the ESPecialist, vol. 36, nº 2 (241-262) 2015 ISSN 2318-7115

ESCRITA ACADÊMICA: UM ESTUDO EXPLORATÓRIO DE QUADRIGRAMAS

Academic Writing: An Exploratory Study of Quadrigrams

Patrícia P. BÉRTOLI (Universidade do Estado do Rio de Janeiro, Brasil)
Tania M. G. SHEPHERD (Universidade do Estado do Rio de Janeiro, Brasil)

Resumo:

Este estudo investiga uma seleção dos quadrigramas-chave de um corpus de escrita de aprendizes de Inglês como Língua Estrangeira (doravante ILE). O corpus foi compilado a partir dos trabalhos prescritos a alunos de Graduação cursando a disciplina 'Ensaio Acadêmico'. A análise levanta a natureza dos quadrigramas em termos de frequência e examina seus entornos. Observa-se que os aprendizes utilizam, coletivamente, determinados quadrigramas em quantidades acima daquelas usadas na escrita acadêmica de nativos. Além disso, empregam os quadrigramas em cotextos inadequados.

Palavras-chave: Quadrigrama-chave, Escrita Acadêmica, *Corpus* de Aprendiz, Inglês como Língua Estrangeira.

Abstract:

This study investigates a selection of key quadrigrams in a learner corpus of English as a foreign language. The corpus consists of written academic essays submitted by undergraduates as assigned homework. The analysis unveils the nature of the quadrigrams in terms of frequency, in addition to examining their cotext. It is argued that, collectively, the novice writers investigated use certain quadrigrams more often than their native counterparts. In addition, they use the same quadrigrams in inadequate cotexts.

Keywords: Academic Writing, Key Quadrigrams, Learner *Corpus*, English as a Foreign Language.

1- Introdução

Em 1997, Leech defendia uma sinergia natural entre Ensino e *corpora*, sendo três os pontos de convergência. Primeiro, coletâneas de textos digitalizados poderiam ser usadas de forma indireta na criação não só de livros didáticos e obras de referência, como também em gramáticas e dicionários, além de materiais de testagem. O segundo ponto de convergência seria o uso pedagógico e direto de *corpora* por linguistas aplicados/Professores com o objetivo de ensinar sobre *corpora* e explorar *corpora* para ensinar. O terceiro ponto incluiria a compilação profissional de *corpora* a partir dos trabalhos produzidos por aprendizes (de Língua Materna e adicional) e línguas para fins específicos. Um ano depois, em 1998, Leech defenderia também a compilação e exploração de *corpora* de aprendiz pelo Professor de Língua Estrangeira (doravante LE), com o propósito de estudar o processo de aquisição da linguagem de seus alunos.

Em 2004, falando a brasileiros, Berber Sardinha chamava a atenção para o fato de que além de estudar a aquisição propriamente dita, o Professor/pesquisador de LE poderia usar *corpora* de aprendizes para:

> 1. Descrição “da linguagem nativa
> 2. descrição da linguagem do aprendiz
> 3. transposição de metodologia de pesquisa acadêmica para a sala de aula
> 4. desenvolvimento de materiais de ensino, currículos e abordagens” (2004, p. 254-255).

O presente trabalho se concentra nos itens 1, 2 e 3 acima e faz algumas reflexões, ainda que breves, sobre o item 4. Parte-se do princípio que a linguagem do aprendiz não é monolítica, mas sim mutante e variável (ELLIS; BARKHUISEN, 2005, p. 4) e, dessa forma, compilar *corpus* de aprendiz implica capturar um momento do processo de aprendizagem. Este trabalho analisa um desses momentos ao investigar um *corpus* de escrita acadêmica de brasileiros do primeiro período de Inglês/Literaturas. O objetivo é extrair o léxico usado por esses alunos de Inglês como Língua Estrangeira (doravante ILE) nesse estágio de sua vida acadêmica. O trabalho utiliza um *corpus* de referência compatível, contendo ensaios de nativos da LE. Para a análise de ambos os *corpora*, o estudo faz uso de pesquisa consagrada sobre n-gramas e, por fim, informa à Professora da classe sobre um aspecto do aprendizado de

LE do grupo: o uso de vocabulário frequente por escritores iniciantes ao lidarem com as demandas de atividades de escrita acadêmica dentro da universidade.

Segundo Hyland (2009, p. 123), o discurso do aluno – e, principalmente, sua escrita - está no centro das práticas pedagógicas de ensino e aprendizagem que ocorrem na Educação Terciária. Isso se dá por duas razões: primeiro porque os conteúdos das disciplinas são acessados por meio do texto escrito; segundo, porque, geralmente, é a partir do texto escrito que o aluno mostra aquilo que aprendeu. É por essa condição de simbiose que o discurso acadêmico vem recebendo tanta atenção por parte de pesquisadores/Professores nos últimos anos.

A situação para o aprendiz, entretanto é nada encorajadora. Para alunos que acabaram de entrar na universidade aprender a escrever um ensaio acadêmico significa adquirir um letramento específico, o assim chamado "*essayist literacy*" (SCOLLON; SCOLLON, 1981). O processo de aquisição é lento, principalmente, se o aprendiz em questão não perceber que tal letramento dar-se-á por "meios de entender e discutir o mundo" (HYLAND, 2009, p. 124). Os desafios da escrita acadêmica aumentam, exponencialmente, quando o ensaio tem de ser escrito em LE. Os problemas enfrentados são inúmeros, os quais vão desde entender como empregar os padrões retóricos da LE, até o conhecimento da língua em uso pela comunidade de prática que escreve nessa LE.

No caso particular do aluno de ILE, a lista de livros didáticos recentes que se propõem a ensinar 'como' escrever um ensaio acadêmico é infindável, mas parece não ajudar muito. Há livros em inglês que têm como alvo um estudante de graduação estrangeiro sem especificação de curso. Essas publicações descrevem o passo a passo, nem sempre claro, de como transformar parágrafos em ensaios (cf. LISS; DAVIS, 2012). Há outro tipo de publicação, geralmente, direcionada ao pós-graduando que necessita escrever artigos para revistas científicas. Esse tipo de livro fornece expressões e frases úteis para a escrita das várias partes componentes de um artigo acadêmico (cf. GLASMAN-DEAL, 2010; SWALES; FEAK, 2012). Dessa forma, ficam à deriva um grupo enorme de alunos de outras disciplinas, bem como o Professor de ILE que se propõe a ensinar o ensaio acadêmico.

O propósito do presente trabalho é analisar o ensaio acadêmico, produto de uma dessas populações esquecidas, o aluno de Letras, aprendiz de ILE. Diferentemente, de outros estudos relevantes existentes no Brasil, o presente estudo visa informar o próprio Professor da classe de ILE sobre as áreas lexicais nevrálgicas de seus alunos. Na tradição de um trabalho na área de Linguística de Corpus, este é um trabalho sobre o léxico de natureza contrastiva, porque

compara *corpora* de ensaios acadêmicos em Inglês escritos por nativos e por aprendizes brasileiros. A proposta aqui é investigar sequências polilexicais frequentes nos dois *corpora*, respondendo às seguintes perguntas de pesquisa:

1- Quais os quadrigramas – chave presentes no corpus de estudo?
2- Como esses itens se configuram em termos de frequência e de uso em ambos os *corpora* estudados?
3- Que cotextos envolvem esses quadrigramas?

Ao responder a essas perguntas focando em termos-chave da escrita acadêmica de universitários brasileiros, pretende-se aqui preencher uma lacuna nos estudos descritivos sobre esse tópico e adicionar os *insights* encontrados neste trabalho.

2. Revisão da Literatura

A pesquisa da escrita acadêmica em Inglês tem sido feita sob duas perspectivas: a Análise do Discurso e a Linguística de *Corpus* (CHARLES *et al*, 2009, p. 1). Tais perspectivas, ainda que distantes, metodologicamente, compartilham o interesse pela linguagem em uso e por sua padronização. Além disso, ambas tendem a relacionar seus achados às práticas sociais, intelectuais ou ideológicas. A Análise do Discurso prioriza textos e se caracteriza como uma abordagem *top-down*. Por outro lado, a Linguística de *Corpus* tem seu ponto de partida na investigação de palavras e frases em dados provenientes de muitos textos em sua forma digital ou digitalizada. Via de regra, os estudos sobre *corpora* digitais mapeiam frequência e distribuição de uma variedade de fenômenos linguísticos e prestam-se à comparação entre registros, gêneros e disciplinas.

Os estudos sobre *corpora* de escrita acadêmica, de forma geral, lidam com tendências de uso de itens lexicais específicos ou de agrupamentos lexicais e tendências fraseológicas. Em termos de agrupamentos lexicais as pesquisas podem se concentrar em grupos lexicais não contínuos (como em *in + substantivo + of*) ou com os assim chamados n-gramas. Segundo Scott e Tribble (2006, p. 131), um feixe lexical[1] (ou n-grama) nada mais é do que um produto artificial oriundo de programas extratores, que existem com base em critérios,

[1] Feixes lexicais contínuos, conhecidos também como pacotes lexicais ou n-gramas, "*clusters*", "*chunks*" e "*bundles*", são unidades extraídas por computador, que aparecem mais, frequentemente, em determinados *corpora*.

puramente, distributivos. Em outras palavras, 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 ‘n-grama’ ou ‘feixe lexical’.

Dois estudos de porte sobre linguagem acadêmica foram executados por Biber (2004) e Biber *et al.* (2004) para a extração de feixes lexicais. Esses estudos trabalham com *subcorpora* compostos de vários registros usados na academia (livros didáticos, aulas, escrita científica) extraindo deles feixes lexicais e desenvolvendo uma sistemática de classificação estrutural e funcional para esses feixes. Ao comparar a distribuição e frequência dos vários feixes, os autores puderam mapear em cada um dos registros estudados a configuração dos feixes. Além disso, por meio da comparação da composição e função de cada agrupamento, estabeleceram características lexicais inerentes a cada registro investigado.

Outras análises de fraseologias na linguagem acadêmica têm fornecido prova irrefutável de que padrões não só mostram diferenças entre registros, como também são fator que diferencia a escrita nas várias disciplinas acadêmicas (HYLAND, 2008a e 2008b). O autor partiu da compilação de quatro *corpora* originários de disciplinas acadêmicas distintas e extraiu os *clusters* das quatro palavras mais frequentes de cada um dos *corpora*. Seus achados de pesquisa evidenciaram que feixes lexicais ou *clusters* são um modo eficiente de se mapear diferenças entre gêneros nas diversas disciplinas e também um excelente modo de caracterizar os contrastes entre a linguagem acadêmica de autores experientes e de alunos cuja Língua Materna não é o Inglês.

A investigação assistida por computador da escrita acadêmica de alunos brasileiros já tem produção, igualmente, numerosa. A pesquisa indutiva de Shepherd (2009) focou em feixes de três e quatro palavras (trigramas e quadrigramas) na produção de brasileiros oriundos de várias universidades, comparando o *corpus Brazilian International Corpus of Learner English* (doravante Br-Icle) e o *Louvain Corpus of Native English Essays* (doravante LOCNESS). O estudo concluiu que a população brasileira investigada não parece ter alternativas para as palavras ditas vagas como ‘*people’* e não demonstra conhecer as possibilidades anafóricas do pronome ‘*this*’. Além disso, parece desconhecer como usar, anaforicamente, substantivos abstratos *(content nouns)*, como ‘problema’, ‘solução’, ‘consequência’ e outros. Diferente estudo importante foi o de Dutra e Berber Sardinha (2013), que investigou n-gramas a partir também do *corpus* Br-ICLE, agora aumentado, do *corpus* LOCNESS e do *International Corpus of Learner English* (ICLE). Os pesquisadores extraíram n-gramas e aplicaram a eles as categorias funcionais de ‘referencial’ (*referential)*,

‘de posicionamento’ (*stance*) e ‘discursivo’ (*discourse*) propostas por Simpson-Vlach e Ellis (2010). Mostraram com seu estudo, por meio de dados estatísticos que os n-gramas referenciais são os mais frequentes no *corpus* de aprendizes brasileiros estudado. Em 2014, Dutra *et al* caminharam com a pesquisa, focando no segundo grupo de n-gramas mas frequente de seu estudo anterior, os chamados n-gramas de posicionamento. O objetivo dessa vez foi discutir o papel ocupado por essas expressões em escrita acadêmica em ILE. Os resultados apontam para uma ausência de marcadores de *hedging* (atenuadores) na escrita acadêmica examinada.

Com relação à análise de padronização em *corpora* de aprendizes de ILE (falantes de diversas línguas que não o Português Brasileiro), em contraste com usuários cuja Língua Materna é o Inglês, há algumas conclusões interessantes. Notam-se diferenças na frequência com que os aprendizes usam a mais ou a menos determinadas expressões. Esses estudos (ver GRANGER *et al*, 2013 e SALAZAR, 2014, para uma revisão detalhada) revelam que, independentemente, da Língua Materna, os aprendizes tendem a:

> a) Usar combinações formadas das palavras menos frequentes na LE (DURRANT; SCHMITT, 2009; LAUFER; WALDMAN, 2011);
> b) Usar em excesso um pequeno conjunto correto de colocações de alta frequência, as quais Nesselhauf (2005, p. 69) denominou "ursos de pelúcia colocacional";
> c) Usar, indevidamente, colocações e coligações típicas na LE (PAQUOT; GRANGER, 2012; NESSELHAUF, 2005), parecendo desconhecer critérios de aceitabilidade, sendo que os usos indevidos mais frequentes são as escolhas de verbos, substantivos e preposições;
> d) Optar por usar colocações ‘seguras’ (PAQUOT; GRANGER, 2012), evitando estruturas que não são típicas de suas Línguas Maternas;
> e) Repetir, excessivamente, os mesmos n-gramas, quando comparados com nativos. Os alunos menos proficientes revelam um repertório lexical próprio, correto, mas, extremamente, restrito (PAQUOT; GRANGER, 2012).

Conhecendo essas conclusões já elaboradas sobre aprendizes de Línguas Maternas diversas do Português, o presente estudo também se propôs a investigar aspectos delineados nos itens listados acima, mas controlando as variáveis o mais que possível: o *corpus* foi coletado numa única universidade, por uma mesma Professora de classe a partir de trabalhos de alunos cujo perfil socioeconômico é semelhante. Além disso, as condições de elaboração dos textos foram as mesmas para todos os alunos, visto que todos os textos foram elaborados fora da universidade e sem controle de tempo para execução. Escolheu-se trabalhar com

expressões-chave a fim de revelar as cadeias de repetição que são, estatisticamente, significativas e que mais caracterizam os textos dos aprendizes.

3. Materiais e Métodos

A quantidade de *corpora* e de unidades fraseológicas em discurso acadêmico é tal que Stubbs (2007) diz, textualmente, que[2] "a única estratégia realista é começar pequeno: usar uma amostra restrita para gerar hipóteses que podem ser testadas em amostras maiores " (p. 92).

Nosso estudo começa pequeno, mas tem uma enorme representatividade, como já explicitado acima. Para a execução do presente trabalho foram necessários dois *corpora*: um *corpus* de estudo e um *corpus* de referência. O *corpus* de estudo foi coletado, especialmente, para esta pesquisa e foi denominado *Corpus* de Ensaios Acadêmicos de Alunos de Letras (EAAL). Como *corpus* de referência foi utilizado *The British Academic Written English* (BAWE) *corpus*.

O *corpus* EAAL é composto por produções acadêmicas de alunos do curso de Letras de uma universidade pública na cidade do Rio de Janeiro. Os alunos fazem Inglês-Literaturas de Língua Inglesa, curso de Licenciatura única que, dessa forma, tem um número de horas de contato com a LE superior ao de cursos com licenciatura dupla. Trata-se de uma coleção de 174 textos produzidos por noventa e dois estudantes de primeiro ano, a partir de tarefas propostas por uma única Professora. A coleta foi feita com a aquiescência por escrito dos alunos, no período entre setembro de 2013 e dezembro de 2014, ou seja, três semestres letivos. Pediu-se que os textos fossem escritos de acordo com os parâmetros de ensaios acadêmicos ensinados em classe, isso é, contendo quatro ou cinco parágrafos, sendo um introdutório e um para conclusão, intermediados por parágrafos argumentativos. Os temas dessas tarefas incluíram ‘cultura inglesa’, ‘inglês como língua franca’, ‘gramática da língua inglesa’, ‘aprendizagem de língua estrangeira’ e o ‘adeus’ na Literatura. Foram adicionados ao *corpus* ensaios avaliativos de uma obra de Literatura de escolha individual, ensaios autobiográficos e uma narrativa recontada. Além disso, não há trabalhos dos mesmos alunos em semestres distintos, o que não influencia esta investigação por não se tratar de pesquisa longitudinal. Os textos foram produzidos como tarefa de casa, fora do horário de aula, e

[2] Nossa tradução para "the only realistic strategy is to start small: to use a restricted sample to generate hypotheses which can be tested on larger samples".

enviados por *e-mail*. Foram arquivados da mesma forma que foram recebidos, tendo-se excluído os nomes dos autores. Para compatibilizá-los com o programa utilizado (Wordsmith Tools. 5.0), foi alterado o formato para arquivo de texto (.txt), além de os títulos dos arquivos terem sido padronizados para a composição e manuseio do *corpus*. O mesmo programa foi usado para a extração de dados do *corpus* de referência.

Dos 174 textos, sessenta e um foram coletados no segundo semestre de 2013, setenta e cinco no primeiro de 2014 e trinta e oito no segundo semestre de 2014. Cada texto tem, aproximadamente, 500 palavras. O EAAL tem um total de 94.343 palavras, sendo 8.949 palavras diferentes (*Tokens*), o que o classifica como um *corpus* pequeno, segundo Berber Sardinha (2005). Todavia, sabe-se que *corpora* pequenos também se revelam interessantes em estudos linguísticos, especialmente, para contextos e registros especializados e para o ensino da escrita, como defendem Flowerdew (2001) e Tribble (2001), já que oferecem oportunidade de descobertas para aquele contexto específico, como é pertinente para as generalizações como a que se propõe esta pesquisa.

O *corpus* de referência BAWE, (disponível mediante licença de http://ota.ahds.ac.uk), contém 6.506.995 itens lexicais, advindos de 2859 trabalhos escritos, incluindo estudos de caso, resenhas, ensaios, explicações, revisão da literatura, seções de metodologia, narrativas, resoluções de problemas, propostas e relatórios de pesquisa. Os trabalhos derivam de trinta e cinco disciplinas submetidos por alunos dos três anos da Graduação e por alunos de Mestrado, incluindo-se aqueles das áreas de Letras e Humanidades.

Poder-se-ia questionar o uso de um *corpus* de referência que é sessenta e oito vezes maior do que o *corpus* de estudo. Segundo Berber Sardinha (comunicação pessoal), caso o pesquisador queira ter informação suficiente para julgar cada palavra ou grupo de palavras do *corpus* de estudo, deveria usar um *corpus* de referência muitas vezes maior do que o que *corpus* de estudo. Tal *corpus* otimizaria a oportunidade de se visualizar uma gama maior de ocorrências do léxico em foco. Um *corpus* de referência pequeno abriria a chance de se acharem ‘falsas’ palavras-chave. Por exemplo, uma palavra que ocorra muitas vezes no *corpus* de estudo e nenhuma no de referência seria considerada chave. Entretanto, se o *corpus* de referência não for, numericamente, representativo, haveria dúvida se a palavra é chave ou se a ocorrência foi acidental devido a problemas de amostragem do *corpus* de referência.

Uma vez que o *corpus* de estudo foi construído e o *corpus* de referência decidido, foram levantados quadrigramas de cada *corpus*. A partir de um ponto de corte de no mínimo

cinco ocorrências, foram obtidos noventa e um quadrigramas do *corpus* EAAL. O quadro abaixo ilustra os dez quadrigramas mais frequentes desse *corpus*.

Figura 1: Quadrigramas mais frequentes no *corpus* EAAL.

N	Quadrigramas	Fre	%	Texto	%
1	on the other hand	34	0.036965061	34	19.31818199
2	it is possible to	29	0.030628195	24	13.63636398
3	one of the most	23	0.024291327	22	12.5
4	the end of the	18	0.019010603	17	9.659090996
5	is one of the	15	0.01584217	14	7.954545498
6	the beginning of the	15	0.01584217	14	7.954545498
7	*fault in our stars*	13	0.013729881	3	1.704545498
8	*learning a second language*	13	0.013729881	8	4.545454502
9	*the fault in our*	13	0.013729881	3	1.704545498
10	at the end of	10	0.010561447	10	5.681818008

Fonte: Elaborado pelas autoras.

Na lista acima os quadrigramas nas posições: sete, oito e nove, estão associados aos temas das tarefas desenvolvidas pelos alunos. Os itens sete e nove referem-se ao livro *The fault in our stars* (A culpa é das Estrelas), que foi escolhido por três alunos diferentes. Isso gerou treze ocorrências de quadrigramas como "*fault in our stars*" e "*The fault in our*". Já o item oito, *learning a second language*, refere-se a um dos temas propostos para as redações, tendo sido usado também treze vezes por oito alunos diferentes.

Do *corpus* BAWE foram extraídos 1997 quadrigramas, tendo como ponto de corte dez ocorrências mínimas. Os quadrigramas dos dois *corpora* foram computados com o auxílio do *software* WordSmith Tools 5.0.

A decisão de analisar elementos chave no *corpus* de estudo se deveu a uma escolha metodológica. Palavras ou expressões-chave são obtidas por meio da comparação de um *corpus* de estudo (menor) e um *corpus* de referência (maior) e representam as palavras cujas frequências são, estatisticamente, superiores (positivas) ou inferiores (negativas) no *corpus* de estudo (cf. BERBER SARDINHA, 2004, p. 111). Por essa razão, sinalizam alguma característica inerente a um conjunto de textos. Trabalhar com itens lexicais-chave como elementos de busca pode ajudar na classificação de textos e registros, mas também servem de ferramentas analíticas do discurso e da interpretação dos mesmos textos (BONDI, 2010, p. 1).

Com relação à opção por analisar quadrigramas ao invés de trigramas, esse parece ser um ponto de partida já consagrado no estudo de diferenças entre escrita de aprendizes e

autores mais proficientes. Cortes (2004, p. 401) explica que a maioria de quadrigramas contém trigramas em sua estrutura e por isso os quadrigramas são uma boa estratégia para eliminar a superposição de expressões do processo analítico.

Por causa do escopo deste trabalho, foram observados os quadrigramas-chave positivos somente. A lista dos quinze mais frequentes está abaixo e a lista completa pode ser encontrada no Apêndice deste artigo:

Figura 2: Quadrigramas-chave positivos no *corpus* EAAL.

n	**quadrigramas**	**EAAL**
1	on the other hand	34
2	it is possible to	29
3	one of the most	23
4	the beginning of the	15
5	fault in our stars	13
6	learning a second language	13
7	the fault in our	13
8	tells the story of	10
9	the tell tale heart	10
10	a walk to remember	9
11	learning a foreign language	9
12	end of the book	9
13	a second language is	8
14	as a second language	7
15	in the end of	7

Fonte: Elaborado pelas autoras.

Os títulos das obras escolhidas pelos alunos para suas redações representam quadrigramas mais frequentes no EAAL em relação ao BAWE: *Fault in our stars*, *The tell tale heart* e *A walk to remember*. Os seguintes quadrigramas estão, diretamente, relacionados com os temas e tipos de tarefas designadas aos alunos: *the beginning of the, learning a second language*, *tells the story of*, *learning a foreign language*; *end of the book, a second language is* e *as a second language*.

Foram retirados da análise os quadrigramas que remetiam aos títulos dos livros ou filmes abordados pelos aprendizes. Escolhemos para este trabalho os outros quatro quadrigramas-chave mais frequentes no *corpus* EAAL: *on the other hand, it is possible to, one of the most* e *in the end of*. Isso se deu a fim de se observar a natureza da chavicidade dessas expressões: uso em excesso, uso errôneo, ou alguma outra característica não prevista.

4. Análise dos Dados

Neste ponto é apresentada a análise de cada um dos quatro quadrigramas-chave selecionados para este estudo. Para que os dados pudessem ser comparados foi necessária uma normalização dos mesmos. Ou seja, o número total de ocorrências de cada expressão de busca, por exemplo, (*on the other hand*) foi multiplicado por 100.000 e dividido pelo número total de quadrigramas de cada *corpus*. No caso do EAAL, obtivemos 1.199 quadrigramas, enquanto que no BAWE existem 464.999. A figura abaixo mostra os dados normalizados para cada uma das expressões de busca.

Figura 3: Quadro comparativo dos quadrigramas-chave.

quadrigrama	EAAL	BAWE	
On the other hand	2836	180	16 X mais
It is possible to	80	80	30 X menos
In the end of	584	4	146 X mais
One of the most	1918	77	25 X mais

Fonte: Elaborado pelas autoras.

Por meio do exame das linhas de concordâncias de cada um dos quadrigramas-chave acima, foi possível identificar se o uso excessivo de cada um deles seria o único problema a transparecer nos trabalhos dos aprendizes. Decidimos, igualmente, classificar os quadrigramas para obter uma ideia das funções das estruturas repetidas no *corpus.*

Dentre as inúmeras possibilidades de classificação funcional para quadrigramas, adotamos a de Hyland (2008) por ser sucinta, mas robusta. Para Hyland (2008) os n-gramas podem apontar para a organização do texto (*text-oriented*), para os tópicos nele discutidos (*topic-oriented*) e para as relações entre os participantes do discurso (*participant-oriented*). Seguindo essa orientação funcional, a análise se concentra em um quadrigrama que aponta para o texto (*on the other hand*), dois quadrigramas que apontam para o tópico e um quadrigrama que aponta para os participantes (*it is possible to* e *one of the most*).

On the other hand

On the other hand pode ser considerado um quadrigrama que aponta para o texto como sinal de transição discursiva, marcando contrastes. É quadrigrama-chave no *corpus* de aprendizes, ou seja, em relação ao número total de palavras do *corpus*, esse quadrigrama é usado em excesso sendo o quadrigrama mais numeroso no *corpus* BAWE. Tal informação converge com os achados de Biber (2006) e Hyland (2008a), que também revelam esse quadrigrama como sendo o mais frequente no discurso acadêmico escrito.

Ao olharmos as trinta e quatro linhas de concordância contendo a expressão *on the other hand* do *corpus* EAAL, verificamos que os padrões de uso preferidos são:

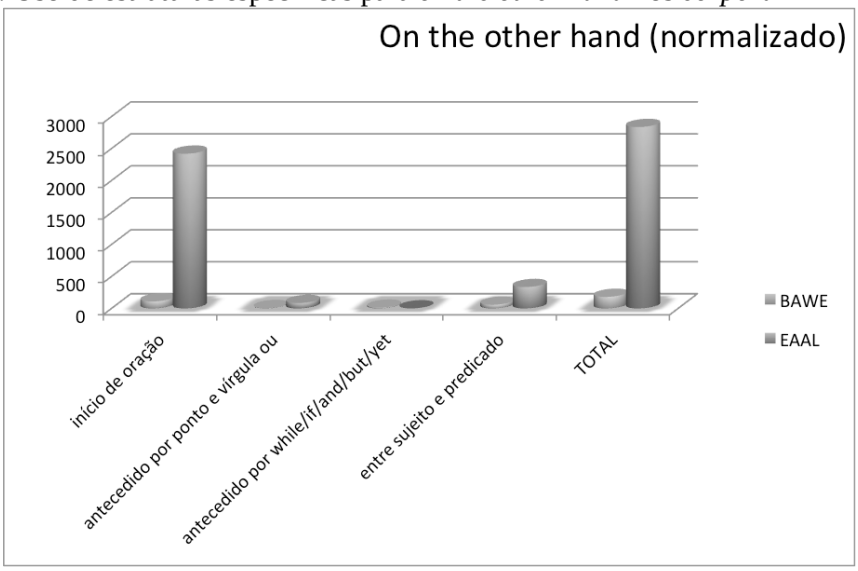
a. *On the other hand* + vírgula + oração
1. On the other hand, this may only happen in developing countries... (1-CB_W3_13.2)

b. Ponto e vírgula + *on the other hand* + oração
2. *the aspects of human relations; on the other hand she is also fragile* (2-RP_W1_13.2)

c. Sujeito + *on the other hand* + predicado
3. *The protagonist of Blue is the warmest colour, on the other hand, doesn't share the same interests…* (2-VA_W1_14.1)

O diagrama abaixo ilustra, comparativamente, os usos da expressão nos dois *corpora*.

Diagrama 1: Uso de estruturas específicas para *on the other hand* nos *corpora* EAAL e BAWE



Fonte: Elaborado pelas autoras.

O diagrama acima é autoexplicativo. Os padrões preferidos dos alunos são aqueles em que o marcador *on the other hand* introduz um contraste entre duas porções do discurso, tanto precedido por ponto ou por ponto e vírgula. Quando o contraste é feito entre dois sujeitos, e o marcador tem de ser inserido entre sujeito e predicado (ver exemplo 3 acima), o número de instâncias dos aprendizes é seis vezes maior, isso é, o padrão é usado em excesso. Mais importante ainda é a inexistência de ocorrências em que *on the other hand* inicia uma oração subordinada precedida dos conectivos *while, if, yet* e das coordenadas *and* (com o sentido de *but*) e *but*. Tal ausência não pode ser explicada pela inexistência do padrão em Português: são inúmeros os casos de 'por outro lado' precedido dos conectivos 'e' e 'mas' na Língua Materna dos aprendizes.

It is possible to

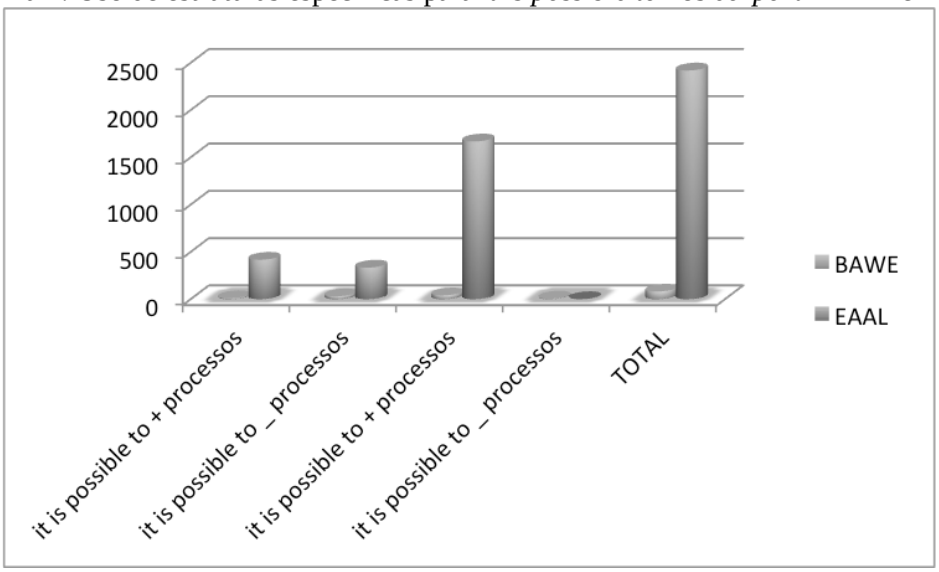
It is possible to é estruturalmente composto da partícula antecipadora *it* e um marcador de modalidade, que aponta para a opinião de um dos interactantes. Por meio de *it is possible to*, o enunciador pode asseverar sua opinião sobre a possibilidade lógica ou especulativa de algo acontecer/ ser feito. Em tese, portanto, qualquer significado verbal poderia vir após a expressão. Entretanto, para os padrões de uso da *escrita acadêmica*, as escolhas semânticas não são ilimitadas. Os padrões para *it is possible to* extraídos do BAWE se concentram em dois campos semânticos: aquilo que é possível dizer sobre um argumento, ou seja, verbos que expressam processos verbais e verbos que expressam aquilo que é possível depreender de algum argumento ou demonstração, isto é, processos mentais. Há também minimamente verbos que expressam eventos possíveis de acontecer ou de ser. Para melhor entender os dados, usamos a metalinguagem de Halliday (1985). O autor separa, semanticamente, os sintagmas verbais em processos existenciais ('haver´), relacionais (ser ou ter), mentais ('sentir', 'conhecer', 'pensar') comportamentais (verbos que descrevem comportamento humano).

Os padrões encontrados no *corpus* BAWE são os seguintes:

a. *It is possible to* + processo verbal (*argue, assert, add, say, claim, disagree, suggest, predict*);

b. *It is possible to* + processo mental (*grasp, identify, see, understand, speculate, observe, infer, know, make sense of , recognize, regard, deduce, estimate, calculate*);
c. *It is possible to* + processo material (*create, set up, send, separate, segment, manage, draw, scan, supplement, begin, match*).

Diagrama 2: Uso de estruturas específicas para *it is possible to* nos *corpora* EAAL e BAWE.



Fonte: Elaborado pelas autoras.

A análise contrastiva sobre a natureza dos processos que acompanham a expressão *it is possible to*, representada pelo diagrama acima, mostra usos excessivos nos processos: verbal, material e mental. Entretanto, o que o diagrama não captura é que em cada um desses processos os aprendizes apresentam uma limitação de vocabulário. O único verbo para a expressão de processos verbais por parte dos aprendizes é *´say´*. Quanto aos processos mentais, há uma preferência quase que preponderante de *´see´*, como sinônimo de ‘*understand*’.

Procuramos estruturas alternativas que pudessem expressar o sentido de *it is possible to*. Talvez a ausência de tais estruturas (x *may/can be argued, claimed*, etc.) na escrita dos aprendizes pudesse explicar o excesso de ocorrências de *it is possible to*. A busca mostra, entretanto, que os aprendizes usam as estruturas alternativas de forma apropriada, mas numerosa, como em:

Her character´s introspective monologues can be seen as depressive and melancholic. (DV_W1_13.2) *(*ao invés de *it is possible to see …)*

> *The same can be said about Howl (*IE_W3_13.2) (ao invés de *it is possible to say…)*
>
> *The book may be divided into 4 great moments* (AY_W1_14.1) (ao invés de *it is possible to divide).*

É somente quando são verificadas outras palavras alternativas para '*it is * to*' que os aprendizes mostram restrição vocabular. No *corpus* BAWE encontramos uma gama de itens lexicais que cumprem a função de expressão da voz do enunciador dentro do quadro *it is * to***.** São eles**:** *accurate, advisable, advantageous, applicable, analogous, appropriate, common, convenient, crucial, desirable, difficult, fair, hard, imperative, important, (im)possible, (un)likely, realistic, (un) reasonable***,** entre outros. A ausência de tal leque de possibilidades dentro do padrão pode ser explicada por desconhecimento e/ou falta de exposição às estruturas comuns do ensaio acadêmico.

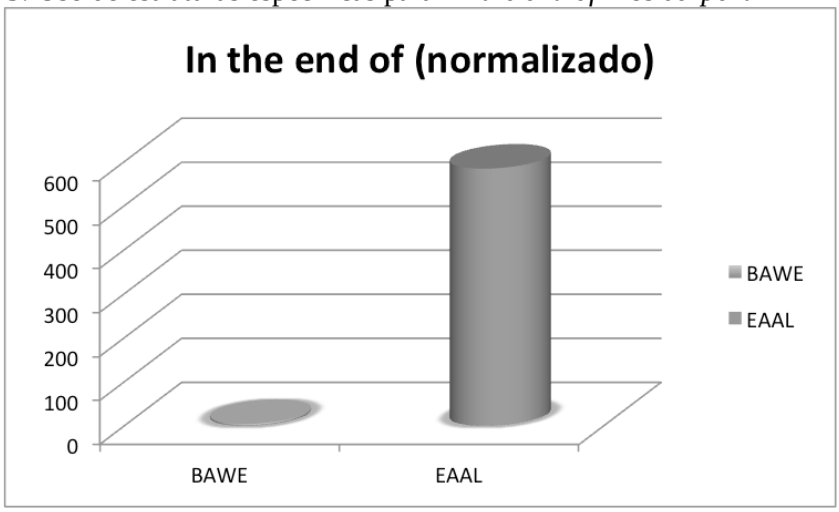
In the end of

A expressão *in the end of* é um caso claro de uso em excesso ('*overuse*') e uso errôneo *(´misuse'*). Os alunos usam-na 146 vezes mais do que no *corpus* BAWE. O cotexto revela que confundem o advérbio *in the end***,** que tem o sentido de *finally***,** com o sintagma preposicional *at the end of* (*the book, the novel, the film*). Alguns exemplos desses usos:

as matter of fact, Susan and Lucy just meet Caspian in the end of the book, therefore , Susan and Caspian does not fall each other, what happened in the movie.<1_AP_W1_14-.txt>

him. In the end of the film, Sophie breaks Howl's spell by placing his heart back in his chest, returning his complete humanity and making him finally ready to engage himself in a serious romantic relationship.<2_IE_W3_13-2.txt>

Diagrama 3: Uso de estruturas específicas para '*in the end of*' nos *corpora* EAAL e BAWE



Fonte: Elaborado pelas autoras.

One of the most

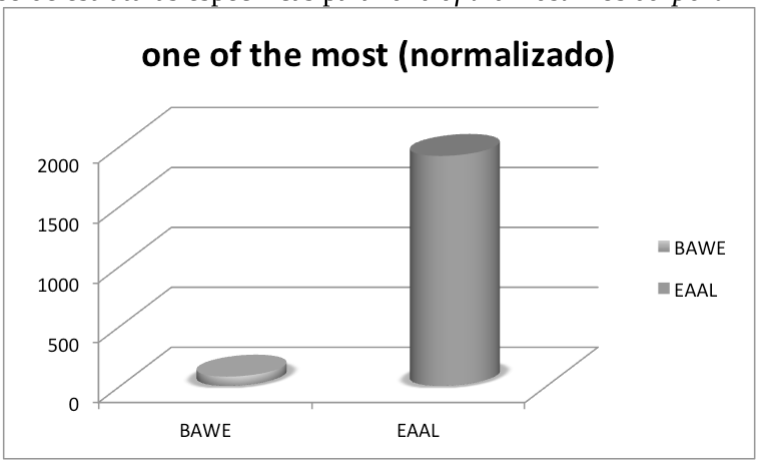
O uso da expressão *one of the most* pelos aprendizes é inadequado em dez exemplos dos vinte e três do total. Nesses dez casos, os aprendizes omitem a outra parte da expressão, ou seja, omitem o sintagma preposicional que deve acompanhá-la, por exemplo, (*one of the most * in the world; one of the most * in history*, etc).

Uma outra característica inerente ao uso de *one of the most* é que a expressão sublinha a voz do enunciador. A expressão é, geralmente, seguida de advérbio intensificador precedendo o adjetivo como em:

> ….*one of the most highly regarded translations of Julius Cesar* (2_MB_ W2_14.2*)*, o único exemplo do *corpus* EAAL.

O que acontece no BAWE, além da diversidade de adjetivos na expressão, é a tendência de os intensificar por meio de advérbios maximizadores como em *one of the most commonly (used, observed, accepted, acknowledged), one of the most extensively/highly studied, one of the most widely recognized/known; one of the most profoundly exciting; one of the most heavily criticized*.

Diagrama 4: Uso de estruturas específicas para '*one of the most*' nos *corpora* EAAL e BAWE.



Fonte: Elaborado pelas autoras.

Acima, pode-se visualizar de forma diagramática a diferença de frequência da expressão, que aparece no *corpus* EAAL vinte e cinco vezes mais.

5. Conclusões e Trabalhos Futuros

Neste trabalho fizemos uma pequena investigação dirigida pelo *corpus*, com caráter exploratório para analisar os quadrigramas-chave em um *corpus* pequeno de escrita acadêmica de aprendizes. A abordagem demonstrou que a chavicidade está contida em quadrigramas que apontam ora para o discurso, ora para a relação entre os participantes do discurso e ora são elementos de referência ao assunto. A chavicidade não reflete somente problemas de uso excessivo (*overuse*). Quando comparados com um *corpus* de referência compatível, os quadrigramas apresentam problemas outros que demandam uma investigação de seus entornos. Quando os entornos são investigados, vemos quatro explicações para o insucesso por parte do aprendiz.

O primeiro problema é que dentro das inúmeras possibilidades de configuração para essas expressões, por exemplo, (*on the other hand*) os aprendizes usam tão somente um número limitado de possibilidades combinatórias. O segundo problema é que essas expressões-chave podem se aglutinar a outros itens semânticos específicos, como em *it is possible to + processos verbais determinados*. Não se pode classificar esse problema com um

uso inadequado de prosódia semântica[3], visto que não se trata de usar prosódia negativa em espaço positivo. Não temos um nome para o fenômeno em que os aprendizes usam os quadrigramas de forma pragmaticamente certa, mas, os acoplam a outros de campos semânticos inadequados (*it is possible to notice*) para a escrita acadêmica ou, simplesmente, 'esquecem' de os completar (*one of the most* x).

Então, o que isso nos diz sobre a fraseologia da escrita acadêmica do aprendiz, aluno de Letras no início de seu curso de Graduação? Pode-se perceber que no pequeno escopo de nossa pesquisa, os achados apontam para características semelhantes às detectadas por Gilquin e Granger (2011). Em estudos com aprendizes de Inglês e de outras Línguas Maternas que não o Português, relatam o uso em excesso de escolhas seguras de colocações frequentes na LE. Pode-se inferir que ao evitar variar, muito provavelmente, a fim de não cometer o erro, os aprendizes apegam-se aos mesmos quadrigramas como a "ursos de pelúcia colocacional", ou seja, jogam sempre para ganhar, conforme apontado por Nesselhauf (2005, p. 69). No entanto, esses ursos de pelúcia nem sempre se combinam entre si. Os aprendizes optam por quadrigramas corretos seguidos de outros quadrigramas corretos, mas as combinações resultantes não são utilizadas pelo falante nativo da LE, como apontado na investigação de *it is possible to*.

Vemos os resultados desta pequena pesquisa, positivamente, tanto como pesquisadoras quanto como Professoras. Esses resultados desvelam características lexicais da escrita específicas de um grupo, podendo contribuir para a elaboração bem informada de atividades pedagógicas que venham a auxiliar o grupo a perceber suas falhas. Como exemplo, podemos citar as inúmeras atividades pedagógicas já desenvolvidas pelo grupo GELC descritas em Berber Sardinha *et al* (2012).

A despeito de seu tamanho, o *corpus* de estudo deixa aberto inúmeros caminhos de pesquisa futura, tais como investigações do uso do léxico dentro de seções específicas do ensaio acadêmico (cf. SALAZAR, 2014), como também a investigação de outros aspectos da escrita de outros aprendizes que ingressam na universidade. Ademais, possibilita a abertura de um caminho para a pesquisa longitudinal sobre aquisição de n-gramas, tão carente na área de *corpora* de aprendiz.

Recebido em: 03/2015; Aceito em 06/2015.

[3] Prosódia semântica refere-se a "associação entre itens lexicais e conotação (negativa, positiva ou neutra)" (BERBER SARDINHA, 2004, p. 40).

Referências Bibliográficas

BERBER SARDINHA, T. 2005. ***Influência do Tamanho do Corpus de Referência da Obtenção de Palavras-chave Usando o Programa Computacional Wordsmith Tools***. *The ESPecialist*, **26.**2: 183-204. São Paulo.

_____. 2004. ***Linguística de Corpus***. São Paulo, Manole.

BERBER SARDINHA, T.; T.M.G. SHEPHERD; D. DELEGÁ-LUCIO; T.L. FERREIRA (orgs.) 2012. ***Tecnologias e Mídias no ensino de inglês: o corpus nas receitas***. Cotia: Macmillan.

BIBER, D. 2006. ***University Language: A corpus-based study of spoken and written registers***. Amsterdam: John Benjamins.

BIBER, D. 2004. ***Lexical bundles in academic speech and writing***. In: B. LEWANDOWSKA-TOMASZCZYD (ed.), ***Practical applications in language corpora*** *(PALC 2003*). Hamburg: Peter Lang. pp. 165-178.

BIBER, D.; CONRAD, S.; CORTES, V. 2004. ***If you look at… : Lexical Bundles***. In: ***University Teaching and Textbooks***. *Applied Linguistics*, **25**.3: 371–405. Oxford.

BONDI, M. 2010. ***Perspectives on keywords and keyness: An introduction***. In: M. BONDI; M. SCOTT (eds.) 2010. ***Keyness in Texts***. 2010. Amsterdam: John Benjamins. pp.1-18.

CHARLES, M; PECORARI, D; HUNSTON, S. 2009 (eds.) ***Academic writing: at the interface of corpus and discourse***. London: Continuum.

CORTES, V. 2004. ***Lexical bundles in published and student disciplinary writing: Examples from history and biology***. *English for Specific Purposes*, **23**, 397–423.

DURRANT, P; SCHMITT, N. 2009. ***To what extent do native and non-native writers make use of collocations?*** *International Review of Applied Linguistics*. **47**.2: 157-177. Berlin.

DUTRA, D. P.; ORFANO, B.; BERBER SARDINHA, T. 2014. ***Stance bundles in Learner Corpora***. In: S. M. ALUISIO; S. E. O. TAGNIN (eds.) ***New Language Technologies and Linguistic Research: A Two-Way Road***. Newcastle upon Tyne: Cambridge Scholars Publishing. pp. 2-17.

DUTRA, D. P.; BERBER SARDINHA, T. 2013. ***Referential expressions in English learner argumentative writing.*** In: S. GRANGER, G. GILQUIN; F. MEUNIER (eds.) 2013. ***Twenty Years of Learner Corpus Research: Looking back, Moving ahead. Corpora and Language in Use – Proceedings 1***, Louvain-la-Neuve: Presses universitaires de Louvain. pp. 117-127.

ELLIS, R.; BARKHUIZEN, G. 2005. ***Analysing Learner Language***. Oxford: Oxford University Press.

FLOWERDEW, L. 2001. ***The exploitation of small learner corpora in EAP materials design***. In M. GHADESSY; R. ROSEBERRY (eds.) 2001. ***Small corpus studies and ELT*** . Amsterdam: John Benjamins. pp. 363–379.

GLASMAN-DEAL. H. 2010. ***Science Research Writing: A Guide for Non-Native Speakers of English***. London: Imperial College Press.

GRANGER, S; GILQUIN, G. 2011. ***From EFL to ESL: Evidence from the International Corpus of Learner English***. In: J. MURKHERJEE; M. HUNDT (eds.) 2011. ***Exploring***

Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap. Amsterdam e Philadelphia: John Benjamins. pp. 55-78.

HYLAND, K. 2009. ***Academic Discourse: English. In A Global Context.*** London: Continuum.

_____. 2008a. ***As can be seen: Lexical bundles and disciplinary variation***. *English for Specific Purposes*, **27**.1: 4–21.

_____. 2008b. ***Academic clusters: text patterning in published and postgraduate writing***. *International Journal of Applied Linguistics*, **18**.1: 41-62

LAUFER, B; WALDMAN, T. 2011. ***Verb-Noun Collocations in Second Language Writing: A Corpus Analysis of Learners***. *Language Learning*, **61**. 2: 647–672. Michigan.

LEECH, G.1998. Preface. In: S. GRANGER (ed.) 1998. ***Learner English on Computer***. London: Longman. pp. xiv-xx

_____1997. ***Teaching and language corpora: A convergence***. In: A. WICHMANN;

S. FLIGELSTONE; T. McENERY; G. KNOWLES (eds.) 1997. ***Teaching and language corpora***. London: Longman. pp 1-23.

LISS, R.; DAVIS, J. 2012. ***Effective Academic Writing***. Oxford: Oxford University Press.

NESSELHAUF, N. 2005. ***Collocations in a Learner Corpus***. **Studies In *Corpus Linguistics***. 14. Amsterdam: John Benjamins.

PAQUOT, M.; GRANGER, S. 2012. ***Formulaic language in learner corpora***. *Annual Review of Applied Linguistics*, **32**: 130-149. Cambridge.

SALAZAR, D. 2014. ***Lexical Bundles in Native and Non-native Scientific Writing***. Amsterdam: John Benjamins.

SCOLLON, R.; SCOLLON, S. 1981. ***Narrative, Literacy and Face in Interethnic***

Communication. Norwood, NJ: Ablex.

SCOTT, M.; TRIBBLE, C. 2006. ***Textual Patterns keyword and corpus analysis in language education***. Amsterdam: John Benjamins.

SHEPHERD, T.M.G. 2009. ***Corpora de aprendiz de língua estrangeira: um estudo contrastivo de n-gramas***. *Veredas*, **13**.2: 100-116. Juiz de Fora.

SIMPSON-VLACH, R.; ELLIS, N. C. 2010. ***An academic formulas list: New methods in phraseology research***. *Applied Linguistics*, **31**.4: 487-512. Oxford.

STUBBS, M. 2007. ***Notes on the history of corpus linguistics and empirical semantics***. In: M. NENONEN; S. NIEMI. (eds.) 2007. ***Collocations and Idioms***. Joensuu: Joensuun Yliopisto. 317-29. Disponível em: < http://www.uni-trier.de/fileadmin/fb2/ANG/Linguistik/Stubbs/stubbs-2007-hist-corp-ling.pdf >. Acesso em 29 mai 2015.

SWALES, J.; FEAK, C. B. 2012. ***Academic Writing for Graduate Students: Essential skills and tasks***. Michigan, Michigan ELT Press. 3rd. ed.

TRIBBLE, C. 2001. ***Small corpora and teaching writing. Towards a corpus-informed pedagogy of writing***. In: M. GHADESSY, A. HENRY, R.L. ROSEBERRY (eds.) 2001. ***Small corpus studies and ELT***. Amsterdam: John Benjamins. pp. 381–408.

Apêndice

Lista dos quadrigramas-chave positivos no *corpus* EAAL em relação ao BAWE, com valores brutos.

	EAAL	BAWE
on the other hand	34	836
it is possible to	29	396
one of the most	23	358
the beginning of the	15	210
fault in our stars	13	0
learning a second language	13	0
the fault in our	13	0
tells the story of	10	0
the tell tale heart	10	0
a walk to remember	9	0
learning a foreign language	9	0
end of the book	9	**6**
a second language is	8	0
as a second language	7	0
in the end of	7	0
a critical review of	7	0
of english as a	7	0
of the book is	7	**16**
of the english language	7	**18**
as a world language	6	0
blue is the warmest	6	0
english as a second	6	0
english as a world	6	0
fifty shades of grey	6	0
gone with the wind	6	0
is the warmest color	6	0
politics as a spectacle	6	0
the art of letting	6	0
the english language is	6	0
this critical review has	6	0
we can see in	6	**9**
to take care of	6	**17**
a psalm of life	5	0
art of letting go	5	0
as a global language	5	0
beginning of the book	5	0
captains of the sands	5	0
conclusion this critical review	5	0
contains a summary of	5	0
critical review has evaluated	5	0
english as a global	5	0

falls in love with	5	0
in the book the	5	0
is not always may	5	0
is possible to notice	5	0
is the one who	5	0
lord of the flies	5	0
of learning a second	5	0
of the fault in	5	0
section contains a summary	5	0
sou surda e no	5	0
summary this section contains	5	0
surda e no sabia	5	0
the full length movie	5	0
the main character and	5	0
this section contains a	5	0
the story of a	5	**6**
all around the world	5	**8**

Fonte: Elaborado pelas autoras.

Dedicamos este artigo ao amigo Tony Berber Sardinha por ser nosso 'Norte' na pesquisa com *corpus*. Nossa dedicatória inclui também os membros do GELC (Grupo de Estudos de Linguística de Corpus), cujos esforços coletivos sempre desafiam as impossibilidades.

Patrícia Bertoli is *Adjunct Professor at the English Department of the University of the State of Rio de Janeiro; PhD in Applied Linguistics from the Catholic University of São Paulo; visiting Professor at Federal University of Minas Gerais (2012); researcher mainly in Corpus Linguistics, TEFL, academic language and discourse analysis. E-mail: patbertolid@gmail.com*

Tania Shepherd is Associate Professor of English and Linguistics at UERJ, where she holds a FAPERJ research grant. She did her PhD at Birmingham University and postdoctoral studies with Tony Berber Sardinha at PUC-SP. She has co-edited Caminhos da Linguística de Corpus, Corpus nas receitas and Linguística da Internet. E-mail: tania.shepherd@gmail.com