A ESCRITA EM LÍNGUA ESTRANGEIRA NA UNIVERSIDADE: UM ESTUDO DE CASO

ACADEMIC WRITING AT UNIVERSITY: A CASE STUDY

Patrícia P Bertoli[1]
Tania M. G. Shepherd[2]

RESUMO: O bacharelando brasileiro de língua inglesa tem entre as inúmeras disciplinas do seu currículo uma cujo objetivo é lhe ensinar a escrita acadêmica. Nela, além de mostrar evidências de aquisição do discurso da língua estrangeira em seu registro escrito, o aluno também tem de evidenciar o domínio do léxico acadêmico. O presente trabalho descreve o léxico na produção escrita de uma turma de bacharelandos em inglês e respectivas literaturas através dos princípios da Linguística de Corpus. Para tal, foi compilado e digitalizado um corpus de escrita acadêmica produzida por aprendizes como parte dos requisitos da referida disciplina de escrita. Como corpus de referência foi usado um dos componentes do BAWE (*British Academic Writing in English*), contendo ensaios em inglês escritos por universitários da área de Humanidades, cuja primeira língua é o inglês. Com o auxílio de meios digitais foram extraídas, contabilizadas e comparadas as expressões lexicais mais frequentes usadas por ambos os grupos de alunos. O estudo faz uso de metodologia consagrada no tratamento de n-gramas (ou *clusters* lexicais) para desvelar um aspecto do aprendizado da língua inglesa escrita por parte do grupo estudado: o uso de vocabulário frequente por escritores iniciantes ao lidarem com as demandas de atividades da escrita argumentativa dentro da universidade.

Palavras–chave: escrita acadêmica; n-gramas; universitário brasileiro; língua inglesa

ABSTRACT: English majors in Brazil must attend a number of subjects as part of their curriculum, including a topic designed to teach them academic writing. In this subject, students need to show evidence of the acquisition of academic discourse in the foreign language but also provide evidence of their mastery of academic lexis. The present work describes the lexis in the written production of a class of English majors in Brazil, using Corpus Linguistics as methodology. To this end, a corpus of freshman academic writing was compiled and digitalized. As a reference corpus one of the components of the BAWE (*British Academic Writing in English*) was used, which consists of essays written in English by Humanities undergraduates, whose mother tongue is English. Digital tools helped extract, count and compare the most frequent lexical phrases used by both Brazilian and British students. The research uses long-established

1 Professora adjunta na Universidade do Estado do Rio de Janeiro; doutora em Linguística Aplicada e Estudos da Linguagem. E-mail: pat@corpuslg.org.

2 Professora associada na Universidade do Estado do Rio de Janeiro; doutora em Letras. E-mail: tania.shepherd@gmail.com.

methodology for the treatment of n-grams to unveil a feature of English language acquisition, namely the frequent use of academic vocabulary on the part of apprentice writers when dealing with argumentative writing as part of their university studies.

Keywords: academic writing; n-grams; Brazilian undergraduates; English language

INTRODUÇÃO

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

A situação que se apresenta para o aprendiz de língua estrangeira em nível terciário, entretanto, é desafiadora. Aprender a escrever um ensaio acadêmico significa adquirir um letramento difícil, porque específico da prática acadêmica, o assim chamado letramento ensaístico (*essayist literacy*) (SCOLLON; SCOLLON, 1981). Tal letramento incorpora a integração de camadas que vão desde a aquisição dos padrões retóricos da língua estrangeira, a habilidade de citar autores sinalizando se o estamos corroborando ou não, até o emprego do mesmo registro acadêmico da comunidade de prática relevante. Como diz Biber (2006:1)3, os alunos que estão começando seus estudos na universidade enfrentam vários obstáculos e ajustes que causam perplexidade, e muitas dessas dificuldades envolvem aprender a usar a linguagem de novas maneiras.

O que pode fazer então o professor de escrita acadêmica? Há inúmeros livros didáticos que se propõem a ensinar 'como' escrever um ensaio acadêmico. Esse estilo de livro, que se assemelha a um 'manual', parece não ajudar muito o aluno brasileiro, visto que tem como alvo duas 'representações' de leitor-ideal: o estudante de graduação estrangeiro que frequenta os cursos de introdução (*pre-sessional courses*) dados por suas respectivas universidades de língua inglesa, em sua grande maioria, e um público-alvo de pós-graduados. Essas publicações descrevem o passo a passo, nem sempre claro, de como transformar orações em parágrafos e parágrafos em ensaios (cf. LISS; DAVIS, 2012). O outro tipo de publicação é geralmente direcionado ao pós-graduando que necessita escrever artigos científicos. Nessas publicações encontram-se expressões e frases úteis para a escrita das várias partes ou movimentos retóricos de produtos acadêmicos (cf. GLASMAN-DEAL, 2010; SWALES; FEAK, 2012) cujo objetivo final é a publicação – algo longo, com múltiplos passos, que incluem a apresentação de problemas de pesquisa e discussão de resultados.

O presente trabalho desenvolvido por duas professoras de escrita acadêmica em inglês em nível terciário originou-se da necessidade de conhecer os pontos fortes e

[3] Tradução nossa de: "*students who are beginning university studies face a bewildering range of obstacles and adjustments, and many of these difficulties involve learning to use language in new ways*" (BIBER, 2006: 1).

fracos da escrita de seu alunado para então poder informar sua prática a partir dos mesmos pontos. O passo inicial da pesquisa foi o mapeamento das 'frases lexicais' existentes na escrita acadêmica dos alunos. As frases lexicais (também chamadas de *clusters*, pacotes lexicais, itens polilexicais) são itens frequentes e funcionalmente importante em textos acadêmicos, que, entretanto, os aprendizes de língua inglesa parecem achar problemáticos. A dificuldade de lidar com esses blocos de informação fica evidenciada no fato de que os aprendizes tendem a usar um repertório limitado dessas frases lexicais de forma correta, e o fazem repetindo as mesmas frases com maior frequência do que seria desejável.

Este trabalho usa uma metodologia calcada na Linguística de Corpus. É um trabalho de natureza contrastiva sobre o léxico, porque estuda comparativamente dois corpora de ensaios acadêmicos em inglês escritos por nativos e por aprendizes dessa língua. A proposta aqui é investigar as frases polilexicais frequentes nos dois corpora, respondendo às seguintes perguntas de pesquisa:

1. Que frases lexicais são encontradas somente no corpus de estudo?
2. Que frases lexicais são encontradas somente no corpus de referência?
3. Como os itens comuns a ambos os corpora estudados se configuram em termos de frequência e de uso?
4. Quais os entornos dessas expressões?

Ao responder a essas perguntas focando em um corpus de estudo de escrita acadêmica de universitários brasileiros, compilado pelas próprias professoras da classe de escrita acadêmica, pretende-se aqui mostrar um modelo de prática pedagógica que se alia com a pesquisa.

ESCRITA ACADÊMICA DE APRENDIZ: SOB O FOCO DO COMPUTADOR

Os estudos auxiliados por computador sobre corpora de escrita acadêmica em língua inglesa, de forma geral, enfocam corpora de textos escritos por a) por sujeitos aprendizes cuja língua materna é o inglês; b) escritores proficientes já publicados; c) falantes de inglês como língua franca, cujos textos já foram publicados ou aceitos para publicação; c) aprendizes da língua inglesa e da variedade expositiva. Essa última pesquisa sobre corpora contendo textos de aprendiz ou 'corpora de aprendiz' foi iniciada por Granger e seus colaboradores nos anos de 1990 (HYLAND, 2002, p. 176). Esses corpora nos fornecem evidência dos recursos léxico-gramaticais e discursivos utilizados por grupos de falantes nativos de uma língua A, ao se expressarem numa determinada língua adicional. Dada a facilidade e rapidez de poder vasculhar de forma confiável enormes quantidades de dados linguísticos autênticos, pode-se examinar com segurança e confiabilidade o que é usado com frequência de forma coletiva por esses grupos de usuários, o que é usado em demasia e o que é pouco ou nada usado. Investigam-se os padrões recorrentes, as tendências de uso de itens lexicais específicos ou de agrupamentos lexicais, além de tendências fraseológicas. Em termos de agrupamentos lexicais as pesquisas podem se concentrar em grupos lexicais não contínuos (como em *in the case of = in the + substantivo + of*) ou com os assim chamados n-gramas (também chamados de *cluster*, feixe lexical ou pacote lexical).

A investigação assistida por computador da escrita acadêmica em inglês de universitários brasileiros já tem produção expressiva derivada do GELC (Grupo de estudos em Linguística de Corpus) sediado na Pontifícia Universidade Católica de São Paulo (ver http://corpuslg.org/gelc/gelc. php). O foco de pesquisa vem se aprofundando no estudo de grupo lexicais de três e quatro palavras (trigramas e quadrigramas) na escrita de brasileiros oriundos de várias universidades. Os corpora usados nesses estudos são o Br-ICLE (*Brazilian International Corpus of Learner English*) em contraste com vários corpora de referência, igualmente escritos por universitários falantes de inglês como língua de herança.

Um dos muitos achados do GELC é que a população de universitários brasileiros investigada parece ter pouca ou nenhuma alternativa para as palavras ditas vagas como '*people*' ou '*thing*' e não demonstra conhecer as possibilidades anafóricas do pronome '*this*' (SHEPHERD, 2009). Além disso, parece desconhecer como usar, anaforicamente, substantivos abstratos (*content nouns*), como 'problema', 'solução' e outros. Um outro aprofundamento tem a ver com a extração e classificação funcional dos mesmos blocos lexicais, a partir das categorias propostas por Simpson-Vlach e Ellis (2010). Dutra e Berber-Sardinha (2013) mostraram que no corpus de aprendizes brasileiros estudado há uma prevalência de blocos com a função referencial sobre a função atitudinal. Em 2014, Dutra, Orfano e Berber Sardinha caminharam com a pesquisa, focando no mapeamento de blocos com função de posicionamento do escritor/aprendiz. O objetivo dessa vez foi discutir o papel ocupado por essas expressões em escrita acadêmica em inglês dos universitários estudados. Os resultados apontaram para a ausência de 'atenuadores' na escrita acadêmica examinada.

O estudo ora descrito foi desenvolvido a partir de um corpus que pode ser considerado pequeno, porque ainda em processo de compilação. Ele é, entretanto, igualmente representativo do contexto e da população específicos estudados em corpora maiores pelo GELC, visto que os sujeitos de pesquisa são universitários estudando a língua inglesa no bacharelado de uma única universidade.

MATERIAIS E MÉTODOS

O corpus deste estudo recebeu o nome de EAAL ou corpus de Ensaios Acadêmicos de Alunos de Letras, sendo composto por textos produzidos por alunos do Bacharelado em Letras Inglês-Literaturas, de uma universidade pública na cidade do Rio de Janeiro. Com consentimento por escrito dos sujeitos de pesquisa, foram coletados 174 textos produzidos por noventa e dois estudantes de primeiro ano, a partir de tarefas propostas por uma única professora, no período entre setembro de 2013 e dezembro de 2014. Cada texto contém aproximadamente 500 palavras e os estudantes tiveram como norma seguir os parâmetros de ensaios acadêmicos argumentativos, contendo quatro ou cinco parágrafos, conforme ensinado nas aulas. Os textos, produzidos como tarefa de casa e entregues à professora via e-mail, incluíram ensaios avaliativos de obra literária de escolha livre, ensaios biográficos e abarcaram também temas como: 'cultura inglesa', 'inglês como língua franca', 'gramática da língua inglesa', 'aprendizagem de língua estrangeira' e o 'adeus' na Literatura. O corpus foi compilado preservando-se a estrutura original dos textos, da maneira como foram entregues, excluindo-se os nomes dos autores, e salvos em arquivos em formato de texto (.txt), a fim de permitir

sua análise lexical por meio do programa *Wordsmith Tools. 5.0*. O mesmo programa foi usado para a extração de dados do corpus de referência.

O EAAL soma 94.343 palavras, sendo 8.949 palavras diferentes. Embora possa ser classificado como um corpus pequeno (cf. BERBER SARDINHA, 2005), é composto por toda a produção escrita de um mesmo grupo de alunos no período de coleta. Vale ressaltar que corpora pequenos também são considerados interessantes em estudos linguísticos (STUBBS, 2005), sobretudo para contextos e registros especializados e para o ensino da escrita, por oferecerem possibilidade de descobertas para contextos específicos, como afirmam Flowerdew (2001) e Tribble (2001), sendo, portanto, pertinentes para a investigação ora descrita.

Como corpus de referência foi utilizado um componente do corpus BAWE (*British Academic Written English*). O BAWE (disponível mediante licença de http://ota.ahds.ac.uk) tem a totalidade de 6.506.995 itens lexicais, advindos de 2859 trabalhos escritos por alunos de graduação e mestrado, por um período de três anos, derivados de trinta e cinco disciplinas diferentes. Os textos incluem estudos de caso, resenhas, ensaios, explicações, revisão da literatura, seções de metodologia, narrativas, resoluções de problemas, propostas e relatórios de pesquisa. Levando-se em consideração que o corpus de estudo da pesquisa ora descrita é de produção textual de alunos de Letras Inglês-Literaturas, optou-se por utilizar apenas a parte compatível do BAWE em relação aos sujeitos estudados. Assim, foram considerados somente os textos das áreas de Letras e Humanidades, das disciplinas específicas de Inglês e Linguística, cuja totalidade é de 362.378 palavras, sendo 19.474 palavras diferentes. Sendo assim um corpus de referência quatro vezes maior que o corpus de estudo, o que se coaduna com os preceitos de Berber Sardinha (comunicação pessoal) quando afirma que, caso o pesquisador queira ter informação suficiente para julgar cada palavra ou grupo de palavras do corpus de estudo deve usar um corpus de referência muitas vezes maior do que o que corpus de estudo (BERTOLI; SHEPHERD, 2015). Tal diferença exponencial na quantidade de palavras possibilita visualizar uma gama maior de ocorrências do léxico em foco.

Selecionados os corpora de estudo e de referência, passou-se para a fase de levantamento de quadrigramas. Um quadrigrama (ou um grupo lexical de quatro palavras) é qualquer combinação de quatro palavras que ocorre com uma frequência estabelecida pelo pesquisador. A razão para a escolha de sequências é que ao se estudar como as sequências são usadas em diferentes corpora, torna-se possível isolar as unidades (pré-fabricadas ou não) de que as diferentes categorias de escritores se utilizam para construir seus argumentos, exemplos e modos de referenciar entre outros (SCOTT; TRIBBLE, 2006, p.ix). Por outro lado, a razão para a escolha de sequências de quatro palavras segue os padrões de Cortes (2004, p. 401) que afirma serem unidades de trabalho produtivas, porque um quadrigrama pode conter o imbricamento de um trigrama.

Os quadrigramas dos dois corpora foram computados com o auxílio do software *WordSmith Tools 5.0*, o qual retornou 274 quadrigramas do corpus EAAL e 2373 quadrigramas do corpus de referência. Entre os quadrigramas mais frequentes no corpus de estudo foram encontradas sequências lexicais que representavam títulos de obras e expressões diretamente ligadas aos temas das tarefas, tais como: "*Fault in our stars*"; "*The fault in our*"; "*The tell tale heart*"; "*the beginning of the*"; "*learning a second language*"; "*end of the book*"; "*English as a second*". Essas expressões foram desconsideradas para a análise por não representarem verdadeiramente destaque de uso, vez que

seria impossível falar de obras ou livros sem mencionar seus títulos, ou discutir seus temas sem os citar.

A partir desses resultados, decidiu-se estudar as sequências lexicais-chave. Palavras-chave (ou sequências de palavras-chave) advêm 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). Dessa forma, indicam uma característica intrínseca a um conjunto de textos, também chamada de chavicidade (ou *aboutness*) A pesquisa com itens lexicais-chave como elementos de busca colabora tanto para a identificação de características de textos e registros, como também para análise do discurso e dos mesmos textos (BONDI, 2010, p. 1) e, portanto, são um elemento essencial na busca de fraseologias típicas.

A extração dos quadrigramas-chave do corpus de estudo, em relação à porção de Letras e Humanidades do corpus BAWE, retornou 162 itens. Em outras palavras, após serem desconsideradas as expressões que designavam temas e tarefas, restaram três grupos de expressões usadas no corpus de estudo: aquelas usadas em excesso, aquelas usadas com escassez e ainda aquelas usadas erroneamente. A tabela a seguir mostra os 30 primeiros quadrigramas-chave extraídos do EAAL em relação ao BAWE-AH acompanhados do número de suas ocorrências.

Tabela 1: Quadrigramas–chave positivos no *corpus* EAAL

N	Quadrigrama	EAAL	BAWE-AH	N	Quadrigrama	EAAL	BAWE-AH
1	ON THE OTHER HAND	35	66	16	IN THE SAME WAY	5	15
2	IT IS POSSIBLE TO	29	41	17	IT IS IMPORTANT TO	5	26
3	ONE OF THE MOST	23	14	18	THE FACT THAT THE	5	37
4	THE END OF THE	18	53	19	THE USE OF THE	5	54
5	THE BEGINNING OF THE	15	30	20	IT IS NECESSARY TO	4	14
6	TELLS THE STORY OF	10	0	21	TO BE ABLE TO	4	26
7	AT THE END OF	10	58	22	BY THE USE OF	3	16
8	END OF THE BOOK	9	0	23	IN THE CASE OF	3	21
9	AT THE SAME TIME	7	26	24	IT IS CLEAR THAT	3	22
10	TO THE FACT THAT	7	27	25	THE MAJORITY OF THE	2	13
11	CAN BE SEEN AS	6	15	26	ALLOWS THE READER TO	2	14
12	THE REST OF THE	6	21	27	THE IMPORTANCE OF THE	2	16
13	AS WELL AS THE	6	23	28	DUE TO THE FACT	2	19
14	AS A RESULT OF	6	28	29	IS AN EXAMPLE OF	2	21
15	AT THE BEGINNING OF	6	34	30	THAT THERE IS NO	2	22

A fim de se buscar ainda mais luz sobre a natureza das expressões extraídas para a análise, esta pesquisa também lançou mão de dois corpora: o COCA (*Corpus of Contemporary American English*), e o BNC (*British National Corpus*). O primeiro com

cerca de 450 milhões de palavras e cuja plataforma se encontra on-line, permite pesquisar volume de frequência e linhas de concordâncias de palavras em expressões de busca (disponível em: <http://corpus.byu.edu/coca/>). O segundo contem 100 milhões de palavras e está disponível em: http://www.natcorp.ox.ac.uk/.

Em resumo, os procedimentos metodológicos para o desenvolvimento desta pesquisa seguiram os seguintes passos. Foram executados a coleta do corpus do estudo (EAAL); a seleção do corpus de referência (BAWE-AH); a extração dos quadrigramas dos dois corpora com a ferramenta *Wordsmith Tools 5.0*; a extração de quadrigramas comparados ou quadrigramas-chave; o levantamento de quadrigramas relevantes (ou eliminação dos quadrigramas com saliência temática) e por fim a análise dos resultados. Esse último procedimento está descrito na seção seguinte.

ANÁLISE DOS DADOS

Esta seção apresenta a análise dos resultados a partir dos dados coletados, seguindo os passos descritos na seção anterior. Num primeiro momento, os quadrigramas de cada um dos corpora de pesquisa foram observados quantitativamente. Dessa forma, foram identificadas as seguintes características:

1. Há quadrigramas-chave que ocorrem apenas no EAAL;
2. Há quadrigramas-chave que ocorrem apenas no BAWE-AH;
3. Os corpora contêm diversas 'molduras' lexicais (expressões que mantém um formato conformação constante com exceção de um item lexical) e
4. O EAAL apresenta sequências semânticas (sequências que se somam a outras sequências) que não se combinam da mesma forma que no BAWE-AH.

No que diz respeito à existência de quadrigramas-chave presentes apenas no corpus de estudo, foram encontradas as expressões "*tells the story of*" e "*end of the book*" (vide figura 1). O não aparecimento de tais sequências lexicais no corpus de falantes nativos BAWE-AH chama a atenção especialmente porque as expressões são gramaticalmente corretas e parecem ser de uso comum na língua inglesa em geral. Comparar listas de frequência de apenas duas fontes demonstrou ser insuficiente. Foi neste ponto que se buscou evidência de uso em outros corpora: o COCA e o BNC.

A comparação da frequência de "*tells the story of*" em diversos corpora pode ser melhor visualizada na tabela a seguir.

Tabela 2: Frequências de "*tells the story of*"

corpus	*frequência*
EAAL	10
BAWE	0
COCA	735
COCA ACADEMIC	137
BNC	100
BNC ACADEMIC	7

Como era antevisto, o quadrigrama "*tells the story of*" ocorre em outras instâncias da língua inglesa. No que diz respeito ao inglês geral, a expressão ocorre 735 vezes no corpus de inglês americano geral (COCA) e 100 vezes no corpus de inglês britânico geral. Quando se examina seu uso em textos exclusivamente acadêmicos, a sequência semântica ocorre 137 vezes no inglês americano e 7 no inglês britânico, todas as ocorrências na área específica de Humanidades. Observa-se, portanto, uma tendência dos aprendizes em adotar formas da variedade americana da língua inglesa. Isto poderia justificar a ausência da expressão "*tells the story of*" no BAWE, que apresenta apenas textos produzidos por estudantes em universidades britânicas. De modo geral, pode-se concluir que, embora ausente no BAWE trata-se de um quadrigrama de uso permissível em inglês de maneira geral e em textos acadêmicos produzidos em inglês americano.

O mesmo procedimento comparativo foi adotado para investigar a ocorrência do quadrigrama "*end of the book*", que aparece nove vezes no EAAL e zero vezes no BAWE-AH, como pode ser visto na figura 3, a seguir:

Tabela 3: Frequências da expressão "*end of the book*"

corpus	*frequência*
EAAL	9
BAWE	0
COCA	179
COCA ACADEMIC	51
BNC	70
BNC ACADEMIC	16

Entretanto, "*End of the book*" demonstrou ser usado na escrita em língua inglesa, tanto de maneira geral quanto especificamente acadêmica. A tabela anterior mostra que a expressão ocorre 179 vezes no COCA e 70 vezes no BNC, assim como ocorre 51 vezes no corpus americano acadêmico e 16 vezes no corpus britânico acadêmico.

Pode-se concluir, portanto que os quadrigramas-chave que ocorrem exclusivamente no corpus EAAL não representam problemas linguísticos para os sujeitos desta pesquisa. As duas expressões não apareceram no BAWE-AH, talvez por pertencerem mais comumente à variedade americana da língua inglesa.

O segundo item identificado pela extração quadrigramas-chave entre o EAAL e o BAWE é a ocorrência de algumas expressões apenas no corpus de referência. Conforme dito anteriormente, a extração de quadrigramas retornou 162 tipos diferentes, dos quais 2 ocorrem apenas no EAAL, 57 (35.2%) estão presentes tanto no corpus de estudo quanto no BAWE-AH, enquanto 103 (ou 63.6 %) ocorrem somente no corpus de falantes nativos. Isso significa que os estudantes de letras Inglês-Literaturas, sujeitos desta pesquisa, compartilham 57 dos 162 quadrigramas utilizados também por graduandos e mestrandos da área de Letras e Humanidades do BAWE. Dos 103 quadrigramas presentes exclusivamente no BAWE-AH, alguns parecem também indicar tópicos, tais como: "*definition of standard English*"; "*modernism in poetry motivations*"; "*In The Waste Land*"; "*male and female speech*"; "*Of the past tense*"; "*The first world war*"; "*in Fletcher and Garman*"; "*in nineteen eighty four*"; entre outros. Vale lembrar que as expressões

pertencentes aos temas das tarefas do corpus EAAL foram desconsideradas para a análise. Todavia, esse procedimento não foi possível para o BAWE-AH em virtude do desconhecimento por parte das pesquisadoras dos temas dos ensaios acadêmicos. Por conseguinte, optou-se por não descartar tais expressões. Contudo, se essas expressões pudessem também ter sido eliminadas na extração de quadrigramas-chave, a diferença de uso de quadrigramas entre os dois corpora poderia ser consideravelmente menor.

Os vinte e nove quadrigramas-chave com maior índice de recorrência no BAWE-AH que não aparecem no EAAL estão dispostos na tabela a seguir:

Tabela 4: Quadrigramas não compartilhados pelos corpora

N	Quadrigrama	EAAL	BAWE-AH
1	CAN BE SEEN IN	0	26
2	THE WAYS IN WHICH	0	24
3	IT COULD BE ARGUED	0	23
4	COULD BE ARGUED THAT	0	21
5	THE CONTEXT OF THE	0	21
6	CAN BE FOUND IN	0	20
7	THE FORM OF A	0	19
8	THE EXTENT TO WHICH	0	19
9	THE NATURE OF THE	0	18
10	TO LOOK AT THE	0	18
11	THE STRUCTURE OF THE	0	17
12	AS CAN BE SEEN	0	16
13	THE ROLE OF THE	0	15
14	THE MEANING OF THE	0	15
15	FOR THE PURPOSES OF	0	15
16	THE WAY WE SPEAK	0	15
17	EXAMPLE OF THIS IS	0	15
18	IN THE CONTEXT OF	0	14
19	AS A MEANS OF	0	14
20	OF THE TONE UNIT	0	13
21	THE USE OF INDEFINITE	0	12
22	REST OF THE POEM	0	12
23	IN THE LIGHT OF	0	12
24	ARE MORE LIKELY TO	0	12
25	OF THE USE OF	0	11
26	COULD BE READ AS		11
27	THIS CAN BE SEEN	0	11
28	CAN BE SAID TO	0	10
29	CAN BE DEFINED AS	0	9

A observação dos quadrigramas não compartilhados pelos dois corpora chama a atenção para duas ocorrências. A primeira tem a ver com o padrão formado com o auxiliar *can/could* + be + particípio de verbos como *see/read* e *argue/say/define*. Apesar de os quadrigramas apresentarem variação no verbo modal (*can/could*), há uma coincidência em termos de campo semântico dos dois grupos de verbos da passiva. No primeiro grupo temos *read* (ler,interpretar) e *see* (ver, entender) que apontam para os processos de entendimento do leitor. Por outro lado temos *argue* (argumentar), *say* (dizer) e *define* (definir) que apontam para processos pertinentes ao escritor. Estes padrões estão ausentes na escrita do aprendiz brasileiro investigado.

A segunda ocorrência é a existência de padrões lexicais descontinuados conhecidos como molduras lexicais (*lexical frames*) como em *the* + substantivo + *of +the*. De acordo com Gray e Biber (2013, p.109), as molduras lexicais são sequências descontínuas de palavras contendo algumas palavras que formam uma constante, as quais envolvem um item lexical variável, ou seja, sequências constantes cuja palavra central é variável como por exemplo em "*the * of the*".

Uma análise simples de valores puros das frequências dos quadrigramas pesquisados indica que 103 dos 162 quadrigramas-chave não ocorrem no EAAL. Entretanto, uma análise mais detalhada de sequências mostra que os mesmos estudantes usam essas sequências descontínuas, utilizando-se de itens lexicais variáveis menos frequentes no corpus de inglês nativo. Essa constatação gerou o terceiro item investigado na pesquisa descrita aqui: a ocorrência, formação e significado de molduras lexicais.

A análise dos quadrigramas-chave com possibilidade de serem molduras lexicais destacou a moldura "*the * of the*" por sua variação e frequência. No corpus BAWE-AH esta moldura aparece com dezesseis variações, ou seja, com dezesseis itens lexicais distintos preenchendo o espaço central variável. As variações foram interpretadas lexicalmente como pertencentes a três campos semânticos distintos: denotação de função ou papel, localização no espaço e quantificação. Essa tentativa de classificação pode ser observada na tabela a seguir:

Tabela 5: Moldura lexical "*the * of the*"

Campo semântico	quadrigrama	EAAL	BAWE(AH)
função, papel	THE USE OF THE	5	54
	THE CONTEXT OF THE	0	21
	THE NATURE OF THE	0	18
	THE STRUCTURE OF THE	0	17
	THE IMPORTANCE OF THE	2	16
	THE MEANING OF THE	0	15
	THE ROLE OF THE	0	15
	THE IMAGE OF THE	1	14
	THE IDEA OF THE	1	10
	THE POWER OF THE	0	9
localização no espaço	THE END OF THE	18	53
	THE BEGINNING OF THE	15	30
quantificação	THE MAJORITY OF THE	2	13
	THE REPETITION OF THE	1	15
	THE REST OF THE	6	21

Embora os textos dos alunos brasileiros demonstrem que eles sabem usar a moldura *"the * of the"* em questão, fica evidente que o fazem com mais frequência quando falam sobre localização espacial e quantificação. No que diz respeito ao uso de "*the * of the*" para falar sobre função ou papel, os dados demonstram frequências baixas, além de um repertório de substantivos centrais restrito. Mais especificamente, os sujeitos de pesquisa parecem saber falar sobre "*use*", "*importance*", "*image*" e "*idea*" em detrimento de (ou por não conhecerem) "*context*", "*nature*", "*structure*", "*meaning*", "*role*" e "*power*". O uso excessivo de um repertório restrito de palavras pode indicar apego por parte dos alunos a expressões que são reconhecidamente corretas para evitar o erro, ao mesmo tempo que pode sugerir que falta aos alunos uma metalinguagem para falar sobre a essência, significado e natureza dos eventos.

Finalmente, o quarto item investigado é a presença de sequências semânticas 'bem formadas'. Sequências semânticas, ou unidades de significados compostas de múltiplos itens lexicais, nada mais do que expressões que demonstram a tendência de certas palavras ocorrerem juntas (HUNSTON, 2002; 2008; 2012) e desempenharem funções discursivas em um texto.

Em um estudo anterior, exclusivamente sobre o uso de quadrigramas na escrita acadêmica, Bertoli e Shepherd (2015) contrastaram as frequências de quadrigramas do corpus EAAL com a totalidade do BAWE – não apenas das disciplinas de Inglês e Linguística da porção de Artes e Humanidades do mesmo corpus, como é o caso da pesquisa aqui descrita – e salientaram que o uso excessivo de certos quadrigramas indica que os aprendizes "usam um número limitado de possibilidades combinatórias" (BERTOLI; SHEPHERD, 2015, p. 257) o que, mais uma vez, reflete "escolhas seguras de colocações frequentes" (BERTOLI; SHEPHERD, 2015, p.258). Naquele estudo, foram investigadas especificamente as expressões "*on the other hand*", "*it is possible to*", "*in the end of*" e "*one of the most*", vez que ocuparam frequência de destaque. As autoras sugeriram que uma investigação dos entornos de uma sequência lexical frequente pode revelar o uso de outras sequências lexicais também frequentes, aumentando assim a dimensão (span) das unidades investigadas.

Um problema que se pode encontrar é que ao utilizarem "quadrigramas corretos seguidos de outros quadrigramas corretos" por se sentirem mais seguros, os aprendizes podem criar combinações impossíveis ou improváveis na língua estrangeira (BERTOLI; SHEPHERD, 2015, p. 258). O presente estudo procurou verificar tal possibilidade por meio da análise dos entornos da sequência "*it is necessary to*", que ocorre três vezes mais no EAAL que no BAWE-AH. No corpus de falantes nativos "it is necessary to" aparece comumente antecedida por "*before **" ; "*to do **" e "*if doing **". Tal disposição de itens lexicais forma uma sequência semântica usada para falar sobre condições, algo que deve ser necessário e suficiente para algo mais acontecer, como pode ser observado nos seguintes exemplos retirados do BAWE-AH:

Before doing X, it is necessary to do Y (Antes de fazer X, é necessário fazer Y)

1. *Before examining those features it is necessary to consider Standard English.*

2. *Before proceeding to the analysis of L1 in relation to my learning experience it is necessary to clarify the context of this learning sit.*

To do X, it is necessary to do Y (Para fazer X, é necessário fazer Y).

3. *To calculate the frequencies of the different pronouns in the political speeches it is necessary to download a concordancer program.*

4. *To study how language is used it is necessary to observe language actually used.*

If doing X, it is necessary to do Y. (Se fizer X, é necessário fazer Y).

5. *If adopting this definition and relating it to the statement above it is necessary to examine exactly who, during the Renaissance.*

Em contraste, no corpus EAAL o quadrigrama "*it is necessary to*" só compartilha sequências semânticas com o corpus BAWE-AH na medida de 50% dos casos, conforme exemplificado pelas seguintes sentenças:

6. *Is damaging and must not be disregarded. For such, it is necessary to motivate people to overcome the genre.*

7. *To understand Norman, it is necessary to comprehend schizophrenia, which is a com.*

8. *And to enter it, or even keep at it, it is necessary to go beyond the native language.*

9. *Of course it is necessary to communicate the understanding of all sides.*

Os exemplos 6 e 7 enquadram-se no frame "*to do **"*, considerando-se "*do*" como qualquer verbo, ou seja, "*para entender Norman*" "é necessário **" e "*para entrar nele ou até mantê-lo*" "é necessário **". Já os exemplos 1 e 4 aparecem precedidos pelas expressões "*For such*" e "*Of course*" que não aparecem no BAWE-AH.

A fim de se verificar se não se tratava de uma característica exclusiva do corpus de estudo, buscou-se pelos mesmos exemplos no COCA e no BNC. Não foi encontrada nenhuma ocorrência de "*for such*" e apenas uma ocorrência de "*of course*", precedida pela palavra "*but*" no BNC, a saber: "*Canon at the octave or double octave is very similar to the above, but of course it is necessary to avoid crossing one voice below the bass.*" O uso de "*but*" na frase anterior indica uma pré-condição para 'aquilo que é necessário'.

Dessa forma, pode-se concluir que os sujeitos desta pesquisa usam o quadrigrama "*it is necessary to*" em seus textos de forma correta no que diz respeito à combinação dessas quatro palavras. Todavia, na construção de sequências semânticas, os mesmos aprendizes desconhecem como implementá-las metade das vezes.

CONSIDERAÇÕES FINAIS

Este trabalho apresentou os resultados de um estudo sobre o uso de quadrigramas, em um corpus de produções acadêmicas de bacharelandos em inglês em comparação com um corpus de produção acadêmica de nativos. A pesquisa enquadra-se no âmbito de investigações e compilação de corpora de linguagem de aprendiz que advogam melhor conhecimento das particularidades da escrita acadêmica do aprendiz (BERBER SARDINHA, 2004; HYLAND, 2009; DUTRA; BERBER SARDINHA, 2013). Esse braço da Linguística de Corpus explora empiricamente as produções de aprendizes para informar as práticas pedagógicas de professores.

Este estudo visou não simplesmente analisar o ensaio acadêmico do aprendiz de uma língua estrangeira de forma coletiva, mas, sobretudo, ilustrar para as professoras das disciplinas de escrita acadêmica questões lexicais conflitantes na escrita de seus alunos. E que questões conflitantes são essas?

A análise feita a partir do levantamento de quadrigramas-chave (ou sequências de quatro palavras características do corpus de estudo) revelou um uso excessivo de alguns quadrigramas. Este achado já foi constatado em outros trabalhos sobre linguagem

de aprendiz. Foi detectado, por exemplo, por Granger e Gilquin (2011) na produção de aprendizes cujas línguas maternas eram outras línguas que não o português. Pode-se somente especular sobre o uso em excesso de determinados quadrigramas que são corretos. Usar algo porque está corretamente estruturado parece oferecer uma maior segurança ao aprendiz no que tange à sua performance linguística. Entretanto fazê-lo em excesso confere ao discurso do aprendiz um aspecto repetitivo, como se lhe faltassem sinônimos ou formas alternativas para construir os mesmos sentidos.

Por último, o estudo mostrou um problema de desconhecimento sobre o uso das chamadas sequências semânticas – ou seja, grupos de itens lexicais que juntados a outros itens lexicais podem servir de âncoras discursivas para o desenvolvimento textual. Ficou evidenciado que nossos alunos empregam seus quadrigramas de forma que combinem entre si, para alavancar ideias dentro do texto. O presente trabalho ratificou a importância da coleta de dados empíricos a partir da produção textual de aprendizes, mas constatamos que para elucidar como o aprendiz realmente molda seu discurso, há que se investigar unidades lexicais cada mais longas e complexas.

REFERÊNCIAS

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

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

BERTOLI, P. P.; SHEPHERD, T. M. G.Escrita acadêmica: um estudo exploratório de quadrigramas. **The ESPecialist**, v. 2, n. 36, pp. 241-262, 2015.

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

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

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

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

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

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

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

GRANGER, S; GILQUIN, G. From EFL to ESL: Evidence from the International Corpus of Learner English. In: MURKHERJEE, J.; HUNDT, M. (eds.). **Exploring Second-Language Varieties of English and Learner Englishes:** Bridging a Paradigm Gap. Amsterdam e Philadelphia: John Benjamins, 2011, pp. 55-78.

GRAY, B.; D. BIBER. Lexical Frames in Academic Prose and Conversation. **International Journal of Corpus Linguistics,** n. 18, pp.109-135, 2013.

HUNSTON , S. Afterword: The problems of applied linguistics. In: HYLAND, K. et al (eds.) **Corpus Applications in Applied Linguistics**. London: Continuum, 2012, pp. 242-248.

____. Starting with the small words: Patterns, lexis and semantic sequences. In **Patterns, meaningful units and specialized discourses**. Special Issue of International Journal of Corpus Linguistics, v. 3, n. 13, pp. 271-295, 2008.

____. **Corpora in Applied Linguistics**. Cambridge University Press, 2002.

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

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

SCOLLON, R.; SCOLLON, S. **Narrative, Literacy and Face in Interethnic Communication.** Norwood, NJ: Ablex, 1981.

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

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

STUBBS, M. Notes on the history of corpus linguistics and empirical semantics. In: NENONEN, M.; NIEMI, S. (eds.) **Collocations and Idioms**. Joensuu: Joensuun Yliopisto, 2007, pp. 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.

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

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

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