Instrumentos e atividades realizadas para a coleta de um *corpus* de aprendizes em língua inglesa para integrar o Br-ICLE (*Brazilian Portuguese Sub-corpus of ICLE*)

(Tools and activities used in data collection for a *corpus* composed of texts written by students of English language to be added to Br-ICLE (*Brazilian Portuguese Sub-Corpus of ICLE*))

Paula Tavares Pinto Paiva[1]

[1]Instituto de Biociências, Letras e Ciências Exatas –
Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP)

paula@ibilce.unesp.br; paulapaivaibilce@gmail.com

Abstract: This paper presents and discusses initiatives taken in a public university in the state of São Paulo in order to collect and organize a corpus composed of argumentative texts to be part of Br-ICLE – a subcorpus of ICLE[1] – which is composed of texts produced by Brazilian students enrolled in the courses of Arts and Languages and Translation. The discussion is based on a three-year project in which we could observe underuse and overuse features in the texts produced by Brazilian undergraduate students.

Keywords: corpus linguistics; learner corpus; underuse and overuse features of English.

Resumo: O objetivo deste trabalho é apresentar e discutir as iniciativas tomadas em uma universidade pública do estado de São Paulo com o intuito de coletar e organizar um corpus de textos argumentativos que ajudará a compor o Br-ICLE – o subcorpus do ICLE, composto por textos em língua inglesa produzidos por alunos brasileiros dos cursos de Licenciatura em Letras e Bacharelado em Letras com Habilitação em Tradução. A discussão tomará como base o desenvolvimento do projeto durante três anos de ocorrência, no qual puderam-se observar características de subuso e sobreuso nos textos dos alunos brasileiros.

Palavras-chave: Linguística de corpus; corpora de aprendizes; Características de subuso e sobreuso da língua inglesa.

Introdução

O objetivo deste trabalho é apresentar e discutir as iniciativas tomadas em uma universidade pública do estado de São Paulo com o intuito de coletar e organizar um *corpus* de textos argumentativos que ajudará a compor o Br-ICLE[2] – o *subcorpus* de textos em língua inglesa produzidos por alunos brasileiros. O projeto geral de compilação do Br-ICLE conta com a participação de equipes de pesquisa de várias universidades brasileiras, dentre elas, PUC-SP, USP, UERJ, UFMG, Unesp, Unicsul e UniCastelo.[3] Trata-se de um acordo feito com o Centro de Linguística de *Corpus* (doravante LC) de Inglês, da Universidade de Louvain, Bélgica, que tem coletado *corpora* de aprendizes de diversos países e visa a desenvolver estudos que descrevam a língua inglesa produzida por aprendizes de diferentes línguas maternas. No caso específico deste trabalho, objetivamos compilar um *corpus* que

1 Disponível em: < http://www.uclouvain.be/en-cecl-icle.html>. Acesso em: 13 fev. 2012.

2 Disponível em: < http://www2.lael.pucsp.br/corpora/bricle/index.htm>. Acesso em: 13 fev. 2012.

3 Para mais informações, acessar < http://www2.lael.pucsp.br/corpora/bricle/team.htm>. Acesso em: 13 fev. 2012.

possibilite a observação de características linguísticas da língua-alvo que são empregadas mais frequentemente (*sobreuso*) ou menos frequentemente (*subuso*) pelos aprendizes de língua inglesa dos cursos de Licenciatura em Letras e Letras com Habilitação em Tradutor, e compará-las à produção de falantes nativos de língua inglesa. A importância da realização de uma pesquisa como esta encontra-se em consonância à opinião de Biber *et al.* (1998), uma vez que os autores afirmam que um estudo sobre os desvios encontrados nas produções de aprendizes aumenta nossa compreensão sobre a aquisição de uma língua estrangeira, fornece dados para outras perspectivas relacionadas aos "erros" (ex. interlíngua e variedades não padrão da língua-alvo), e fornece evidências para decisões instrucionais. Uma das hipóteses a ser observada a fim de justificar a presença de traços de sobreuso[4] e/ou subuso seria a influência da língua materna dos aprendizes em suas produções escritas.

Para que o estudo fosse realizado, iniciamos, ao final do ano de 2009, a coleta dos textos produzidos pelos alunos dos cursos supracitados, que são enviados a uma das pesquisadoras responsáveis pela coleta e organização do Br-ICLE, Denise Delegá Lúcio. Importante ressaltar que os textos são enviados sem que nenhuma correção tenha sido feita, a fim de manter as características da escrita original dos alunos, entretanto, como a coleta é feita por nós no primeiro e no segundo semestre, descreveremos na seção de materiais e métodos as atividades de caráter didático-pedagógica realizadas por nós para que a escrita dos alunos seja melhorada com o passar do ano.

Fundamentação teórica

A observação e a descrição da linguagem em uso em diferentes contextos, como no de ensino e aprendizagem de línguas, têm sido favorecidas pela LC em virtude de sua abordagem teórico-metodológica apresentar um caráter interdisciplinar. Berber Sardinha (2004) destaca as vantagens que a LC oferece tanto para o professor quanto para o aluno, por permitir o acesso à tecnologia computacional e à linguagem utilizada nos textos contidos em um determinado *corpus*.

As aplicações da LC se fazem sentir tanto na área da Lexicografia, quanto nos estudos sistemáticos do uso da língua, em trabalhos dos Estudos da Tradução, Linguística Aplicada e em Processamento de Linguagem Natural. Tal pressuposto deve-se à proposta de Sinclair (1978, apud BERBER SARDINHA, 2004) ao afirmar que a língua deveria ser estudada a partir de exemplos reais de uso e não somente a partir de textos criados com propósitos de exemplificação. A interdisciplinaridade constitutiva dessa área de estudos tem possibilitado a troca de experiências e uma real colaboração entre profissionais de diferentes áreas do conhecimento.

Berber Sardinha discute em sua obra o papel da LC uma vez que esta abordagem:

> Ocupa-se da coleta e exploração de *corpora*, ou conjuntos de dados linguísticos textuais que foram coletados criteriosamente com o propósito de servirem para a pesquisa de uma língua ou variedade linguística. Como tal, dedica-se à exploração da linguagem através de evidências empíricas, extraídas por meio de computador. (BERBER SARDINHA, 2004, p. 3)

4 De acordo com Delegá-Lucio (2006, p. 2), "O sobreuso de palavras diz respeito à utilização de uma mesma palavra mais vezes do que um falante nativo normalmente o faria". O subuso seria o oposto, ou seja, um uso menos frequente de palavras que são comumente utilizadas pelos falantes nativos.

Importante destacarmos que o estudo ora proposto pretende descrever a linguagem do aprendiz de uma língua estrangeira, sendo assim, o *corpus* adequado para tal propósito é o *corpus* de aprendiz, cuja compilação e propósito é discutido por Delegá-Lucio ao afirmar que:

> Um corpus de aprendiz é aquele formado por textos naturais falados ou escritos por estudantes de uma língua estrangeira, que atenda a critérios que possibilitem seu estudo e que tenha sido coletado de modo que possa ser lido por computador. Os textos de um corpus de aprendiz são considerados naturais porque são produzidos por estudantes sem controle do que escrevem, ou seja, nenhum ponto gramatical (por exemplo) específico deve ser elicitado. Há, no entanto, critérios para a coleta desse corpus de acordo com o tipo de texto que se quer representar. (DELEGÁ-LUCIO, 2006, p. 21)

A diferença entre o *corpus* de aprendiz e o *corpus* nativo se dá pelo fato de os textos serem escritos por não nativos de uma língua estrangeira. Ao se utilizar o *corpus* de aprendiz, de acordo com a autora, pode-se mostrar aos alunos, por exemplo, o uso excessivo que fazem dos mesmos adjetivos, como mostra seu próprio estudo sobre relexicalização de adjetivos (DELEGÁ-LUCIO, 2006). Dessa forma, acredita-se que os alunos serão conscientizados sobre este problema. Ao mesmo tempo, ao se usar materiais informados por um *corpus* de falantes nativos, pode-se expor aos alunos às variedades de adjetivos que se associam aos substantivos para os quais há adjetivos sobreusados, ou seja, um corpus de aprendiz e um de textos de falantes nativos podem ser usados como instrumentos de aprendizagem de língua estrangeira.

Leech (1998) também destaca algumas questões que podem ser levantadas a partir da observação de *corpora* de aprendizes, dentre elas:

a) Quais características linguísticas da língua-alvo são empregadas com mais (sobreuso) ou menos (subuso) frequência em comparação com falantes nativos?

b) Qual é a extensão da influência da língua nativa (transferência) na produção dos aprendizes?

c) Em que áreas eles tendem a usar estratégias de evitação deixando de explorar a fundo o potencial da língua-alvo?

d) Em que áreas eles tendem a demonstrar desempenho nativo ou não-nativo?

e) Quais são as áreas nas quais os aprendizes de um dado país parecem necessitar de mais ajuda para desenvolver sua produção na língua-alvo? (LEECH, 1998, p. xiv)

Corpora de aprendizes são constituídos por textos de falantes não-nativos, os quais são utilizados para o estudo da linguagem que produzem. De acordo com Haan (1992), o tamanho ideal de um *corpus* de aprendizes vai depender do tipo de pesquisa a ser realizada. Ainda segundo o autor, um corpus de 20 mil palavras pode ser suficiente para se realizar um estudo. *Corpora* de aprendizes que variam de 20 a 200 mil palavras costumam ser mais específicos em relação a *corpora* maiores, no que se refere aos seus tópicos e gêneros discursivos. Para Granger (1998), 200 mil palavras foi o número ideal que se determinou para a constituição do ICLE (*International Corpus of Learner English*). De acordo com a autora, *corpora* de aprendizes maiores seriam mais destinados a pesquisadores que visam, por exemplo, a compilação de dicionários para aprendizes de uma língua estrangeira.

Tomando como base as questões destacadas acima, este estudo vem sendo guiado pelos dados fornecidos pelos *corpora* de aprendizes uma vez que procura identificar, com mais exatidão, em quais áreas os aprendizes brasileiros costumam ter mais dificuldade. Adicionalmente, citamos um estudo diacrônico baseado em *corpus*, e discutido por Biber *et al.* (1998), no qual foram evidenciados desvios de uso da língua inglesa utilizada por alunos estrangeiros em relação à concordância verbo-nominal e marcação do plural. Após quatro anos de acompanhamento desse estudo, os autores verificaram que houve uma diminuição considerável de erros relacionados à concordância de sujeito e verbo nas orações produzidas pelos aprendizes. No entanto, os erros ligados à morfologia verbo-nominal ocorriam mais frequentemente do que os erros de concordância. Mais recentemente, observamos o sobreuso do artigo definido "the" (ORENHA et al., 2007), que poderia ser considerada uma das evidências observadas nas produções em língua inglesa de alunos brasileiros e que, possivelmente, é influenciada pela língua materna dos aprendizes. Por esse motivo, temos observado tais características tomando como base estudos sobre a aquisição e aprendizagem de língua estrangeira que tratam do assunto (SHUMANN, 1992; MITCHELL; MYLES, 1998).

Ainda sobre estudos a partir de *corpora* de aprendizes, baseamo-nos em diversos trabalhos realizados no exterior e no Brasil (GRANGER 1993, 1998, 2009; BERBER SARDINHA, 2004; DELEGÁ-LUCIO, 2006; MEUNIER, 2011; DUTRA; SILEIRO, 2010, 2012).

Neste ponto, cabe destacar que, em termos de nomenclatura, embora alguns pesquisadores prefiram apresentar a diferenciação entre "*corpus-based*" (*corpus* usado para provar uma teoria ou posição a priori) e "*corpus-driven*" (*corpus* utilizado para permitir contraprova a posições iniciais assumidas pelos pesquisadores em geral) Beber Sardinha (1999) enfatiza que o primeiro termo, hoje em dia, tem sido mais utilizado para se referir às pesquisas a partir de *corpora* eletrônicos.

Metodologia e *corpora*

Os alunos selecionados para participar da coleta do Br-ICLE são regularmente matriculados nas disciplinas de Língua Inglesa III e IV dos terceiro e quarto anos, uma vez que os textos devem ser escritos por aprendizes em nível avançado. No início de cada ano letivo, os alunos conhecem o projeto Br-ICLE por meio de uma apresentação na qual eles têm contato com a terminologia específica da Linguística de *Corpus*, ou seja, os alunos que não conhecem esta abordagem teórico-metodológica passam a conhecer a definição de *corpus* e conhecer sua bases teóricas.

Em seguida, eles preenchem o Learner Profile (anexo) que descreverá seus perfis como aprendizes de língua inglesa. Em nossa universidade, os alunos fazem as redações em casa e utilizam todas as ferramentas que quiserem como dicionários mono e bilíngues, glossários e internet. Depois de preenchidos e assinados os formulários, os alunos escolhem dois temas para serem trabalhados no primeiro semestre, e dois para o segundo. Para tanto, eles recebem a lista dos tópicos a serem utilizados como tema para suas composições, conforme sugerido no projeto de Granger (1993). Os requisitos para a escritura das composições são: a) conter no mínimo 500 palavras e no máximo 1000; b) o aluno deve ter nível avançado de inglês; c) o aluno não pode ser nativo de língua inglesa.

Os temas sugeridos para as redações do Br-ICLE são:

(1) Crime does not pay

(2) The prison system is outdated. No civilised society should punish its criminals: it should rehabilitate them.

(3) All armies should consist entirely of professional soldiers: there is no value in a system of military service

(4) Most university degrees are theoretical and do not prepare students for the real world. They are therefore of very little value

(5) A man/woman's financial reward should be commensurate with their contribution to the society they live in

(6) In the 19th century, Victor Hugo said: "How sad it is to think that nature is calling out but humanity refuses to pay heed. "Do you think it is still true nowadays?

(7) Some people say that in our modern world, dominated by science technology and industrialisation, there is no longer a place for dreaming and imagination. What is your opinion?

(8) In the words of the old song "Money is the root of all evil"

(9) The Gulf War has shown us that it is still a great thing to fight for one's country.

(10) Feminists have done more harm to the cause of women than good.

(11) In his novel Animal Farm, George Orwell wrote "All men are equal: but some are more equal than others" How true is this today?

(12) The role of censorship in Western society.

(13) Marx once said that religion was the opium of the masses. If he was alive at the end of the 20th century, he would replace religion with television.

A seguir, apresentaremos a composição do *corpus* de aprendizes de nossa universidade conforme sua coleta, iniciado em 2010. Também apresentaremos algumas amostras das palavras mais frequentes, palavras-chave[5] e linhas de concordância,[6] que são analisadas por nós e pelos alunos durante as aulas de língua inglesa com o auxílio de *corpora* computadorizados.

Composição do *Corpus* em 2010 e algumas análises com o AntConc[7]

Abaixo mostramos a tabela 1 com os dez itens mais frequentes no primeiro subcorpus, composto por textos de alunos do terceiro ano do curso de Licenciatura em Letras:

5 Palavras-chave são estatisticamente mais altas e, portanto, relevantes no corpus de estudo em comparação a um corpus de língua geral.

6 Linhas de concordância são todas as linhas de um corpus de estudo que contenham uma palavra de busca inserida em seu contexto de uso.

7 Disponível em < http://www.antlab.sci.waseda.ac.jp/index.html>. Acesso em: 15 fev. 2012

Tabela 1: Número total de formas e itens do *corpus* de 2010 e os dez itens mais frequentes

Total No. of Word Types: 4635		
Total No. of Word Tokens: 8287		
1	4398	the
2	2654	to
3	2514	of
4	2362	and
5	1810	a
6	1740	is
7	1618	in
8	1420	that
9	890	it
10	886	are

Além das palavras que são, em sua maior parte, palavras funcionais, temos o número total de formas do *corpus*, isto é, o número total de palavras não repetidas, (4.635 formas), assim como o número total de palavras, chamadas aqui de itens (8.287).

Composição do *Corpus* em 2011 e algumas análises com o AntConc.

Similarmente ao *corpus* descrito anteriormente, no ano de 2011, coletamos as redações dos alunos do terceiro ano do curso de Licenciatura em Letras, conforme demonstrado na Tabela 2:

Tabela 2: Número total de formas e itens do *corpus* de 2011 e os dez itens mais frequentes

Total No. of Word Types: 4639		
Total No. of Word Tokens: 8752		
1	4399	the
2	2655	to
3	2517	of
4	2363	and
5	1811	a
6	1741	is
7	1619	in
8	1421	that
9	891	it
10	887	are

Neste ano, o número de formas (4.639) e itens (8.752) foi semelhante ao ano anterior.

Composição do *Corpus* em 2012 e algumas análises com o AntConc

Diferentemente dos *corpora* anteriores, como mudamos de *campus* e de curso, desta vez, a turma que atenderia ao perfil requisitado no Br-ICLE era o quarto ano do curso de Bacharelado em Letras com Habilitação em Tradutor. Abaixo apresentamos a tabela com os dados gerados pelo *corpus* composto por sua produção:

Tabela 3: Número total de formas e itens do *corpus* de 2012 e os dez itens mais frequentes

Total No. of Word Types: 4639		
Total No. of Word Tokens: 9217		
1	4400	the
2	2656	to
3	2520	of
4	2364	and
5	1812	a
6	1742	is
7	1620	in
8	1422	that
9	892	it
10	888	are

Embora o número de itens tenha sido mais alto no *corpus* composto por alunos de Tradução (9.217), coincidentemente, o número de formas foi o mesmo que o empregado no ano anterior pelos alunos de Letras (4639). O próximo passo, ao analisarmos os *corpora* detalhadamente, será o de analisar se o emprego lexical também foi semelhante.

Nesta fase do estudo temos trabalhado com os *corpora* separadamente, ou seja, com as listas geradas a partir das composições coletadas nos anos de 2010, 2011 e 2012. Entretanto, ao juntarmos os três subcorpora teremos um número total de itens de 26.256 itens, que seria considerado um tamanho aceitável para a realização de uma pesquisa com *corpus* de aprendiz, de acordo com o pesquisador Haan (1992), citado anteriormente.

Participação de Assistentes de Língua Inglesa (ETAs) na análise manual do Br-ICLE

No ano de 2010, recebemos em nossa universidade dois assistentes de língua inglesa (*English Teaching Assistants* – ETAs) pelo programa da Fulbright/Capes. Assim como o programa de leitores estrangeiros que auxiliam os professores de línguas estrangeiras no Departamento de Letras Modernas, os ETAs auxiliaram os professores da área de língua inglesa desempenhando várias funções, dentre as quais, destacamos a assistência na correção das composições dos alunos. Os ETAs também desempenharam um papel fundamental ao discutirem, nas aulas de língua inglesa, quais são as características mais frequentes nos textos dos alunos brasileiros de nossa universidade, que não soam naturais em língua inglesa. Nosso estudo nos mostrou, em primeiro lugar, que do total de treze temas a serem escolhidos pelos alunos brasileiros de nossa universidade, a maior parte tem optado pelos seguintes títulos:

a) Most university degrees are theoretical and do not prepare students for the real world. They are therefore of very little value;

b) Some people say that in our modern world, dominated by science technology and industrialisation, there is no longer a place for dreaming and imagination. What is your opinion?;

c) In the words of the old song “Money is the root of all evil”;

d) In his novel Animal Farm, George Orwell wrote “All men are equal: but some are more equal than others”. How true is this today?

Interessante notar que esses quatro temas foram repetidamente escolhidos pelos alunos no primeiro semestre dos três anos em que o projeto tem sido desenvolvido. Com o intuito de aumentar a variedade de temas utilizados e, por consequência, aumentar a diversidade lexical empregada em suas composições, pedimos aos ETAs que debatessem questões culturais da realidade estadunidense como, por exemplo, o sistema penitenciário e a pena de morte nos Estados Unidos, a seleção de jovens durante o alistamento militar, a participação dos Estados Unidos na Guerra do Golfo e outros temas mais gerais como o movimento feminista nos dias de hoje, a censura e as diferentes religiões que coexistem atualmente nos Estados Unidos.

Sabemos que a prévia discussão desses temas pode ter influenciado a escrita dos alunos em língua inglesa, assim como, tê-los incentivado a fazer uma busca mais refinada do vocabulário a ser empregado em suas redações, o que poderia de certa forma, invalidar os resultados finais do Br-ICLE. Entretanto, percebemos que as redações no início do projeto não eram bem escritas e que os temas se repetiam. A entrada dos ETAs no projeto foi justamente para que isso não acontecesse. Acreditamos que os alunos brasileiros puderam entrar em contato, de uma forma mais dinâmica e palpável, com a realidade cultural estadunidense por meio dos relatos e discussões com jovens que a vivenciaram e que tiveram seus familiares envolvidos em tais questões. Esse fato aguçou a curiosidade dos aprendizes brasileiros para a pesquisa de um vocabulário mais avançado a ser empregado em suas redações. Acreditamos que essa realidade acabou levando-os a, pelo menos, pensar sobre os diferentes títulos e, dessa forma, enriquecer seu léxico na língua estrangeira que estudavam. A nosso ver, essa "intervenção" indireta foi muito positiva para o desenvolvimento de um aprendizado mais autônomo, o que era o objetivo primordial de nossa investigação, ou seja, por meio da observação de "desvios", características de sobreuso e subuso, pretendíamos direcionar o ensino da língua inglesa a fim de suprir lacunas do conteúdo programático das disciplinas de Língua Inglesa III e IV.

Discussão

Até o momento, em uma breve análise das linhas de concordância, observamos o sobreuso do artigo definido "the" que, conforme destacado anteriormente, pode ter sido influenciado pela língua materna dos aprendizes, como, por exemplo, o uso do artigo com certas entidades geográficas que, na língua inglesa, não deveria ter sido usado.

Também observamos a influência da língua portuguesa nas colocações verbo-nominais, como em "have time to", "have classes", "have an appointment", "have chances" e "have opportunity to" ("ter tempo para", "ter aulas", "ter um compromisso", "ter chances" e "ter oportunidade para"). Este fato não é um problema, mas percebemos o subuso de outras possibilidades de combinatórias como em "have a laugh"e "have a look at", que são altamente empregadas na língua inglesa por seus nativos.

Na análise manual, ou seja, sem o auxílio de ferramentas computacionais, observamos a presença de "dangling modifiers", isto é, o emprego do particípio em posições erradas na oração e um desconhecimento do uso da pontuação na língua inglesa. O próximo passo será verificar esses itens no *corpus* geral com os programas de análise lexical.

Como uma possível solução para estes itens, temos apresentado aos alunos, por meio de atividades didáticas, linhas de concordância em *corpora* on-line a partir de textos

produzidos por nativos de língua inglesa que exemplificam o uso dessas estruturas. Dessa forma, os alunos entram em contado com uma diversidade maior de colocações verbo--nominais e com a estrutura da língua inglesa e se tornam cientes de outras possibilidades de combinatórias desconhecidas que são apresentadas junto ao seu contexto de uso.

Conclusões e encaminhamentos

Uma vez que o projeto de coleta do Br-ICLE, durante três anos, já está completo em nossa universidade, enviaremos todas as composições aos coordenadores do projeto no Brasil. Em seguida, continuaremos as análises do *corpus* compilado com o auxílio de ferramentas computacionais, conforme os passos elencados no trabalho de Dutra e Sileiro (2010) utilizando o *Error Tagger 1.0*, disponível em <www.corpuslg.org> e a análise mais aprofundada das palavras-chave com o auxílio do programa *WordSmith Tools 4.0* (SCOTT, 2004). Acreditamos que os dados apontados até o momento e as ações tomadas no sentido de discutir os temas do Br-ICLE e propor exercícios de pesquisa com *corpora* disponibilizados gratuitamente na internet já têm rendido frutos positivos para um aprendizado mais autônomo da língua inglesa.

REFERÊNCIAS

ANTHONY, L. *AntConc* Disponível em < http://www.antlab.sci.waseda.ac.jp/software.html>. Acesso em: 15 jul. 2011

BERBER SARDINHA, A. P. *Linguística de Corpus*. Barueri, SP: Manole, 2004.

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

DELEGÁ-LUCIO, D. *A relexicalização de adjetivos nas redações de alunos de inglês* – um estudo baseado em corpus de aprendiz. 2006. Dissertação (Mestrado em Linguística Aplicada e Estudos da Linguagem) – Pontifícia Universidade Católica de São Paulo. São Paulo.

DUTRA, D. P.; SILERO, R. P. Descobertas linguísticas para pesquisadores e aprendizes: a Linguística de Corpus e o ensino de gramática, 2010. *Revista Brasileira de Linguística Aplicada*, Belo Horizonte, v. 10, n. 4, 2010.

_____. O uso de For: uma análise de itens linguísticos em corpus de aprendizes brasileiros. In: SHEPHERD, T. M. G.; BERBER SARDINHA, T.; PINTO, M. V. (Org.) *Caminhos da linguistica de corpus*. Campinas: Mercado de Letras, 2012. p. 325-341.

GRANGER S. The International Corpus of Learner English. In: AARTS, J.; DE HAAN, P.; OOSTDIJK, N. (Org.) *English Language Corpora*: Design, Analysis and Exploitation. Amsterdam: Rodopi, 1993. p. 57-69.

_____. *Learner English on Computer*. London and New York: Longman, 1998.

_____. The contribution of learner corpora do second language acquisition and foreing language teaching: a critical evaluation. In: ALIJMER, K. (Org.) *Corpora and language teaching*. Studies in Corpus Linguistics 33. Amsterdam: John Benjamins, 2009.

HAAN, P. The Optimum Corpus Sample Size? In: LEITNER, G. (Org.) *New Directions in English Language Corpora*. Berlin: Mouton de Gruyter, 1992. p. 3-19.

LEECH, G. Preface – Learner corpora: what they are and what can be done with them. In: GRANGER S. *Learner English on Computer*. London and New York: Longman, 1998. p. xiv – xx.

MITCHELL, R.; MYLES, F. Functional/Pragmatic Perspectives on Second Language Learning. In: *Second language learning theories*. London: Arnold, 1998.

ORENHA, A.; PAIVA, P. T. P.; CAMARGO, D. C. O uso de corpora de aprendizes no ensino de produção escrita em língua inglesa. *MOARA*, v. 26, p. 281-293, 2007.

SCOTT, Mike. *WordSmith Tools*: Version 4. Oxford: Oxford University Press, 2004.

SHUMANN, J. La adquisición de lenguas segundas: la hipótesis de la pidginización. In: LICERAS, J. M. *La adquisición de las Lenguas Extranjeras*. Madrid: Visar, 1992. p. 123-141.

ANEXO

LEARNER PROFILE

==========

Text code: (do not fill in)
Essay: ……..……..……..……..……..……..
Title: ……..……..……..……..……..……..……....
Approximate length required: +500 words
Conditions: timed **( )** untimed **( )**
Examination: yes **( )** no **( )**
Reference tools: yes **( )** no **( )**

What reference tools?
Bilingual dictionary: ……..……..
English monolingual dictionary: ……..……..
Grammar: ……..……..
Other(s): ……..……..

==========

Surname: ……..…….. **First names**: ……..……..
Age: ……..…….. Male **( )** Female **( )**

Nationality: ……..……..
Native language: ……..……..
Father's mother tongue: ……..……..
Mother's mother tongue: ……..……..
Language(s) spoken at hom: (if more than one, please give the average % use of each)
……..……..……..……...
Education:
Primary school - medium of instruction: ……..……..
Secondary school - medium of instruction: ……..……..
Current studies: ……..……..
Current year of study: ……..……..
Institution: ……..……..
Medium of instruction: ……..……..
English only **( )**
Other language(s) **( )** (specify) _____
Both **( )**

==========

Years of English at school: ……..……..
Years of English at university: ……..……..

Stay in an English-speaking country:
Where? ……..……..
When? ……..…….. How long? ……..……..

==========

Other foreign languages in decreasing order of proficiency:
1……
2……
3.……
4.……

==========

I hereby give permission for my essay to be used for research purposes

Date: .....…..... Signature: ……..…….....