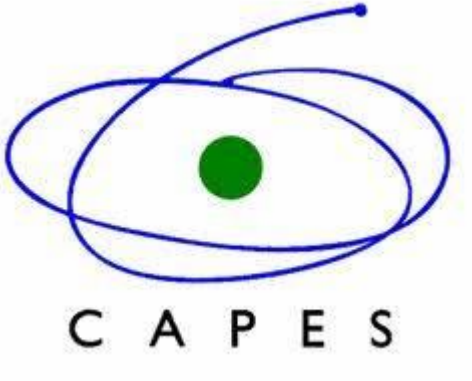
EXPLORANDO DISFLUÊNCIAS NO CÓRPUS COBRA-7, UM CÓRPUS EXTENSO DE PROCESS-WRITING DE APRENDIZES DE INGLÊS COMO LÍNGUA ESTRANGEIRA

Pesquisa de mestrado
Autor: Wendel Mendes Dantas
Orientador: Tony Berber Sardinha
Instituição: PUC-SP
Linha de pesquisa: Linguagem, tecnologia e educação
Início da pesquisa: 02/2010
Previsão de defesa: 02/2012
Andamento: análise piloto do corpus disponível
E-mail: wendel@bilinguismo.org
Grupo de estudos: http://corpuslg.org/gelc/gelc.php

Agradecimentos



Introdução

Há alguns anos temos observado a produção escrita de alunos brasileiros no idioma inglês. Porém, pouco tem sido investigado sobre quais os maiores problemas apresentados na produção escrita de alunos brasileiros nesse idioma e em qual proporção esses problemas são solucionados nos níveis seguintes.

Objetivo

O objetivo deste trabalho é verificar as características da linguagem de brasileiros na produção em inglês (dados de uso de linguagem) com vistas a informar a produção de materiais didáticos.

Corpora utilizados

Córpus de estudo: Córpus Brasileiro de Aprendizes de Inglês como Língua Estrangeira – Seven Idiomas (COBRA-7):

1. Composto por produções escritas (versões 1, 2 e 3) de aprendizes de ambos os sexos, adolescentes e adultos, nos níveis pré-intermediário, intermediário e intermediário superior da grande São Paulo.

2. Contém até o momento 572 redações (114.065 itens) criadas entre 2008 e 2009. Objetivamos coletar o mínimo de 1000 redações até o final do primeiro semestre de 2011.

3. A coleta foi feita através de acesso via nome de usuário e senha (*Web invisível*, cf. Bergh & Zanchetta, 2008, p. 311).

Corpora utilizados

Córpora de referência:

1. BNC (British National Corpus):

Contém 100 milhões de palavras e é composto de textos orais e escritos da variedade britânica do inglês colhidas a partir da segunda parte do século XX na Grã-Bretanha. Por ser distribuído, este córpus permite manuseio e análise utilizando-se diversas ferramentas, como, por exemplo, o programa computacional Wordsmith Tools 5.0 (Scott, 2008) e suas ferramentas.

Corpora utilizados

Córpora de referência:

2. COCA (Corpus of Contemporary American English):

Contém aproximadamente 400 milhões de ocorrências e é composto de textos orais e escritos da variedade norte-americana do inglês compilados a partir de 1990, incluindo textos ficcionais, revistas, jornais, periódicos acadêmicos, e transcrições de entrevistas de televisão e programas de rádio. Como se trata de um córpus *online* (http://www.americancorpus.org), a análise dos dados fica restrita aos meios disponíveis no *site*.

Corpora utilizados

Córpora de referência:
Quadro comparativos dos dois córpora de referência:

	BYU-BNC*	COCA**
Número de textos	4.052	Mais de 160.000
Número de ocorrências	Aproximadamente 100 milhões	Mais de 410 milhões
Número de itens	Aproximadamente 658.316	414.771.808

* Dados BYU-BNC cf. PINTO (2008).
** Dados COCA cf. Davies (2008) (linhas 1 e 2) e Mark Davies (comunicação pessoal) (linha 3).

Perguntas de pesquisa

(1) Quais são os padrões léxicogramaticais mais típicos no córpus COBRA-7?

(2) Quais são as frequências dos padrões léxicogramaticais típicos do COBRA-7 nos córpora de referência?

Metodologia

1. Escolha dos córpora de estudo e referência;

2. Compilação do córpus de estudo COBRA-7 (Córpus Brasileiro de Aprendizes de Inglês como Língua Estrangeira – Seven Idiomas);

3. Feitura de uma lista *Index* de palavras do córpus de estudo para cada nível utilizando o comando "*Make/Add to index"* da ferramenta *Wordlist*, do programa computacional Wordsmith Tools versão 5.0 (Scott, 2008) (doravante WST5);

4. Feitura de uma lista *Index* de palavras do BNC (*British National Corpus*) (córpus de referência) utilizando o comando "*Make/Add to index"* da ferramenta *Wordlist*, do programa computacional WST5);

Metodologia

5. Feitura de listas de *clusters* de três palavras para cada nível analisado e para o córpus de referência (BNC), usando a ferramenta *Wordlist*, do programa computacional WST5, comandos "*Compute/Clusters"* .

6. Feitura de listas de *clusters*-chave para cada um dos níveis analisados utilizando como córpus de referência o BNC, por meio da ferramenta *Keywords*, do programa computacional WST5.

7. Padrões dos *clusters*-chave: Observação dos padrões léxicogramaticais das 10 primeiras ocorrências dos *clusters* de três palavras na lista de *clusters*-chave do nível pré-intermediário, comparando-os com os dois outros níveis.

Metodologia

8. Como os 10 primeiros *cluster*s-chave da lista correspondente ao nível pré-intermediário não foram todos encontrados nos outros dois níveis pelo aplicativo *Keywords* do programa computacional WST5, realizou-se, então, uma busca desses *clusters*-chave para esses níveis utilizando o aplicativo *Concord*, do WST5.

9. Normalização dos resultados por 1000.

10. Busca dos *clusters*-chave no COCA, conseguidas digitando-os no campo de busca do *site* e selecionando-se a opção CHART, que fornece um gráfico estatístico dos subcórpora nos quais os *clusters* são encontrados.

Metodologia

11. Normalização dos resultados dos córpora de estudo e referência por 1000 para criar um quadro comparável adequado.

12. Observação dos resultados listados.

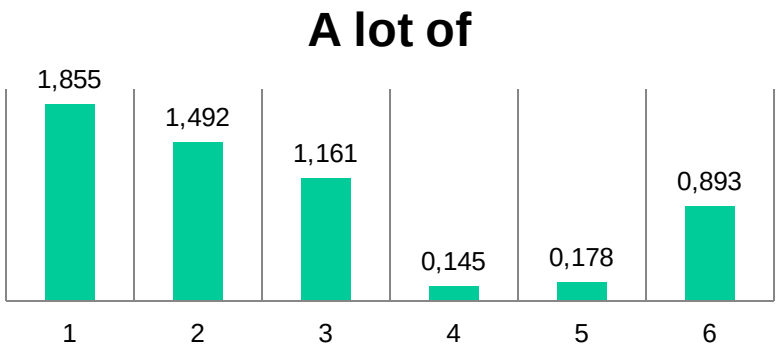
Piloto

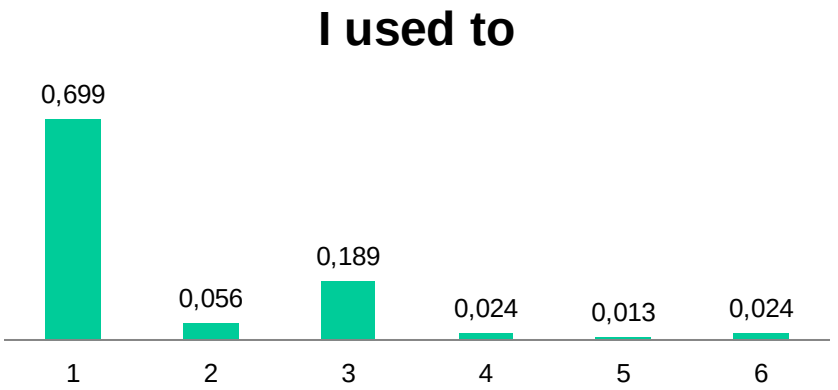
Padrões Léxicogramaticais presentes nas listas de *clusters*-chave da versão 1 das redações nos três níveis analisados no córpus de estudo (valores normalizados por 1000):

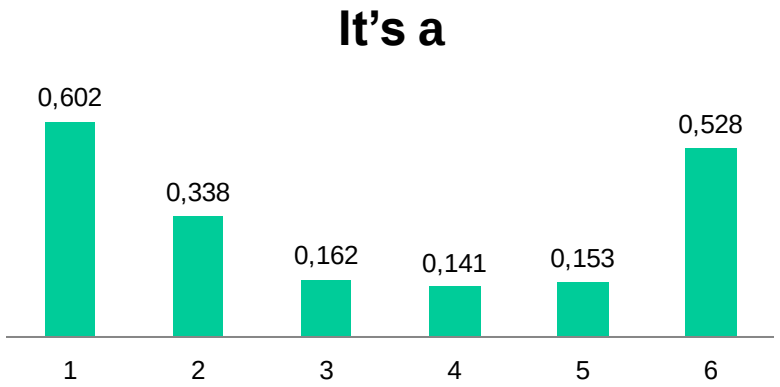
Padrões léxicogramaticais	Freq. COBRA-7			Freq. BNC	Freq. COCA*	
	Pre-inter	Inter.	Inter-sup		Escrito	Oral
Verbo + **A lot of** + subst. sing/pl	1,855	1,492	1,161	0,145	0,178	0,893
I used to + verbo	0,699	0,056	0,189	0,024	0,013	0,024
It's a + adj./subst.	0,602	0,338	0,162	0,141	0,153	0,528
I want to + verbo/fim de oração	0,554	0,507	0,459	0,048	0,060	0,243
Suj./to + **Go to the** + subst.	0,530	0,450	0,300	0,032	0,030	0,074
Verbo + **To work with** + obj..	0,506	0,056	0,081	0,010	0,015	0,026
I don't + verbo	0,506	0,676	0,486	0,357	0,312	1,010
I was very + adj./verbo particípio	0,506	0,113	0,189	0,007	0,004	0,018
It was a +subst./adj.	0,506	0,422	0,378	0,134	0,111	0,202
subst./pron./verbo + **To go to** + obj.	0,482	0,370	0,320	0,040	0,040	0,105

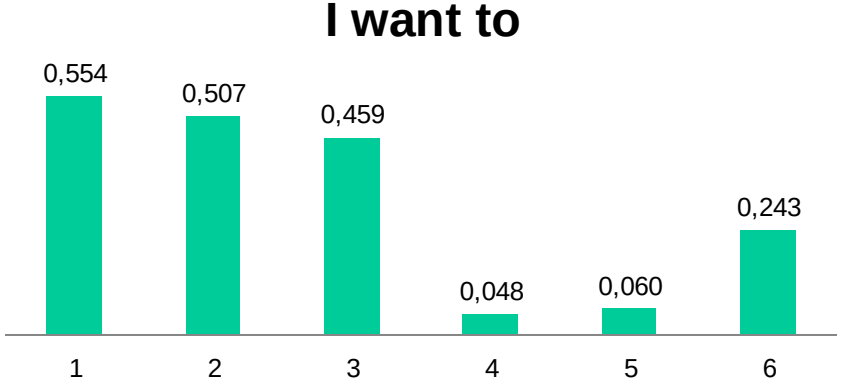
* Este valor foi dividido por 1000, pois o COCA é normalizado por 1 milhão.

Piloto

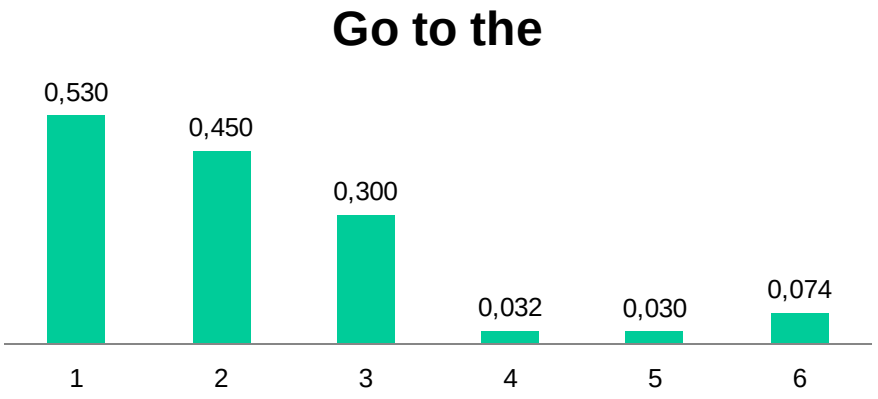








1 = Pre-intermediário
2 = Intermediário
3 = Intermediário superior
4 = BNC
5 = COCA (Escrito)
6 = COCA (Oral)



Piloto

Resultados auferidos até o momento:

Os dados apontam um uso dos *clusters*-chave pesquisados nas composições dos alunos superior àquele observado nas composições dos falantes nativos do inglês, coletados nos córpora de referência.

Mostram ainda que o uso dos *clusters*-chave pesquisados no córpus de estudo se aproxima mais da oralidade dos falantes nativos do inglês do que da produção escrita.

Porém, observa-se entre os níveis pré-intermediário e intermediário superior uma maior aproximação do padrão de ocorrência revelado pelos córpora de referência.

Referências bibliográficas

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

BERGH, G.; ZANCHETTA, E. Web linguistics. IN: LÜDELING, A.; KYTÖ, M. (Ed.). **Corpus Linguistics: an International Handbook**. Berlin/New York: Walter de Gruyter, 2008. p. 309-327.

DAVIES, M. **The Corpus of Contemporary American English (COCA): 410+ million words, 1990-present**. (2008-) Disponível em: <http://www.americancorpus.org> Acesso em: 04 out. 2010.

PINTO, M. V. **O uso de things, thing, anything, something e everything em corpora de aprendiz**. 2008. 221f. Dissertação (Mestrado em Linguística Aplicada) – LAEL – Programa de Estudos Pós-Graduados em Linguística Aplicada e Estudos da Linguagem, Pontifícia Universidade Católica de São Paulo, São Paulo, 2008. Disponível em: <http://www4.pucsp.br/pos/lael/lael-inf/teses/tese_marcia.pdf>. Acesso em: 04 out. 2010.

SCOTT, M. **Wordsmith Tools version 5**. Liverpool: Lexical Analysis Software, 2008.

Obrigado

wendel@bilinguismo.org