilaram e analisaram um corpus oral no qual alunos de graduação verbalizaram apreciação sobre textos literários. Da mesma forma, Balocco et al. (2005) utilizaram a LC como metodologia para avaliar a atitude de professores de inglês (usuários proficientes) com relação a

um componente de curso de especialização recém-cursado pelos mesmos professores. Um outro exemplo prático sobre o uso do computador como metodologia de extração e tratamento de dados encontra-se em Almeida (2007), que verificou como aprendizes de língua inglesa usam o modal *can* em textos escritos, ao comparar suas preferências com as preferências de alunos cuja língua materna é o inglês. A autora extraiu dos corpora todas as instâncias de uso do modal em questão; em seguida, rotulou cada uma das ocorrências, de acordo com a função (epistêmica ou deôntica) desempenhada pelo modal. Ao fim do processo, comparou a frequência de cada uma das funções do modal *can* em cada um dos dois grupos.

Os trabalhos cujo ponto de entrada é dirigido pelo corpus propriamente dito, que são muitos, focam o léxico e suas combinações sintagmáticas descritas acima. Há trabalhos sobre colocações, coligações, prosódia e preferência semântica extraídos de corpora especialmente compilados. Entretanto, esses estudos vêm se especializando na extração e análise de grupos polilexicais ou os chamados n-gramas. Muitos trabalhos analisam como funcionam as unidades formadas por vários itens lexicais,

Em termos de blocos formados por grupos lexicais, as unidades analiticas podem incluir blocos relativamente fixos ou blocos cujos componentes podem variar. 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)[7]. Se os blocos contêm elementos que são mero resultado de programas extratores, podem ser conhecidos como 'n-gramas' (SINCLAIR, 2004b), 'agrupamentos´(*clusters)*, pacotes ou feixes lexicais (BIBER, 2004 e BIBER et al., 2004).

Segundo Scott e Tribble (2006, p. 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'.

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 postulação de Sinclair (1991, p. 109-110), verificada empiricamente por Erman e Warren (2000) de que os usuários de uma língua, em sua forma escrita ou falada, podem recorrer a conjuntos lexicais que contêm duas ou mais palavras que, por sua vez, podem ter um significado único. 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.

A grande contribuição para a LC dada pela abordagem dirigida pelo corpus foi a verificação empírica de que ao vasculharmos qualquer corpus eletrônico com programa apropriado, podemos extrair agrupamentos com mais de um item lexical, os chamados n-gramas, que tendem a aparecer com regularidade em determinados corpora. Esses padrões frequentes podem fornecer evidência do principio 'idiomático', ou das unidades ouvidas/lidas e internalizadas pelos sujeitos que deram origem aos textos. Scott e Tribble (2006, p. 132) vão mais além, afirmando que um exame cuidadoso de uma lista de agrupamentos lexicais pode inclusive 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, ao darmos um exemplo prático da LC dirigida pelo corpus, ou seja, da LC como caminho para a Linguística.

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

O estudo que reportamos abaixo sobre corpora de aprendiz, fornecendo quase que um passo a passo analítico, utiliza dois corpora. O corpus de estudo, chamado Br-ICLE (*Brazilian International Corpus of Learner English*)[8] é formado de 127 composições argumentativas escritas por universitários brasileiros, aprendizes de língua inglesa em nível avançado, cursando do quinto período 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 e se o sujeito da pesquisa usou ou não material de consulta. 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, é considerado pequeno segundo os parâmetros postulados por Berber Sardinha (2004, p. 26).

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.

O estudo usa a abordagem dirigida pelo corpus, isto é, não lança mão de categorias linguísticas pré-estabelecidas para confirmação de hipóteses. Aliás, no início dessa pesquisa, pouco ou nada se sabia em relação à população de estudo e seus hábitos de escrita. O estudo segue os preceitos de Scott e Tribble (2006) para a análise de corpora de aprendiz, ou seja 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 erros de toda a espécie.

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 palavras-chave. Os autores alegam que um exame detalhado dessas listas ajuda a iluminar não só a preferência por determinados itens lexicais por parte de 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 dos 22[9] itens mais frequentes dos dois corpora, listado no Quadro 1 acima, evidencia uma coincidência de itens como

artigos definidos, pronomes pessoais demonstrativos e preposições, formas frequentes na língua inglesa em geral. Entretanto, chama a atenção o item *people*, usado no BrICLE quase duas vezes mais 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.

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 (não mostradas aqui por razão de espaço) que os sujeitos do Br-ICLE usam o item com sentido indeterminado. Fica também claro que os colocados mais frequentes da palavra 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 são verbos lexicais (*people believe*, *people do not/have*). Quando há elementos modificadores para *people*, esses consistem de adjetivos quantificadores 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*), em todos os casos havendo uma tentativa de colocar ´*people*´em um compartimento. Diferentemente dos sujeitos do BrICLE, que não usam adjetivação para *people*, os sujeitos do LOCNESS tendem a caracterizar quem são essas ´pessoas´ (people) a que se referem nos *essays*.

Quando se fala na quantidade de 'pessoas', o leque de opções feitas no LOCNESS é bem específico em termos de coligação[10] 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[11], 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 de processos mentais e verbais à direita de *people,* (a única escolha é *people think*) não permite aos sujeitos brasileiros a opção de trazerem ao discurso opiniões outras além de suas próprias.

A investigação de listas de itens lexicais 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 para *this* e *that* 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 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, outra vez ocorrem os processos verbais e mentais mencionados acima. Somente há padrões no corpus brasileiro com *believe*; *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. *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 talvez compensada pelo uso de *I think* , um bigrama frequente nesse corpus.

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, p. 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 no mesmos textos.

BR	Word	Freq.	%	Locness	Word	Freq.	%
1	IN ORDER TO	79	0,12	1	THE FACT THAT	162	0,05
2	THE FACT THAT	35	0,05	2	IN ORDER TO	130	0,04
3	IT IS NOT	34	0,05	3	ONE OF THE	123	0,04
4	AS WELL AS	33	0,05	4	THAT IT IS	105	0,03
5	ON THE OTHER	32	0,05	5	BE ABLE TO	94	0,03
6	ONE OF THE	28	0,04	6	THERE IS NO	94	0,03
7	THE OTHER HAND	28	0,04	7	THE RIGHT TO	85	0,03
8	THERE IS NO	25	0,04	8	IT IS NOT	84	0,03
9	*THEY DO NOT*	*25*	*0,04*	9	DUE TO THE	82	0,03
10	THE END OF	22	0,03	10	THE END OF	82	0,03
11	IT IS A	21	0,03	11	BECAUSE OF THE	80	0,02
12	*MORE AND MORE*	*19*	*0,03*	12	THERE IS A	78	0,02
13	THE NUMBER OF	19	0,03	13	THE IDEA OF	77	0,02
14	*THE ONES WHO*	*19*	*0,03*	14	AS WELL AS	76	0,02
15	BE ABLE TO	18	0,03	15	END OF THE	70	0,02
16	*IN OTHER WORDS*	*18*	*0,03*	16	IT IS A	70	0,02
17	THERE IS A	18	0,03	17	THE USE OF	69	0,02
18	*AT THE SAME*	*17*	*0,03*	18	THIS IS A	68	0,02
19	*IT IS POSSIBLE*	*17*	*0,03*	19	SHOULD NOT BE	66	0,02
20	*OF THE WORLD*	*17*	*0,03*	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, procede-se a 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 frequentes 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 as vezes de palavras-chave) com um estudo cuidadoso dos textos de onde elas foram extraídas" (SCOTT E TRIBBLE, 2006, p. 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 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[12] 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.

Por fim, o exemplo detalhado fornecido acima sugere que a) pode-se abordar um corpus sem idéias pré-formuladas ou categorias gramaticais (ou discursivas) pré-estabelecidas; b) esse tipo de abordagem dirigida pelo corpus é de orientação lexical; c) as extrações de n-gramas (meros agrupamentos encontrados pelo computador) podem revelar muitos insights.

4. Conclusão

O presente trabalho começou com um levantamento, ainda que pequeno, do dilema identitário enfrentado pelos linguistas de corpus por transitarem em arena que é, por vezes, vista como 'mera' metodologia e, por outras, reconhecida como sendo área da Linguística com objeto de investigação próprio, as relações lexicais de natureza sintagmática. O artigo tentou sugerir que é infrutífera a discussão, ao elencar estudos que tanto partem da Linguística para o corpus, como fazem o percurso invertido, ambos tendo feito contribuições inegavelmente valiosas para a Linguística. De forma compatível com todos os avanços tecnológicos (como o microscópio ou o telescópio eletrônico), muitas vezes a nova tecnologia instaura um novo *status quo* para um campo científico: com o computador e programas extratores podemos ver mais coisas, mais de perto, além de podermos retificar aquilo que percebemos anteriormente de forma obscurecida. Com a LC é possível, portanto reescrever as descrições existentes para a linguagem de forma mais clara ou elaborar novas descrições.

No Brasil, a despeito dos esforços pontuais de alguns, os trabalhos com corpus digitalizado em geral ainda são poucos, o que torna a inclusão do capítulo sobre LC neste volume da Matraga especialmente auspiciosa.

Recebido em 15/04/09
Aprovado em 07/05/09

ABSTRACT

The present paper problematizes empirical research based on corpus, in the light of the role played by Corpus Linguistics. The article opens with a discussion of the place of Corpus Linguistics itself, i.e., whether it can be considered part of Linguistics proper or whether it is no more than a methodology utilizing the computer in the investigation of linguistic phenomena. Without favoring either position, the discussion focuses on practical examples of corpus-based and corpus-driven approaches carried out in recent studies of both Portuguese and English as a Foreign Language. The work argues in this way that the status of a "branch of Linguistics" or linguistics "data mining methodology" may be established at the onset of the analysis of digital data. Whichever the role opted for, the results yielded make important contributions to the ways we view linguistic phenomena.

KEYWORDS: Corpus Linguistics; corpus-driven approach; corpus-based approach.

REFERÊNCIAS

AIJMER, K. *English discourse particles*: evidence from a corpus. Amsterdam: John Benjamins, 2002. 299p.

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.

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. p.37-54.

BAKER, P. *Using corpora in discourse analysis*. London: Continuum. 198p.

BALOCCO, A.E.; CARVALHO, G.; SHEPHERD, T.M.G. What teachers say when they write or talk about discourse analysis In: BARTELS, N (ed.), *Applied linguistics and language teacher education*. New York: Springer, 2005, p.119-134.

BERBER SARDINHA, T. A. *Linguística de corpus*. São Paulo: Manole, 2004. 410p.

_____; SHEPHERD, T. M. G. 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.

BERNARDINI, S; STEWART, D; ZANETTIN, F. Introduction. In: BERNARDINI, S; STEWART, D; ZANETTIN, F. (eds.) *Corpora in translator education*. Manchester: St Jerome. 2003. p. 1-13.

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.

_____; 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.

_____; CONRAD, S.; REPPEN, R. *Corpus linguistics: investigating language structure and use*. Cambridge: Cambridge University Press. 1998. 300p.

_____; JOHANSSON, S; LEECH, G.; CONRAD, S.; FINEGAN, E.; *Longman grammar of spoken and written English*. London: Longman, 1999.

DE COCK, S. et al. An automated approach to the phrasicon of EFL learners. In: GRANGER, S. (ed.). *Learner English on computer*. London: Longman, 1998. p. 67-79.

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, p. 29-62, 2000.

FIRTH, J R. Modes of meaning. *Essays and studies. The English association, 118-149*, 1957.

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.

_____ 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.

_____ (ed.). *Learner English on computer*. London: Longman, 1998c. 228p

_____; HUNG, J. PETCH-TYSON, S. (eds.) *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam: John Benjamins, 2002. 257p.

HOEY, M. *Lexical priming*. London: Routledge, 2005. 202p.

HYLAND, K. As can be seen: lexical bundles and disciplinary variation. *English for Specific Purposes*. 27, p. 4-21, 2008.

KENNEDY, G. *An introduction to corpus linguistics*. London: Longman. 1998. 309p.

KRISHNAMURTHY, R. (ed.) *English collocation studies: the OSTI report*. London: Continuum, 2004. 208p.

MCENERY, A.; WILSON; A.; *Corpus linguistics*. Edinburgh: Edinburgh University Press.

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.

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 adavanced learner English: a cross-linguistic approach. In.: GRANGER, S. (ed.) *Learner English on computer*. London: Longman, 1998. 228p.

SCOTT, M. *Wordsmith Tools*. Oxford: Oxford University Press, 1999.

_____; TRIBBLE, C. (2006). *Textual patterns: keywords and corpus analysis in language education*. Amsterdam: John Benjamins, 2006. 214p.

SEMINO, E.; SHORT, M. *Corpus stylistics*. London: Routledge. 256p.

SHEPHERD, T.M.G; ZYNGIER, S.; VIANA, V. A tale of two cities: lexical bundles as indicators of linguistic choices and socio-cultural traces In: JEFFRIES, L. ; MCINTYRE, D.; BOUSFIELD, D.(eds.). *Stylistics and social cognition*. Amsterdam: Rodopi, 2007. p.216-235.

SINCLAIR, J. *Corpus, concordance, collocation*. Oxford: Oxford University Press, 1991. 197p.

_____. Interview. In.: KRISHNAMURTHY, R. (ed.) *English collocation studies: the OSTI report*. London: Continuum, 2004a. 208p.

_____. Preface. In: LEWANDOWSKA-TOMASZCZYK, B. (ed.). *Practical applications in language and computers*. Frankfurt: Peter Lang. 2004b. p. 7-11.

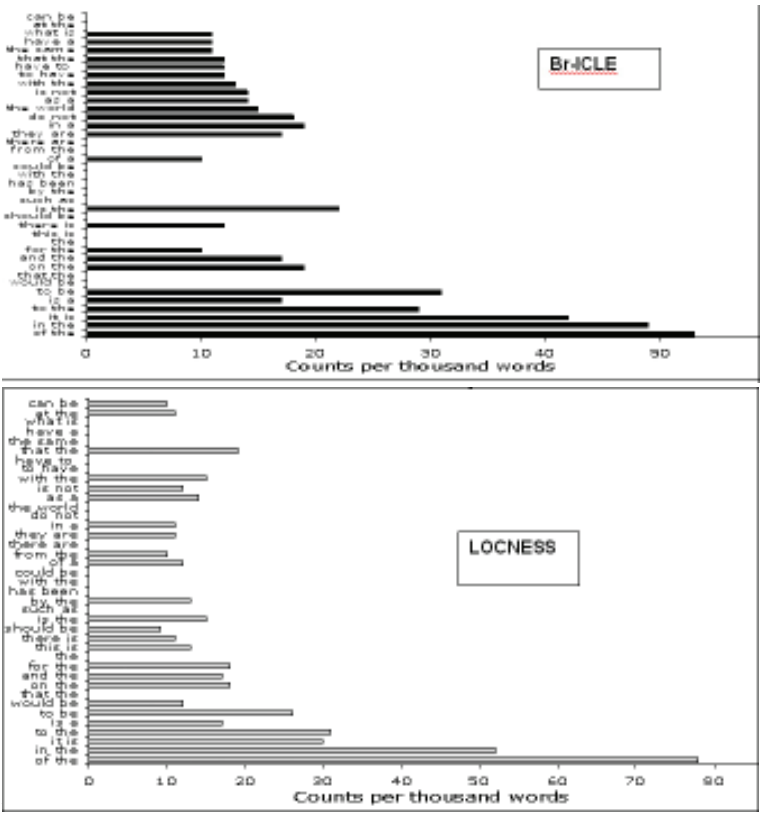
TEUBERT, W. Language and corpus linguistics. In: TEUBERT, W.; CERMAKOVA, A. (eds.) *Corpus linguistics: a short introduction*. London: Continuum. 2004. P. 1-58.

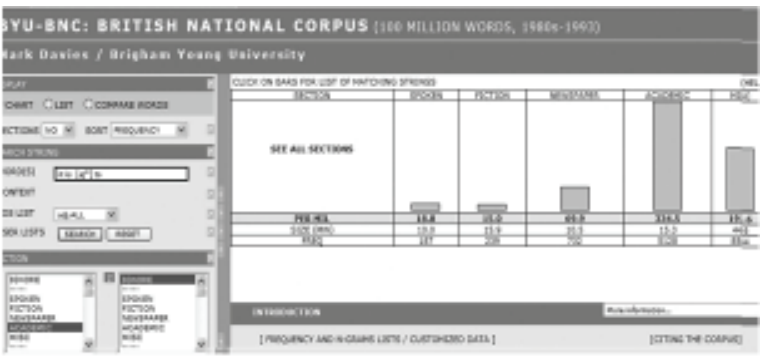
TOGNINI BONELLI, E. *Corpus linguistics at work*. Amsterdam: John Benjamins, 2001. 223p.

WRAY, A. Formulaic language in learners and native speakers. *Applied Linguistics* (32), p. 213-231, 1999.

_____. *Formulaic language and the lexicon*. Cambridge: Cambridge University Press, 2002, 332p.

ZYNGIER, Sonia; SHEPHERD, T. M. G., What is literature, really? A corpus-driven study of students' statements. *Style*, v.37, p.14 - 26, 2003.





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

NOTAS

[1] Ver emails contendo discussão sobre o que pensam os linguistas de corpus, os pesquisadores de linguagem natural e os linguistas aplicados em http://www.ling.lancs.ac.uk/groups/crg/files/CRG_discussion_prompts_w3.pdf

[2] Essa e as demais traduções do inglês são de minha responsabilidade. O texto original é "Is Corpus Linguistics a branch of Linguistics? The answer to this question is both yes and no." (McEnery e Wilson (1996: 2).

[3] "...corpus linguistics is now firmly established as a research area and a methodology." (Bernardini et al., 2003: 1)

[4] "You shall know a word by the company it keeps" (Firth, J. R. 1957, p.11).

[5] "corpus evidence was ignored, spurned and talked out of relevance until its importance became too obvious for it to be kept out in the cold". Sinclair (2004, p. 2)

[6] Como exemplo podem ser citadas três iniciativas de compilação de corpora de língua portuguesa, feitas por 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 1 bilhão de palavras (ver http://www2.lael.pucsp.br/corpora para maiores informações).

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

[8] 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, sendo preservados erros de ortografia.

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

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

[11] O termo 'processo' é usado aqui no sentido a ele atribuído pela gramática sistêmico-funcional de Halliday.

[12] Este enquadramento faz parte dos quadrigramas mais frequentes no subcorpus de línguagem acadêmica do British National Corpus.(ver http://site.ebray.com/pub/benjamins/docDetail.action?docID=10126062&P00=scott%20tribb1e)