A PADRONIZAÇÃO DA LINGUAGEM E OS COLOCADOS: UMA ANÁLISE DE ERROS COMUNS APRESENTADOS POR APRENDIZES DE LÍNGUA INGLESA

Jane Marian[1]
Ronaldo Lima[2]

RESUMO

Entre as diversas áreas que contribuem para o ensino-aprendizagem de línguas estrangeiras, a Linguística de Corpus (LC) ocupa um lugar privilegiado, pois permite a localização automática de fenômenos a serem considerados. O objetivo deste artigo é apresentar instrumentos da LC e ferramentas computacionais que, permitindo acesso às colocações, coligações ou mesmo à prosódia semântica das palavras e frases, disponibilizam subsídios para o profissional deste ramo. Por meio da LC será sugerida uma metodologia de pesquisa que possa auxiliar mestres e aprendizes a discriminar usos paralelos àqueles oferecidos por suportes tradicionais, como as gramáticas prescritivas e normativas. Segundo Gavioli (2005), o ponto alto da LC é explicitar contextos de uso em situações que muitas vezes não poderiam ser explicadas apenas com as teorias gramaticais de uso cotidiano.

Palavras-chave: Ensino de Línguas Estrangeiras. Linguística de Corpus. Corpora. Colocados e Coligações.

[1] Doutora em Estudos da Tradução pelo Centro Universitário FAE. Professora da FAE Centro Universitário. *E-mail:* jane.marian@fae.edu

[2] Pós-doutor pela USP. Doutor em Sciences du Langage pela Université de Nice Sophia Antipolis. Universidade Federal de Santa Catarina -UFSC. *E-mail:* ronaldoearly@gmail.com

INTRODUÇÃO

Segundo Sardinha (2004), Tagnin (2005), Sinclair (1999) e Baker (1995), existem padronizações da língua que não podem ser explicadas pelas regras gramaticais, por exemplo: *merry christmas* e *happy christmas*. Seria inadequado falar *happy christmas*? De acordo com Tagnin (2005), existem convenções linguísticas que nos indicam que "*merry*" é mais adequado aos padrões da língua em determinada cultura. De fato, há explicações culturais para tais fenômenos. As variações e mudanças são fenômenos naturais que acabam cristalizando as novas composições. A lexicalização é um processo inexorável. A língua é uma entidade viva em constante mutação. Situação similar ocorre com *happy birthday* e *happy anniversary*. Para idade utiliza-se *happy birthday*, mas quando se trata do aniversário de casamento, usa-se *happy anniversary*.

O ensino-aprendizado de línguas estrangeiras se tornou uma disciplina acadêmica face à importância do setor. Todavia, trata-se de um campo permeado e que abrange todas as outras áreas do conhecimento. Não há um só domínio do conhecimento que não se desenvolva a partir do uso das linguagens e, por *default*, das línguas. Logo, muito mais do que multidisciplinar, o ensino-aprendizagem de línguas se caracteriza pela transdisciplinaridade. Neste sentido, a Linguística de Corpus, estando intimamente ligada com o processamento digital de dados de natureza linguística, contribui com suas especificidades para estudos intralinguísticos e interliguísticos.

Segundo Sardinha (2004), a LC é uma área que se ocupa da coleta e análise de textos eletronicamente armazenados, chamados de corpus (no plural: corpora). Estes devem ser coletados de acordo com os critérios da pesquisa que se vise realizar. No caso de estudos envolvendo duas ou mais línguas, naturalmente, caberá considerar cada um dos universos linguísticos envolvidos.

Para Sardinha (2004) o surgimento da LC se deu devido às necessidades de estudiosos em se apoiarem em contextos de uso para realizarem generalizações e constituírem teorias sobre o funcionamento da língua com base em suas estruturas linguísticas. Para Kitao (1994), anteriormente ao advento da LC, boa parte dos linguistas estudava a gramática por autoridade ou intuição de forma racionalista. Diferentemente, a LC surgiu para reforçar as mudanças de paradigma relativamente aos aspectos funcionais das linguagens. A abordagem empirista toma os dados da língua como um sistema probabilístico, ou seja, os traços linguísticos não ocorrem de forma aleatória, sendo possível evidenciar e quantificar padrões. Segundo Sardinha (2004) na Linguística de Corpus essa padronização se evidencia, como cita acima, pelas colocações, coligações e prosódia semântica. De acordo com Sinclair (1991), a LC se opõe à divisão tradicional entre léxico e gramática e considera que há um nível do sistema linguístico que engloba

tanto o vocabulário quanto a gramática, passando então a considerar a existência de dados de natureza léxico-gramátical.

Segundo Sardinha (2004), se a produção de textos envolve escolhas sobre o léxico e tais seleções estão obviamente ligadas a uma probabilidade, quanto maior a quantidade de palavras envolvidas no corpus maior a probabilidade de coocorrerem as palavras de baixa frequência. Sardinha (2004) define essas coocorrências de itens lexicais como colocados e coligações. Os colocados são as palavras que ocorrem com frequência significativa uma ao lado da outra. Por exemplo, *happy new year*, *merry christmas*, *credit card*, *police officer*, *pet shop*, *ice cream*, *pay attention*, *driver's license*, *hot dog* etc. E as coligações são itens lexicais e gramaticais que ocorrem repetidamente um ao lado do outro, por exemplo: *depend on*, *talk to*, *look at, kind of*, listen to etc.

Sinclair (1991) define colocação como a ocorrência de duas ou mais palavras no âmbito de um pequeno espaço entre elas em um contexto. Na colocação aparecem coocorrências que se repetem frequentemente ou que são estatisticamente consideráveis. A colocação é a evidência de que as palavras não se combinam por acaso. Sinclair (1991) acredita que as palavras não ocorrem aleatoriamente em um texto e que temos grande número de frases pré-construídas que constituem escolhas que fazemos quando elaboramos uma conversação. De acordo com Sinclair (1991), a significação das palavras no contexto afeta a escolha do léxico usado para o restante da frase.

A Prosódia Semântica, segundo Sardinha (2004) e Sinclair (1991) consiste em associações entre os itens lexicais e a conotação de campos semânticos. A prosódia prepara o interlocutor para que tipos de sons ou estrutura fraseológica que virão a seguir, ou seja, prepara o leitor para o conteúdo semântico que estará por vir. Esta ocorrência pode justificar a fluência ou não em determinada língua, já que os traços linguísticos e o léxico criam "relações de expectativas" que devem ser mantidas pelo falante para manter o padrão de naturalidade. Muitas vezes um falante nativo, ou fluente em certo idioma, sabe utilizar certa estrutura, todavia, não sabe explicar as razões por estar empregando determinada fórmula linguística. Essa padronização se justifica pelos colocados e coligações.

A prosódia semântica se justifica pelo campo semântico. Em certos casos, ao iniciar uma frase, muito antes de o falante enunciar o que pretende é possível prever o que segue. Isto significa que devido às suas escolhas lexicais iniciais já existem indícios suficientes para anunciar os colocados que o falante poderá adotar. As palavras apresentam campo semântico negativo, positivo ou neutro. Por exemplo, ela *sofreu um acidente/uma queda/por desgosto/ por amor* ou isso *causou* um acidente/danos/ uma *tragédia/ódio/prejuízo* etc. Nestes casos, a palavra "sofreu" e "causou" têm campo semântico negativo, pois estão relacionadas a palavras de valor negativo.

A descoberta de que uma palavra está padronizada, tanto lexical quanto gramaticalmente, baseia-se na estatística de concordâncias. As concordâncias são listas de orações extraídas de um corpus, completas ou não, nas quais uma ou mais palavras aparecem centralizadas. A palavra central recebe o nome de nódulo e o que fica à direita e à esquerda do texto chama-se de colocados ou coligações. As concordâncias processadas eletronicamente permitem a análise de um grande número de dados a partir dos quais é possível calcular a frequência de coocorrência de palavras. Este tipo de organização possibilita analisar dados e perceber fenômenos que não seriam detectáveis de forma analógica ou quando se examina pequenas porções de textos.

Apresenta-se, então, nas linhas que seguem, uma metodologia para que mestres e aprendizes se questionem a respeito dos "colocados" e, em medida similar, sobre fatos ligados à padronização da linguagem de forma empírica por meio de linhas de concordâncias que evidenciam contextos de uso efetivo da língua. Para esta pesquisa foram analisados três itens lexicais, a saber: *wrong, mistakes* e *error,* que Jacob (1999) apresenta em sua obra *Como não aprender inglês - erros comuns do aluno brasileiro*.

1 METODOLOGIA

A presente metodologia consiste na coleta de amostras de dados apresentados por Jacob (1999) na referida obra. Sobre este autor, cabe dizer que nasceu em Londres, Inglaterra, e após se formar em Engenharia Industrial, desembarcou no Brasil em 1967. Atuou em várias empresas e, em 1989, passou a se dedicar ao ensino da língua inglesa. A obra é baseada em suas experiências com o ensino da língua inglesa para brasileiros.

Segundo Sinclair (2004), os corpora de aprendizes, que são textos produzidos por alunos, cujo objetivo é justamente analisar os problemas apresentados por estudantes no processo da aprendizagem, é um fenômeno ainda recente e pouco explorado. Até recentemente ao discutir o uso de corpora se falava apenas de textos autênticos produzidos por falantes nativos. No entanto, a possibilidade de verificar de forma empírica os resultados do processo de aprendizagem tem gerado diversas pesquisas na área. Jacob (1999) apresenta os dados de seu livro a partir de suas experiências pessoais de forma intuitiva, o que é contrário às bases da LC, que oferece evidências concretas da língua em uso efetivo. O objetivo nessas linhas é de apresentar ferramentas que incitem o interesse pela LC face aos problemas que mestres e aprendizes enfrentam ao se deparar com frases ou colocados para os quais os instrumentos de suporte tradicionais não oferecem explicações gramaticais. Quando isto ocorre, normalmente geram-se bloqueios no processo de ensino-aprendizagem devido a essa padronização que Tagnin (2005) chama de convencionalizações de fatos da língua.

Após a etapa dedicada à escolha de um corpus, e posterior coleta dos dados, foi gerada uma busca por linhas de concordâncias relativamente às palavras destacadas. Para fazê-lo, recorreu-se a três plataformas disponíveis *online* com vistas à realização de análises linguísticas: (i) Lextutor, (ii) WebCorp e (iii) British National Corpus (BNC).

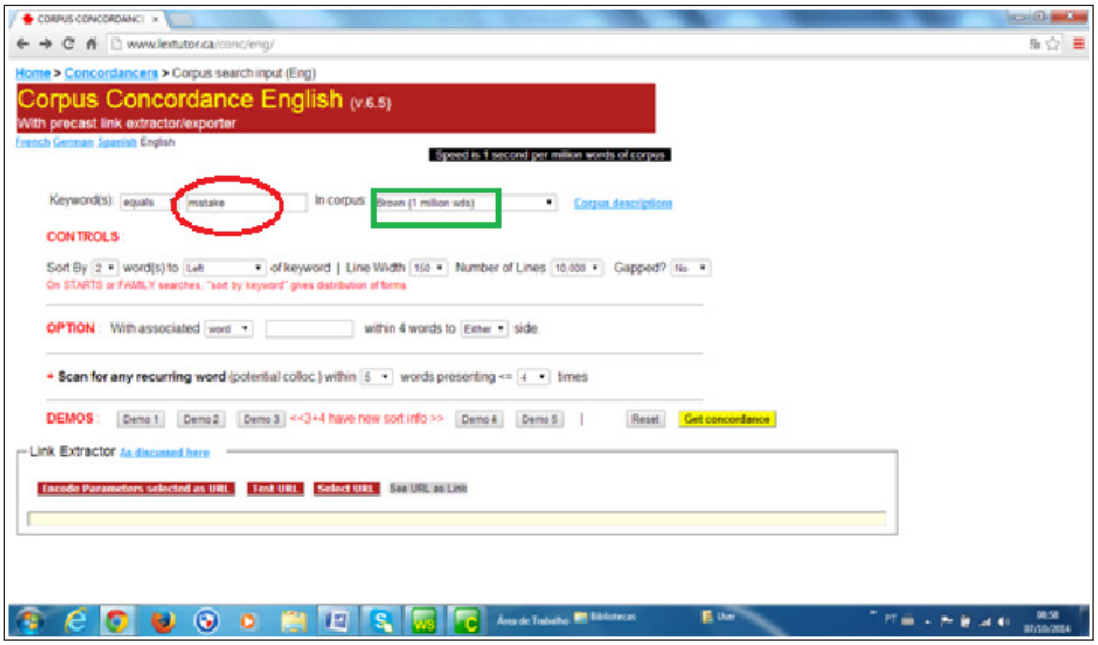
A plataforma Lextutor[3] apresenta, além do concordanciador de análises linguísticas, outras ferramentas para aprendizes, mestres e pesquisadores. Para acessar a ferramenta basta entrar no site e clicar em "*concordance*", logo abaixo de "*researchers*".

FIGURA 1 - Plataforma Lextutor Online

[3] Disponível em <http://www.lextutor.ca> Acesso em outubro de 2014.

Ao clicar em "*concordance*" será preciso selecionar a língua concernente que, no presente escopo, é o inglês. Após selecionar o ícone "*concordance*" surge uma interface, conforme apresentado a seguir.

FIGURA 2 - Interface Lextutor – *Concordance*



Em "*keywords*" digita-se a palavra requisitada e "*in corpus*" seleciona-se o corpus que se pretende analisar. A busca foi realizada a partir do corpus Brown, o primeiro corpus para pesquisas linguísticas compilado no final da década de sessenta por Kucera e Francis, na Brown University. O corpus contém 1 milhão de palavras e foi constituído através da compilação de vários tipos de textos. No centro aparece a palavra "nódulo" e ao redor os "colocados". A ferramenta computacional "*concordancer*" do Lextutor apresenta uma relação dos colocados à direita e à esquerda da palavra de busca. Além disso, se o pesquisador desejar obter dados em relação à sua fonte (origem) ou verificar o contexto no qual a palavra está inserida, basta clicar na palavra-chave e uma página com o contexto será aberta.

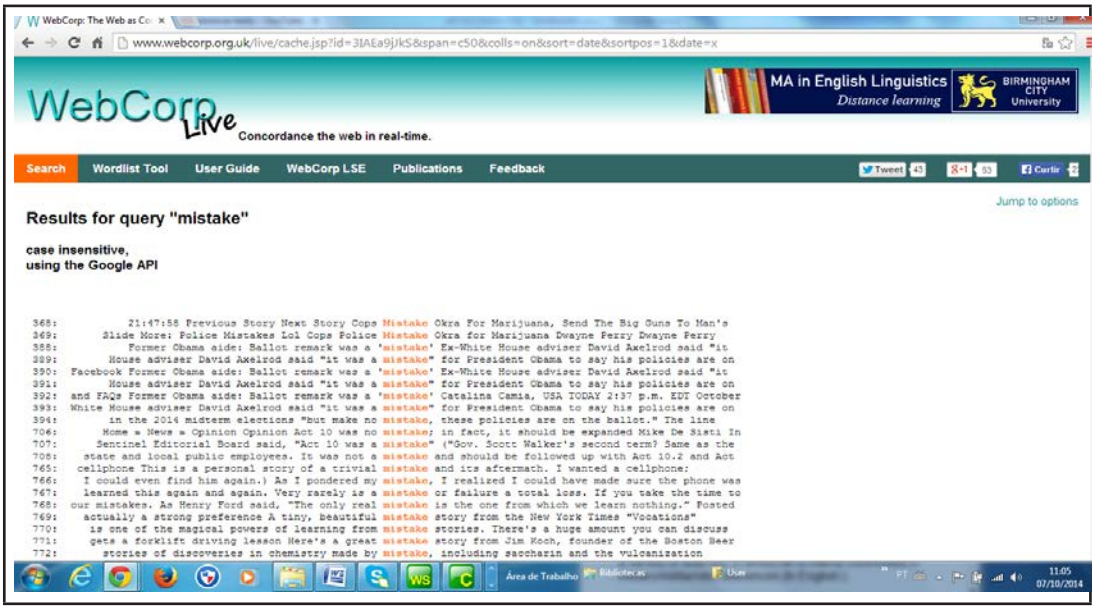
FIGURA 3 - Linha de Concordância - Lextutor

As linhas de concordância, segundo Alston (1995 apud SINCLAIR, 2004, p.18), podem destacar padrões de repetição ou variações em um texto, favorecendo também a análise de grandes quantidades de dados. Os corpora podem despertar *insights* e a curiosidade dos aprendizes na descoberta de novas possibilidades de usos e padrões da linguagem. O aprendiz, por sua vez, desenvolve papel de pesquisador e, de certa forma, torna-se responsável por investir em sua formação de forma autônoma. Ao professor, a seu turno, cabe o papel de orientar o aprendiz em seus novos processos de pesquisa, incitar a curiosidade por meio da motivação e permitir que os padrões da língua possam ser acessados de modo diferente, sobretudo em relação às mudanças de paradigmas em relação às práticas de ensino-aprendizagem tradicionais.

O segundo programa utilizado foi o concordanciador do webcorp[4], uma ferramenta computacional que faz uma busca em toda a *web* das ocorrências da palavra de busca. Essa ferramenta permite também visualizar a fonte dos corpora e apresenta algumas possibilidades de seleção de acordo com os objetivos da pesquisa.

4 Disponível em <http://www.webcorp.org.uk/live/> Acesso em outubro de 2014.

FIGURA 4 - Interface WebCorp – Linhas de Concordância



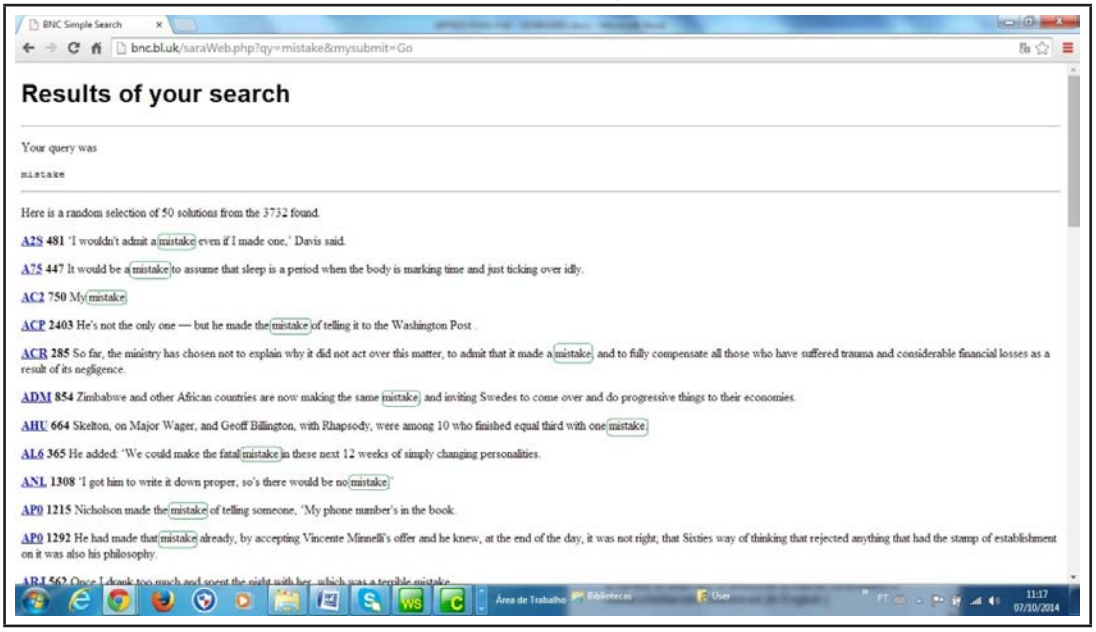
A terceira ferramenta de análises linguísticas é o *British National Corpus*[5] (BNC). Esta plataforma comporta o maior corpus compilado disponível gratuitamente. Somam-se cem milhões de palavras representativas tanto da fala quanto da escrita de textos como: revistas, jornais, artigos, etc. Nesta base de dados, a palavra de busca é digitada, conforme apresentado a seguir:

FIGURA 5 - Interface British National Corpus – BNC

[5] Disponível em <http://www.natcorp.ox.ac.uk/> Acesso em outubro de 2014.

A desvantagem desta plataforma é que a palavra de busca não aparece centralizada e sim dentro de um pequeno contexto apresentado. O pesquisador deverá correr os olhos para encontrar a palavra desejada. Todavia, os dados são confiáveis e, assim como nas outras plataformas, é possível verificar a fonte da qual o item faz parte.

FIGURA 6 - Resultados obtidos do British National Corpus - BNC



2 RESULTADOS

A análise de palavras-chave em contexto, associadas às linhas de concordâncias em que a palavra-chave fica centralizada, apresenta algumas vantagens. Sinclair (2004) diz que uma delas é não precisar ler toda a linha de concordância para procurar pela palavra-chave. Outra vantagem é a visualização textual, isto é, o pesquisador pode identificar facilmente padrões da linguagem, analisando os colocados e as coligações. Jacob (1999), ao distinguir as três palavras "*error*", "*mistakes*" e "*wrong*" apresenta alguns exemplos a partir de experiências de sala de aula.

> Excuse my wrongs, disse um aluno para mim em seu primeiro dia de aula. Um erro comum, pois o ideal seria: Excuse my mistakes. Wrong é adjetivo, e não um substantivo. (JACOB, 1999, p.28)

Analisando as linhas de concordância no WebCorp, Lextutor e BNC, verificam-se vários exemplos do item lexical "wrong" como adjetivo, por exemplo: "even when you have strong reason to believe a file contains *wrong* information, you have no right to check it"; "there was nothing *wrong* with them"; "but you're *wrong* about the rest of it" etc. Também percebeu-se que "*mistakes*" ocorre frequentemente como um substantivo, por exemplo: "Do you remember the *mistakes* she made?"; "meaning that *mistakes* can simply be over painted when required"; "we can begin with just one family and can learn from our *mistakes*" etc. As análises dos dados de forma empírica auxiliam na confirmação da hipótese de Jacob (1999), todavia, não foi encontrado nenhum exemplo de "*Excuse my mistakes*". Segundo Tagnin (2005), pode-se concluir que essa combinação não é convencionalmente utilizada no idioma. Jacob (1999) também fala que "*error*" e "*mistake*" não apresentam uma diferença muito grande, mas que "*error*" é menos grave do que "*mistake*". As linhas de concordância apontam que "*error*" está mais ligado a palavras das áreas exatas, como cálculos, computadores, erros humanos, políticos e falhas de impressão ou digitação. Por exemplo: "data not transfered due to a previous *error*"; "a computer *error* knocked out the system for the second time in a month", "… and in which the social and economic cost of a small calculation *error* can be great" etc. Segue abaixo uma tabela com os principais colocados de "*wrong*", "*mistakes*", "*error*":

QUADRO 1 - Lista de colocados e coligações – WebCorp/BNC/Lextutor

WRONG	MISTAKES	ERROR
right and wrong	serious mistake	message error
quite wrong	second mistake	computer error
am/are/is/were/was wrong	make/made/making mistakes	error 520
something wrong	tiny/big mistake	calculation error
nothing wrong	serious mistake	a grave error
completely wrong	number of mistakes	a serious error
totally wrong	learn from our mistakes	margin of error
simply wrong	same mistake	caused by human error due to human error
entirely wrong	political mistakes	translation error
absolutely wrong	past mistakes	risk of error
very wrong	several mistakes	a typing error
going/go/gone/goes/went wrong	few mistakes	printing error
wrong about	previous mistakes	technical error
the wrong adversary	avoid/no mistakes	mozilla/microsof error
the wrong approach	twenty mistakes	programming error
wrong message/decision/reasons	administrative mistakes	natural error

Os colocados e coligações apresentados anteriormente são os mais frequentes no âmbito dos corpora disponibilizados para o processamento das linhas de concordância, isto é, WebCorp, BNC e Lextutor. As três colunas apresentam, em sua maioria, colocados e coligações diferentes, evidenciando que dependendo do conteúdo semântico existe uma escolha lexical diferente. Essa percepção de colocados e campo semântico é automática, ou seja, quando se trata do idioma materno esse processo já está internalizado, no entanto, ao aprender um segundo idioma a evolução é mais lenta e, portanto, quanto mais exposição à linguagem natural, melhores as chances de internalização do processo ensino-aprendizagem.

CONCLUSÕES

Neste artigo foram apresentados três programas de processamento de linhas de concordâncias disponíveis online: o WebCorp, Lextutor e BNC. Foram examinados com vistas a verificar se podem constituir instrumentos de auxílio de aprendizes de línguas estrangeiras no que tange às dúvidas relacionadas às colocações, coligações e ao uso efetivo da língua falada ou escrita.

O presente estudo apresentou resultados relevantes de acordo com as pesquisas baseadas em corpora. Sinclair (1991), Sardinha (2004), Baker (1995) e Tagnin (2005) falam sobre a padronização da linguagem e as evidências que somente um corpus pode apresentar. Ao analisar os exemplos de Jacob (1999), percebe-se que mesmo para um falante nativo, citar exemplos e explicar a língua nem sempre é uma tarefa evidente ou elementar.

Essa pesquisa, de forma simples, teve como objetivo apresentar algumas ferramentas computacionais existentes, entre tantas outras, na perspectiva de motivar o desenvolvimento da autonomia e, sobretudo, da pesquisa com base em padrões envolvendo o uso de novas tecnologias.

REFERÊNCIAS

BAKER, Mona. **Corpora in Translation Studies**: An Overview and some Suggestions for Future Research. 1995, Target 7.2, 223-243.

GAVIOLI, Laura. **Exploring Corpora for ESP Learning**. Amsterdam: John Benjamins Publishing Company, 2005.

KITAO, Kenji. **Developing Resources for Corpus Linguistics**. Journal of Culture and Information Science, 1(1), -19, June/2004. Disponível em: < http://www.cis.doshisha.ac.jp/kkitao/library/article/corpus/resource.pdf> Acesso em: jul. 2010.

SARDINHA, Tony Berber. **Linguística de Corpus.** São Paulo: Manole, 2004.

SINCLAIR, John. **Corpus, concordance, collocation**. Oxford: Oxford University Press, 1991.

_____. **How to Use Corpora in Language Teaching**. Amsterdam: John Benjamins Publishing Company, 2004.

TAGNIN, Stella. **O jeito que a gente diz:** expressões convencionais e idiomáticas inglês e português. São Paulo: Ed. Disal, 2005.

_____. **Os Corpora: instrumentos de autoajuda para o tradutor**. Florianópolis: Cadernos de Tradução, 2002, v.9, n. 2002/1, p.191-213.


FAE