O Léxico do Inglês Escrito: Um Estudo Contrastivo de Inglês Como Língua Materna E Segunda Língua

Phaedra de Athayde (PIBIC/UERJ)[1]

Resumo: O presente trabalho analisa escolhas lexicais feitas por um grupo de aprendizes brasileiros de língua inglesa. Para tal compara um corpus digitalizado formado de 105 redações em inglês escritas por esses alunos, com outro corpus (LOCNESS), composto por redações escritas por alunos cuja língua materna é o inglês americano e britânico. O estudo tem como objetivo obter uma 'radiografia' da natureza lexical desses conjuntos de textos, que aqui são analisados comparativamente em termos de suas freqüências lexicais. A proposta se baseia nos princípios da Lingüística de Corpus e principalmente na metodologia proposta por Ringbom (1998).

1) Introdução

A produção escrita de alunos de inglês como língua estrangeira é um dos desafios que mais freqüentemente se coloca para o pesquisador da área de Lingüística Aplicada. Ainda há muito a ser estudado sobre a escrita desses aprendizes em termos de natureza retórica, em nível macro, e de escolhas lexicais, em nível micro.

O presente trabalho pretende preencher esta lacuna, ainda que de forma singela. A pesquisa enfoca 105 redações em inglês escritas por alunos brasileiros e tem como objetivo obter uma 'radiografia' da natureza lexical desses textos, que aqui são analisados em termos de suas freqüências lexicais. Esta proposta é completada por uma análise comparativa, baseada na metodologia proposta por Ringbom (1998), em seu trabalho **"**Vocabulary frequencies in advanced learner English: a cross-linguistic approach**"**, no qual foram comparadas e contrastadas redações em língua inglesa como língua materna e redações em língua inglesa escritas por aprendizes de inglês de sete países da União Européia. A motivação para a presente pesquisa partiu das propostas encaminhadas por Ringbom na conclusão do artigo acima citado, de que o inglês escrito por outras nacionalidades deveria ser objeto de investigações semelhantes (Ringbom, 1998: 51). Como aprendizes de inglês, falantes de português como língua materna não haviam feito parte de seu estudo, decidimos proceder a uma análise semelhante, tendo como corpus de estudo as redações em inglês de aprendizes brasileiros. Assim, a partir da comparação entre as freqüências lexicais obtidas em dois corpora – de alunos que têm o inglês e o português como língua nativa, respectivamente – procuramos responder a algumas das questões relativas à natureza das redações desses alunos.

[1] Orientadora: Profa. Tania Shepherd

Por conta do tipo de abordagem empregada, pode-se dizer que este trabalho se insere em três áreas, a saber: a Análise do Discurso (doravante AD), o ensino de língua estrangeira (doravante LE) e a Lingüística de Corpus.

1.1) Análise do Discurso

Este estudo faz parte da área denominada Análise do Discurso (AD), já que uma das atribuições desta é ajudar a entender as escolhas lingüísticas que fazemos em termos de uso. É ainda através da AD que obtemos ferramentas para analisarmos padrões de vocabulário e organização textual típicos de cada gênero textual (Paltridge, 2000). Por conta disso, quando fazemos uma análise comparativa entre as freqüências lexicais nas redações de usuários de uma determinada língua como L1 ou L2[2] , na verdade, estamos nos apoiando em algumas orientações teóricas propostas pela AD. Além disso, através da AD é possível obtermos informações sobre quem produz o discurso e quando – dados que definem o enunciador – e sobre as características próprias a este discurso – dados que definem o enunciado. Neste estudo, o enunciador é formado pelo conjunto de 105 alunos brasileiros estudantes de inglês como língua estrangeira. O enunciado, por sua vez, é o conjunto das redações produzidas por estes alunos para avaliação final de seus cursos, aqui analisadas em termos de suas escolhas lexicais. Nossa abordagem nos faz chegar ao ´coletivo´ , ou seja, quem é esse enunciador, o aprendiz de língua estrangeira que estamos investigando, e como ele escreve.

1.2) Ensino e aprendizagem de L2

A segunda inserção deste estudo se acha na área a que chamamos ensino e aprendizagem de L2. Quando compilamos, para estudo, um ‘corpus de aprendiz’ (denominação adotada por Granger, 1998), habilitamos o pesquisador (e indiretamente o professor) além do aprendiz a desenvolverem várias tarefas, já que há inúmeras vantagens da utilização de um corpus de aprendiz em aplicações pedagógicas. Segundo Tognini Bonelli (1996: 9), a utilização desse tipo de corpus é possível tanto por parte de professores, como por parte de alunos de todos os níveis, pois não requer prática ou embasamento teórico prévios. O corpus de aprendiz pode ser uma ferramenta diagnóstica útil tanto para alunos quanto para professores na medida em que identifica

[2] Neste trabalho usaremos L1 para designar o inglês enquanto língua materna e LE para o inglês como língua estrangeira.

os padrões característicos da produção dos alunos em termos de seus erros e acertos na produção de determinados textos, por exemplo.

Para o professor de LE, as aplicações pedagógicas são possíveis em termos da facilidade de manuseio do corpus. Essa utilização caracterizaria um aprendizado **"bottom-up"** (Tognini Bonelli, 1996: 14), visto que, para o aluno, a utilização de um corpus de aprendiz passa pela conscientização e construção do aprendizado. Mais especificamente, é através da comparação de próprio texto a textos autênticos em LE (jornais, revistas etc.) e do desenvolvimento da capacidade de descobrir por si mesmo as similaridades e as diferenças entre sua escrita e a escrita dos nativos da língua estrangeira, que o aprendiz desenvolve uma conscientização sobre o que escreve. Além disso, o corpus de aprendiz também pode ser usado para detectar os erros e acertos individuais de cada aluno (Tognini Bonelli, 1996: 9).

Em termos da construção do aprendizado, o trabalho desenvolvido a partir de um corpus de aprendiz pode conferir ao aluno autonomia para efetivamente construir o conhecimento e, de forma bem orientada, é capaz de fazer do aluno um verdadeiro pesquisador, capaz de identificar, compreender e corrigir seus erros mais comuns (Tognini Bonelli, 1996: 9).

1.3) Lingüística de corpus

A ultima área de inserção do presente estudo é a da Lingüística de Corpus, por três razões principais:

a. trata-se de um estudo empírico, cujo ponto de partida é a coleta de dados autênticos para posterior análise;

b. este estudo mantém o foco na linguagem em uso; e

c. este estudo utiliza-se de novas tecnologias para análise lingüística, mais especificamente de um programa de computador. (Tognini Bonelli, 1996: 2)

A respeito das vantagens da utilização do computador, Berber Sardinha (2004: 85) nos apresenta algumas delas:

> "Maior emprego de computadores na investigação da linguagem seria benéfico. Em primeiro lugar, são consistentes. Os computadores não se cansam e podem fazer tarefas tediosas (...) de modo eficiente e confiável. Em segundo lugar, permitem maior abrangência na quantidade de dados que se pode lidar. (...) Uma outra vantagem diz respeito à possibilidade da descoberta de fatos novos, ou mesmo da contestação de opiniões e crenças estabelecidas. (...) Além de permitir enxergar fenômenos novos, o computador pode também modificar o modo de se enxergar a linguagem."

Entretanto, as vantagens da utilização do computador na análise lexical, vai mais além do que a identificação de fenômenos novos. Segundo Sinclair ( ) o olho humano é capaz de perceber diferenças sem maiores dificuldades, mas para o olho humano é difícil apreender no meio de muitos dados, aquilo que é repetido, padronizado. Esta tarefa é desempenhada pelo computador sem maiores dificuldades. A Lingüística de Corpus e seus conceitos são, portanto, o instrumento ideal para olhar a padronização de escolhas lexicais em corpus alentado, como o que acontece com o presente estudo.

1.4) Perguntas de pesquisa

Tendo como ponto de partida a pergunta mais abrangente: "Como se compara a prosa argumentativa em inglês como LE e como L1 em termos de léxico?", este estudo se propôs a pesquisar e comparar mais detidamente três aspectos específicos dos corpora de aprendizes compilados. São eles:

a. os 10 itens mais freqüentes em ambos os corpora, ou seja, as 10 palavras mais usadas por ambos os conjuntos de alunos investigados;

b. os 100 itens mais freqüentes em ambos os corpora em termos da sua distribuição percentual em relação ao vocabulário total; e, finalmente,

c. a densidade lexical dos dois corpora em questão.

2) Aporte teórico

Nossa pesquisa é calcada em Ringbom (1998), em cujo trabalho "Vocabulary frequencies in advanced learner English: a cross-linguistic approach**"**, a autora compara as freqüências lexicais em textos escritos por alunos de L1 e de aprendizes de inglês como LE. Para tanto, ela faz uso de dois corpora digitais. Um deles, denominado **LOCNESS** (Louvain Corpus of Native Speaker Essays) consiste de redações escritas em L1 por alunos norte-americanos e britânicos. O outro, denominado **ICLE** (International Corpus of Learner English) é composto por um conjunto de redações escritas em inglês por alunos da União Européia, cujas línguas maternas são o francês, o espanhol, o finlandês, o finlandês-sueco, o sueco, o holandês e o alemão[3].

Determinados assim os corpora, Ringbom faz um levantamento minucioso das freqüências lexicais em cada um deles. Em seguida a autora apresenta uma seqüência de

[3] Não há um corpus de inglês escritos por aprendizes cuja L1 é o português europeu. A coleta de dados de aprendizes de inglês que falam português do Brasil como L1 ainda está em progresso.

tabelas que efetuam comparações entre os itens mais freqüentes em cada um dos corpora, desde, por exemplo, os itens de classes fechadas, como artigos e pronomes, até os verbos lexicais mais constantes, estabelecendo uma radiografia de cada um dos grupos estudados. em contraste com o corpus LOCNESS.

Para tal, Ringbom lança mão de um conceito da Lingüística de Corpus denominado *type/token ratio simples*, para calcular a densidade lexical de cada um dos seus corpora. Como no presente trabalho este conceito também é empregado quando da determinação das densidades lexicais, faz-se necessário esclarecer o que significa cada um dos componentes desta razão.

A definição de *type* compreende o vocábulo em si, o tipo de item, em suma, cada um item lexical diferente que compõe um texto. A definição de *token* corresponde ao número total de palavras de um texto, consideradas todas as repetições destas palavras (Scott, 1999). De acordo com Berber Sardinha (2004: 94):

> Na prática, a razão vocábulo / ocorrência indica a riqueza lexical do texto. Quanto maior o seu valor, mais palavras diferentes o texto conterá. Em contraposição, um valor baixo indicará um número alto de repetições, o que pode indicar um texto menos rico, ou variado, do ponto de vista de seu vocabulário.

Outro conceito empregado por Ringbom e incorporado ao presente trabalho é o conceito de *type/token ratio padronizada*. A razão *type/token* (vocábulo/ocorrência) na forma padronizada é calculada em intervalos regulares e permite a comparação de textos de tamanhos diferentes, a partir da determinação de um intervalo comum para o cálculo (Scott, 1999). O resultado obtido é a média de todas as densidades parciais, calculadas tantas quantas forem as vezes em que o intervalo esteja presente ao longo do conjunto de textos. Ainda segundo Berber Sardinha (2004: 95), a forma padronizada seria utilizada sempre que fosse necessário neutralizar a influência de corpora de tamanhos distintos. Isto ocorre porque textos maiores, por sua natureza, apresentam mais repetições de palavras (ou um número maior de tokens) e, por isso, tendem a possuir valores para esta razão mais baixos do que textos curtos.

Na seção abaixo explicamos os materiais usados na presente pesquisa passo a passo da investigação.

3) Materiais e métodos

Como Ringbom (1998), utilizamos o LOCNESS como corpus de referência. LOCNESS é um corpus comercial, que consiste de ´essays´ argumentativos, descritivos

e literários, disponível para venda através de *download*. Entretanto, somente uma seção do corpus LOCNESS foi utilizada na presente pesquisa, ou seja, a seção de composições argumentativas. Tal decisão foi tomada a fim de haver compatibilidade entre o corpus de referência e corpus compilado, totalmente argumentativo.

A seção escolhida para formar o corpus de referência consistia de redações argumentativas de alunos norte-americanos de nível universitário (constituindo um total de 149.574 *tokens*) e redações argumentativas de alunos britânicos à época do A-level – exame semelhante ao ENEM, Exame Nacional do Ensino Médio – constituindo um total de 60.209 *tokens*. No total a seção perfazia 12.608 *types* (tipos de palavras diferentes) e 210.075 *tokens* (número total de palavras).

O corpus de aprendiz foco do presente estudo foi compilado a partir de trabalhos em inglês escritos por falantes de português brasileiro como língua materna. Para a coleta dos dados, foram compiladas 105 composições em língua inglesa escritas por brasileiros estudantes de inglês de nível avançado. As redações foram produzidas durante um momento de avaliação e os temas foram fornecidos aos alunos, de forma semelhante ao que ocorreu com o corpus LOCNESS. As 105 redações foram obtidas em escolas de idiomas da cidade do Rio de Janeiro. Este corpus de estudo consiste de 2.870 *types* (tipos de palavras diferentes) e 30.261 *tokens* (número total de palavras que compõem o corpus), um número inferior ao corpus de referência.

Para que fossem lidas por computador, as composições foram digitadas, houve correção de erros ortográficos e exclusão do título nos casos em que este tivesse sido fornecido pelo professor. Essa exclusão foi necessária para que o número de repetições lexicais não fosse inflacionada a partir desses títulos. Ao final do processo de digitação, cada uma das composições foi salva como arquivo individual (.txt) e recebeu uma numeração de acordo com o sexo, idade e escola de inglês de origem do aprendiz.

O software utilizado nesta pesquisa foi o programa *Wordsmith Tools*, versão 3.0. desenvolvido por Mike Scott. (1999). Para a realização deste trabalho em particular, o uso da ferramenta *Wordlist* – capaz de extrair listas de freqüências, estatísticas e percentagens – constitui-se como a principal fonte dos resultados obtidos.

4)_Análise dos dados

Retomando a pergunta mais abrangente feita no início do trabalho: “Como se compara a prosa argumentativa em inglês como LE e como Língua materna em termos de léxico?” podemos agora nos voltar para os resultados deste estudo concernentes aos

três aspectos específicos que foram analisados nos corpora em questão. O primeiro deles, que trata da obtenção dos 10 itens mais freqüentes em ambos os corpora, está ilustrado na tabela 1, abaixo:

Posição	LOCNESS	Corpus de redações de brasileiros
1º	the	the
2º	to	to
3º	of	and
4º	and	a
5º	a	is
6º	is	of
7º	in	that
8º	that	**I**
9º	be	in
10º	it	it

Tabela 1: 10 itens mais freqüentes em ambos os corpora

Analisando-se a tabela 1, é possível verificar-se que os 10 itens mais freqüentes são exatamente os mesmos, com uma única exceção: o pronome pessoal de 1ª pessoa do singular "I". Este ocupa o 8º lugar entre as 10 palavras mais freqüentes utilizadas pelos alunos brasileiros de inglês. Quando a posição deste mesmo pronome é aferida em relação ao corpus de redações argumentativas do LOCNESS, verifica-se que ele aparece na 28ª posição.

A segunda pergunta de pesquisa, sobre o comportamento dos 100 itens mais freqüentes em termos de sua distribuição percentual, é esclarecida com a observação dos dados da tabela 2. Esta tabela apresenta a distribuição percentual dos 100 itens mais freqüentes em relação ao vocabulário total. Aqui, os 100 primeiros itens têm sua distribuição apresentada em intervalos determinados de 1 a 100.

	LOCNESS	105 redações de alunos brasileiros
1 (the) 6,35	4,41	
1-10	24,77	24,56
1-30	36,83	40,63
1-50	42,99	48,96
1-70	47,01	54,95
1-100	51,14	61,23

Tabela 2: Distribuição percentual dos 100 itens mais freqüentes em cada um dos corpora em relação ao vocabulário total

A partir da Tabela 2, nota-se que o emprego das 100 palavras mais freqüentes, inicialmente similar quando considerados os 10 itens mais comuns, apresenta uma distribuição diferenciada para os dois grupos a partir do intervalo 1-30. Verifica-se que, a partir deste ponto, os conjuntos de palavras definidos dentro destes intervalos são consistentemente maiores para os alunos brasileiros. Ou seja, estes estudantes vão lançar mão destas palavras numa intensidade muito maior do que os nativos de língua inglesa, o que indicaria menos recursos lexicais à sua disposição.

Quando se considera a densidade lexical, proposta da terceira pergunta de pesquisa, há duas abordagens possíveis, como mencionado anteriormente na seção de Aporte Teórico. A densidade lexical pode ser calculada em sua forma simples ou em sua forma padronizada. Calculada em sua forma simples, ela é maior para o corpus de redações de brasileiros (9,48) em relação ao corpus de referência (6,00). De certa forma, este achado já era esperado, pois o número de palavras que compõem o corpus de redações brasileiras é significativamente menor (30.261) em relação ao número de palavras que compõem a seção do LOCNESS (210.070) considerada neste trabalho.

Conforme citado anteriormente, a densidade lexical foi calculada também em sua forma padronizada, com o objetivo de neutralizar a influência da discrepância no tamanho dos textos. Para tanto, foram definidos três intervalos para cálculo, de respectivamente 100, 500 e 1.000 palavras. A tabela 3 apresenta os resultados obtidos.

	LOCNESS	redações de alunos brasileiros
1-100	69,61	66,72
1-500	47,83	43,32
1-1000	39,99	35,19

Tabela 3: Valores obtidos através do cálculo das densidades padronizadas

É interessante notar que, quando calculada a densidade lexical em sua forma padronizada, ela é sempre maior no corpus de referência. Vale ainda mencionar que quanto mais o intervalo se aproxima do tamanho de um intervalo padrão (por exemplo, definido em 1.000 palavras no programa *Wordsmith Tools*) maior a diferença percentual entre as densidades obtidas.

5) Conclusões preliminares e encaminhamentos

Em relação aos 10 itens mais freqüentes em ambos os corpora, verifica-se que são as mesmas palavras de classe fechada. Entretanto, chama a atenção o uso pronunciado do pronome de 1ª pessoa singular "I" na escrita dos alunos brasileiros: 8º lugar versus 29º lugar para a seção analisada do LOCNESS. Tal uso marcado do pronome ´I´ em composições de natureza argumentativa é no mínimo estranho, revelando uma tendência dos aprendizes a se inscreverem no argumento. Um estudo mais aprofundado dos 'colocados'[1] (itens lexicais que aparecem na vizinhança do item de busca) do pronome pessoal 'I' talvez pudesse esclarecer as razões que determinam esta escolha por parte dos alunos brasileiros investigados. A presença de verbos de cognição como 'believe' e 'think', por exemplo, poderia sinalizar ausência de comprometimento em relação ao que é afirmado nestas redações. Outra alternativa, também passível de ser investigada através dos ´colocados' seria a utilização do pronome pessoal para ilustrar a inserção do 'eu' na argumentação através de narrativas de cunho pessoal, cujo objetivo é fornecer argumentos extras para o argumento.

Quanto ao emprego das 100 palavras mais freqüentes, uma comparação qualitativa indica que 75% delas são exatamente as mesmas em ambos os corpora estudados. Entretanto, uma comparação mais detalhada desses números sugere diferenças marcantes. Note-se que o emprego das 100 palavras mais comuns do vocabulário total é maior em 10% por parte dos alunos brasileiros em relação ao LOCNESS. Conforme mencionado na seção anterior, uma intensidade maior no uso das 100 palavras mais comuns por parte do grupo de alunos brasileiros estudados seria um indicativo de menos recursos lexicais. Estes, por desconhecerem as alternativas possíveis à disposição dos usuários de inglês como língua materna, lançam mão de um repertório mais limitado de palavras. Além disso o fazem numa intensidade maior do que os nativos de língua inglesa. Assim, as 100 palavras mais freqüentes são coincidentes nas duas listas, mas são mais utilizadas por alunos brasileiros do que por nativos de língua inglesa.

Finalmente, a densidade lexical é maior para o LOCNESS quando calculada em intervalos determinados, ou seja, quando calculada em sua forma padronizada. Estes resultados indicariam uma maior riqueza lexical na produção textual dos nativos de língua inglesa em relação aos textos dos alunos brasileiros, ao menos em termos dos corpora em questão.

Com relação aos encaminhamentos possíveis para o trabalho, estamos considerando a realização de estudos contrastivos que focalizem o emprego dos pronomes **"I", "my", "me", "our"**, por exemplo, no discurso escrito argumentativo em língua inglesa de alunos brasileiros e nativos de língua inglesa. A realização de estudos dos verbos lexicais e não lexicais presentes nos corpora também é um dos encaminhamentos possíveis. Por último, diríamos que atualmente nossos estudos apontam para um interesse maior sobre questões como vagueza e evidências de **overuse** (uso excessivo) e **underuse** (uso escasso) de itens lexicais em relação ao corpus de referência, estas últimas podendo ser verificadas através de uma abordagem estatística.

Bibliografia

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

GRANGER, S. *Learner English on computer*. London/New York: Longman, 1998.

PALTRIDGE, B. *Making Sense of Discourse Analysis*. NSW: Gerd Stabler, 2000

RINGBOM, H. Vocabulary frequencies in advanced learner English. In GRANGER, S. (org.)., *Learner English on computer*. London/New York: Longman 1998. p 41-52

SCOTT, M. *Wordsmith Tools*, versão 3.0. Oxford: Oxford UP, 1999.

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

SINCLAIR, J.; JONES, S. DALEY, R. The OSTI Report. In KRISHNAMURTHY, R. *English Collocation Studies*. London: Continuum, 2004.

TOGNINI BONELLI, E. *Corpus linguistics at work*. Amsterdam/Philadelphia: John Benjamins Company, 1996.