http://periodicos.letras.ufmg.br/index.php/textolivre
Ano: 2009 – Volume: 2 – Número: 2

ALIMENTAÇÃO DE UM BANCO DE DADOS DO SEMIOFON, VIA *CORPUS* NILC/SÃO CARLOS, COM PALAVRAS CUJAS VOGAIS MÉDIAS SE ENCONTRAM EM POSIÇÃO TÔNICA

Ana Cristina Fricke Matte/Universidade Federal de Minas Gerais
Adelma Lúcia de Oliveira Silva Araújo/Universidade Federal de Minas Gerais

RESUMO: Neste trabalho, apresentamos como recursos da web são usados em uma pesquisa sobre vogais médias desenvolvida pelo grupo SEMIOFON. Inicialmente, apresentamos brevemente o *corpus* NILC/São Carlos. Em seguida, detalhamos as etapas da pesquisa desde a coleta dos dados até a transcrição fonológica dos dados seguindo a conversão letra-fonema de Albano & Moreira (1996).

PALAVRAS-CHAVE: *Corpus* NILC/São Carlos. Vogais médias. Transcrição fonológica letra-fonema.

ABSTRACT: This paper presents how web resources are used to develop a research on a middle vowel investigation by the SEMIOFON group. Initially, the project AC/DC: corpus NILC/São Carlos is briefly described, then the different stages of our investigation is detailed, starting with the date collecting and finishing with the letter-phoneme phonological transcription according to Albano (1996).

KEYWORDS: Corpus NILC/São Carlos. Middle vowel. Letter-phoneme phonological transcription.

INTRODUÇÃO

O grupo de pesquisa Semiose e Fonoestilística (na época chamado SEMIOFON, atualmente Texto Livre: Semiótica e Tecnologia) da Universidade Federal de Minas Gerais (UFMG) desenvolve um projeto de alimentação de um banco de dados com palavras que contêm as vogais médias em posição tônica. Com o objetivo de levantar estes dados, realiza uma vasta pesquisa na internet em *corpora* do português online. Durante esta busca, chegou ao site do projeto/serviço Linguateca, cujo objetivo central é disponibilizar todos os *corpora* em seu sítio, inclusive o *corpus* do AC/DC, fonte de dados desta pesquisa.

Paralelamente ao objetivo central, o projeto/ serviço Linguateca, disponível em http://www.linguateca.pt/, propõe, também, melhorar as informações associadas a esses *corpora* ao desenvolver uma interface gráfica de simples utilização pelo usuário; facilitar a disponibilização e utilização dos recursos existentes via rede mundial de computadores; fornecer programas que facilitem a obtenção dos *corpora* através da Internet; criar *corpora* de tamanho suficiente que possam ser usados em pesquisas de grande magnitude, contribuir e facilitar a pesquisa e o processamento de *corpora* do português conforme declara Santos & Bick (2000).

O *corpus* NILC/São Carlos do Núcleo Interinstitucional de Linguística Computacional

da Universidade de São Paulo (NILC/USP), usado neste trabalho, tem aproximadamente 24 milhões de palavras coletadas a partir de textos do Jornal Folha de São Paulo, tratado pelo núcleo acima citado e disponibilizado através da Linguateca. Este núcleo dedica-se ao estudo do processamento computacional da língua portuguesa com sede no Instituto da Universidade de São Paulo em São Carlos. A formação integral de seu *corpus* contém textos de diversos gêneros coletados em jornais, material didático, epistolar e redações de alunos conforme atestam Rocha & Santos (2000). No site da Linguateca mostra-se a composição do projeto AC/DC:corpo NILC/São Carlos subdividindo em 03 diretórios: (a) macro: congrega os textos corrigidos, os semi-corrigidos e os não corrigidos; (b) micro: diferencia os textos por gêneros; e (c) subdiretórios: especifica os textos que contêm fontes diferentes. Associados aos subdiretórios existem, ainda, as pastas que categorizam especificadamente os textos. É desta forma que estão estruturados os *corpora* no sítio Linguateca. Dentre as organizações virtuais visitadas na busca de material para coleta e análise de dados e que trabalhassem com *corpora* do português, elegemos a Linguateca como sítio de busca para nossa pesquisa.

Segundo a descrição do próprio sítio da Linguateca, ela é uma organização virtual constituída por cinco pólos localizados em Oslo, Praga, Porto, Lisboa e Coimbra, ou seja, centros de pesquisa conhecidos internacionalmente e com uma vasta experiência em processamento da língua portuguesa. Através de seu sítio é possível fazer o levantamento de dados acessando os diferentes *corpora* nele existente (um por vez), com base em expressões de procura até que as informações linguísticas necessárias sejam encontradas. O resultado obtido pode ser, por exemplo, uma concordância em contexto dos objetos que satisfaz a expressão ou a distribuição, no *corpus* selecionado, desses mesmos objetos. A partir do conhecimento da composição dos *corpora* e da crescente oferta de dados para pesquisas em ambiente virtual, especialmente do português brasileiro, tem aumentado significativamente o fomento às pesquisas, facilitando e potencializando o trabalho de coleta de dados de muitos pesquisadores que necessitam manipular um volume vultuoso de dados.

1 APRESENTAÇÃO DO SÍTIO DA LINGUATECA

Na pesquisa, ainda em andamento, precisávamos para a formação do *corpus* de palavras que contivessem as vogais médias em posição tônica. Inicialmente, fizemos uma pesquisa em dicionário eletrônico que não se mostrou muito produtiva, pois, embora encontrássemos palavras com o ambiente vocálico pretendido, eram, em sua maioria, palavras com pouca frequência de uso, ou consideradas esdrúxulas. A fim de solucionarmos parcialmente este problema, resolvemos procurar em *corpora* online o ambiente pretendido através do critério frequência da palavra. Por meio de consulta pelo site de busca Google, chegamos ao site da Linguateca. Inteiramo-nos das possibilidades de uso e começamos as tentativas de busca do corpus almejado. O escolhido foi o NILC/São Carlos, por conter uma amostra significativa de dados do português brasileiro. Na página principal da Linguateca, encontram-se descritas tanto toda a estrutura funcional quanto a história do desenvolvimento deste vultoso projeto. Para compreenderem um pouco a metodologia de busca utilizada, mostraremos o passo-a-passo de como fizemos nossa busca e coleta de dados.

Inicialmente, entramos na página da Linguateca utilizando a URL: http://www.linguateca.pt/. Posteriormente, clicamos em > acesso a recursos, do lado esquerdo da tela. Em seguida, do mesmo lado, clicamos no nome do corpus desejado: AC/DC. Uma janela pertencente a esse corpus se abriu e à sua esquerda lemos o título: acesso a corpus de português:

Projeto AC/DC, seguida de uma breve introdução e, novamente, de outro título: breve descrição do corpus feito por meio de tabela. Dentro dessa, escolhemos o corpus com o qual trabalhamos - NILC/São Carlos – clicando sobre ele. Por fim, abrimos outra janela à direita para filtrarmos a busca dentro do próprio corpus especificado.

Como muitas são as palavras do acervo, optamos por buscar todas as formas do verbo reunir marcando no corpus a distribuição de formas deste verbo. Para a busca é necessário saber as nomenclaturas utilizadas. Neste caso, inserimos a seguinte nomenclatura [lema="reunir"], que é a notação para procura do lema reunir neste corpus. No passo seguinte, marcamos o resultado pretendido pela distribuição de formas. Um detalhe interessante é que caso seja importante este resultado em ordem alfabética devemos marcar a opção referente. O quadro após a marcação ficou da seguinte forma:

Para seguir a explicação, necessário se faz entrar na página da Linguateca através da seguinte url: <http://www.linguateca.pt/>. Abre-se uma página e no seu lado esquerdo deve-se clicar em > acesso a recursos, em seguida, do mesmo lado, clica-se no nome do corpus desejado. Nossa escolha ateve-se ao AC/DC. Uma janela pertencente a esse corpus se abrirá e à sua esquerda lê-se o seguinte título: acesso a corpus de português: Projeto AC/DC. Tem-se uma breve introdução e a seguir o título: breve descrição dos corpus feito por meio de tabela. Dentro dessa, escolhe-se o corpus com o qual se quer trabalhar – NILC/São Carlos – clicando-se sobre ele. Uma outra janela é aberta à direita para sua busca já dentro deste corpus. A página encontrada tem este formato de busca. Para demonstrar a facilidade de se trabalhar com corpora online segue um exemplo de busca.

Escolhi para minha busca todas as formas do verbo reunir marcando no corpus a distribuição de formas deste verbo. Para a busca é necessário saber as nomenclaturas utilizadas. Neste caso inseri a seguinte nomenclatura [lema="reunir"], que é a notação para procura do lema reunir neste corpus. O passo seguinte foi fazer a marcação do resultado pretendido pela distribuição de formas. Um detalhe interessante é que caso seja importante este resultado em ordem alfabética deve-se marcar esta opção. O quadro após minha marcação fica da seguinte forma:

Procurar:
Resultado:
Concordância
Distribuição das formas
Distribuição dos lemas
Distribuição da categoria gramatical (PoS)
Distribuição do tempo verbal e/ou do caso pronominal
Distribuição de pessoa e/ou número
Distribuição do gênero
Distribuição da função sintática
Opções
Resultados por ordem alfabética (só distribuições)
Fonte: http://www.linguateca.pt/ACDC/

No resultado instantâneo desta procura, apareceram 6991 ocorrências de frequência de distribuição de formas com o lema "reunir", conforme especificado no resultado a seguir. Vale salientar que esta pesquisa foi realizada por amostra quantitativa, por palavras, e sem distribuição de categoria gramatical, apenas por distribuição de lemas.

Resultado da busca:
Tue Apr 21 03:56:07 WEST 2009

Procura: [lema="reunir"].
Pedido: Distribuição das formas
Corpus: NILC/São Carlos v. 8.0
6991 ocorrências.

Na distribuição apareceram 45 valores diferentes de forma do verbo reunir com frequências que variaram de 2374 presenças no *corpus* para o verbo reunir na terceira pessoa do singular e de apenas 1 presença atestada para a forma desse verbo no plural.

2 RESULTADO E DISCUSSÃO DOS DADOS

A título de exemplo detalharemos uma busca, no *corpus* linguístico do AC/DC, por palavras com a letra "x" no ambiente final. Inserimos em *procura* a notação ".*x.*" e marcamos como seleção *a distribuição das formas*. Obtivemos uma lista de palavras com a respectiva frequência que cada uma fora atestada no *corpus* pesquisado. Qual é a importância da frequência da palavra em um estudo desta natureza? O estudo da frequência da palavra é de extrema relevância por mostrar através dos resultados quantitativos a relação que se pode estabelecer entre a frequência de uma palavra e a mudança linguística, por exemplo, das normas gramaticais, como afirma Biderman (1998, citado por Berber Sardinha, 2004, p.163), "a norma linguística nada mais é do que a média dos usos frequentes das palavras que são aceitas pelas comunidades dos falantes".

O *corpus* NILC/São Carlos oferece dois sistemas de busca ou consulta: o primeiro, refere-se ao *pedido de concordâncias*, mais precisamente às frases reais do *corpus* que caracteriza o fenômeno que se pretende pesquisar; e o segundo, ao *pedido de distribuição de formas*, ou seja, às quantidades detectadas de um ou de vários fenômenos existentes no *corpus* inteiro. Para sua obtenção, marca-se concordâncias, distribuição e frequências simples e/ou complexas, como esclarecido no exemplo anteriormente apresentado.

O acesso ao banco de dados pode ser feito segundo a sua composição, a partir dos seus dados quantitativos, pela contabilização de multipalavras, mas também através de sua distribuição gramatical. Observe, nas tabelas a seguir, os exemplos dos argumentadores na coleta do *corpus* por dados quantitativos e por classe gramatical presente no banco de dados do NILC/São Carlos que trabalhamos nesta pesquisa.

Argumento	**Nº de formas**	**Nº de tipos**
Unidades	1198015	68784
Palavras	997695	68138
Palavras em minúscula	7570013	37765
Palavras com inicial maiúscula	122771	19250
Palavras todas em maiúsculas	7422	1539

Números	8655	777
Palavras com números	1146	522
Palavras mistas	1452	960
Pontuação	68106	637

Tabela 1: Dados Quantitativos
Fonte: http://www.linguateca.pt/ACDC/

Categoria gramatical	**Expressão de procura**	**Número de tokens**	**Em percentagem**
Substantivos	N|N[^U].*	204299	22,89%
Verbos	V.*	132470	14,84%
Adjetivos	ADJ.*	57646	6,46%
Pronomes pessoais	.*PERS.*	15539	1,74%
Preposições	PRP.*	176541	19,78%
Conjunções	K.*	40590	4,55%
Advérbios	ADV.*	53817	6,03%
Determinantes	.*DET.*	183852	20,60%
Numerais	.*SPEC.*	17511	1,96%

Tabela 2: Distribuição por categoria gramatical
Fonte: http://www.linguateca.pt/ACDC/

2.1 Para se fazer uma busca no corpus AC/DC existe uma nomenclatura específica para procura?

Na tabela 2, acima, vê-se na segunda coluna *expressão de procura* várias nomenclaturas utilizadas na busca dos dados, ou seja, há sim uma linguagem específica, mas que nos parece flexível ao entendimento conforme exposto no tutorial IMS-CWB disponível na URL: http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench. Os dados devem ser manipulados com base nos nomes dos atributos (expressão de procura) e dos seus valores, esses identificadores foram atribuídos pelos desenvolvedores do programa ao codificar os *corpora*. Santos (2000) afirma que no banco de dados do projeto Linguateca existe dois tipos de atributos: os estruturais e os posicionais. Estes estariam relacionados à categoria gramatical (POS) ou ao lema (LEMA) e aqueles aos conjuntos de frases (s) e conjunto de parágrafo (p). Esse último estaria relacionado à categoria gramatical (POS) ou ao lema (LEMA). Cada um desses atributos está descriminado nas páginas de contabilização dos diversos *corpora* do Projeto. Tanto os nomes dos atributos posicionais quanto os valores que estes atributos podem assumir foram determinados no programa e podem ser analisados nas páginas de anotação do AC/DC. Para um maior aprofundamento sobre essa sintaxe no português vide Bick (2000).

Com o objetivo de explorar o potencial do *corpus* NILC/São Carlos, fizemos algumas buscas no acervo desses dados. Como exemplo, citamos algumas notações que estão descritas na tabela 3 , a seguir.

Argumento	Procurar como
Palavras que contenha “x”	“.*x.*”
“x” em início de palavras	“x.*”
“x” em final de palavras	“.*x”.
Formas do verbo reunir	[lema=”reunir”]

Tabela 3: Exemplos de busca no *corpus* NILC/São Carlos
Fonte: http://www.linguateca.pt/ACDC/

2.2 Procedimentos utilizados para obtenção da lista de palavras sobre vogais médias no *corpus* do NILC/São Carlos

Inicialmente, é preciso comentar que cada serviço oferecido pelos *corpora* da Linguateca tem uma vasta documentação e um ambiente amigável facilitador da interação entre o programa e o usuário. De fácil explicação, os artigos citados na biblioteca do site dão suporte ao pesquisador em suas buscas e as formas mais eficientes de como proceder desde a etapa de busca até a obtenção de dados. No entanto, este projeto está em pleno desenvolvimento fazendo com que muitos casos ainda não sejam encontrados em seus *corpora*.

Dessa forma, pode acontecer de um pesquisador não chegar aos dados de sua busca pelos seguintes motivos: (a) há dados que ainda não se encontram catalogados no acervo; (b) há dados que mesmo que estejam catalogados encontram-se em uma quantidade irrelevante para uma pesquisa mais aprofundada, por exemplo, dados referentes a um *corpus* infantil; e (c) há dados que se encontram no acervo mas não com o nome especificado pelo pesquisador em sua busca, por exemplo, os verbos intransitivos. Neste caso, deve-se seguir uma forma alternativa de busca que seria a procura de contextos em que um dado verbo é usado de modo intransitivo, vale salientar que se pode obter algumas vezes como resultado um conjunto de verbos que o serviço considerou como sendo intransitivo, mesmo não o sendo. Este projeto está em pleno desenvolvimento fazendo com que muitos casos ainda não sejam encontrados nos corpora. Como exemplo de procura não confirmada pelos corpora do projeto podemos citar a presença dos verbos intransitivos. Neste caso, deve-se seguir uma forma alternativa de busca que seria a procura de contextos em que um dado verbo é usado de modo intransitivo, vale salientar que se pode-se obter algumas vezes como resultado um conjunto de verbos que o utilizador. O que fazer neste caso? Rocha & Santos (2000) aconselham que, neste caso específico, deve-se esquecer a rapidez e agilidade dos dados colhidos online e se selecionar manualmente os resultados confirmados. Para se evitar situações como essas, aconselha-se, também, utilizar apenas argumentos que sejam reconhecíveis pelo utilizador fazendo com que o mesmo possa detectar na medida certa o que você quer/precisa procurar.

Para a busca específica dos dados necessários a nossa pesquisa, fizemos uma procura sistemática e específica apenas das palavras com o contexto fonético requerido, vogais médias

abertas ㄱㅇㄷ e ㄱㅞㄷㄹ. Essa busca, na verdade, nos deu uma listagem das vogais abertas inseridas na sílaba tônica das palavras em ordem alfabética e por frequência. Para realizarmos essa busca, seguimos as recomendações descritas anteriormente utilizando o argumento correto: para a busca da vogal aberta "e" utilizamos o argumentador **.*é.+** e na busca da vogal média aberta "o" escrevemos o argumentador .***ô.+,** obtendo-se, por conseguinte, a coleta dos dados com as vogais médias abertas. Como proceder, então, para a obtenção das palavras com vogais fechadas ㄱㅌㄷ e ㄱㅞㄷㄹ na sílaba tônica da palavras? Como a Linguateca não possui em seus *corpora* os dados anotados/classificados em termos de sílaba, tivemos que fazer uma busca geral manualmente e selecionar os dados que julgássemos relevantes para nossa pesquisa. Nesta busca, utilizamos os seguintes anotadores: **.+e.+** para o ㄱㅌㄷ e **.+o.+** para o ㄱㅞㄷ no meio da palavras. Os dados selecionados estão assim constituídos: para o anotador ".+e.+"**.** encontrarmos 248980 ocorrências, sendo 31159 valores diferentes de forma, com uma frequência de 10.040. Na busca do anotador ". +o.+". verificamos 1.114 diferentes formas, com 217.645 ocorrências. Já a distribuição do anotador ".+ó.+" nos revelou um valor de 25.383 diferentes formas, totalizando 9.185 ocorrências. Finalmente, chegamos a distribuição do ".+é.+", com 10.040 ocorrências e 1.114 valores diferentes de forma.

Concluída a pesquisa e com o nosso *corpus* já estabelecido, passamos à fase de transcrição, a qual segue o esquema de conversão arquissegmental letra-fonema, proposto por Albano & Moreira (1996, p.3). Conforme as autoras o objetivo principal deste tipo de transcrição é o de representar as distinções lexical e pós-lexicalmente, embora deixe a realização fonética dos segmentos altamente variáveis relativamente aberta. Para a adoção do sistema fonológico de escrita, fizemos a escolha da transcrição fonológica seguindo a decisão de Matte et al. (2006). As razões que embasam esta escolha são: (a) a possibilidade real de automatização da transcrição fonológica da escrita ortográfica; e (b) a ortografia portuguesa ser quase totalmente fonêmica, o que torna sua conversão uma tarefa de fácil execução, exceto, segundo Albano & Moreira (1996), naqueles casos em que a alofonia é percebida onde os níveis fonéticos e morfológicos conduzem a uma ambiguidade nas análises fonológicas. De acordo com essas autoras a conversão letra-fonema objetiva tratar apenas os níveis de análises considerados abstratos, sem querer, entretanto, discutir a questão levantada pela Fonologia dos anos 70 de como deveria ser esse nível de representação. Com esta conversão letra-fonema transcrita, pretendemos alimentar um banco de dados com os anotadores para vogais médias presentes e possíveis de serem encontrados no *corpus* do NILC/São Carlos.

3 CONSIDERAÇÕES FINAIS

Neste trabalho, apresentamos as possibilidades que se abrem para a pesquisa linguística quando usamos os recursos disponíveis na rede mundial de computadores. Descrevemos, detalhadamente, o passo-a-passo da pesquisa sobre vogais médias realizada com o uso de *corpora* online. Deixamos claro os inúmeros aspectos positivos de se realizar pesquisas através de tarefas com um volume de dados gigantesco de um banco virtual de dados de modo seguro e eficiente. Trabalhamos os dados graças a este fácil acesso aos *corpora* e a sua disponibilidade de uso.

Além dessas vantagens, adicionamos o fato de que a análise de *corpora* como esses permitem adicionalmente a descoberta de elementos novos, citamos aqui uma análise minuciosa nos detalhes oferecidos pelo valor das frequências da palavra, pois, dependendo do tópico a ser

analisado, essa frequência poderá ajudar a desmistificar crenças linguísticas preestabelecidas. Outro fato relevante que destacamos é que ao se fazer pesquisas com *corpora*, ajuda-se, também, a mudar o velho paradigma já preestabelecido de que os linguistas "fogem" da análise quantitativa e da estatística.

Finalmente, gostaria de utilizar este espaço para agradecer a valiosa ajuda concedida pela Drª Diana Santos, da Linguateca – PT, que sempre solícita ajudou-nos a esclarecer todos os questionamentos, deixamos aqui, portanto, nossos sinceros agradecimentos.

REFERÊNCIAS BIBLIOGRÁFICAS

ALBANO. E. & MOREIRA, A.A. Archisegment-based letter-to-phone conversion for concatenative speech synthesis in Portuguese. In: *International Conference on Spoken Language Processing*, 1996, Filadelfia. Proceedings ICSLP 96. Filadelfia - EUA, 1996. v. 3. p. 1208-1711. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.16.3931>. Acessado em 10 de setembro de 2009.

BICK, Eckhard. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Dr. Phil. Thesis. Aarhus University. Aarhus, Denmark: Aarhus University Press. November 2000.

MATTE, A; MEIRELES, A; FAGUAS, C. *SIÇWeb – syllabic -accentual phonological parser of witten texts.* Rev. Est. Ling., v.14, n.1, p.31-50, jan./jun.2006.

ROCHA, Paulo Alexandre & SANTOS, Diana. "CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa". In Maria das Graças Volpe Nunes (ed.), *V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000)* (São Paulo, Brasil, 19-22 de Novembro de 2000), São Paulo : ICMC/USP, p. 131-140.

SANTOS, Diana. O projecto Processamento Computacional do Português: Balanço e perspectivas. In: Maria das Graças Volpe Nunes (ed.). *V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000)* (São Paulo, Brasil, 19-22 de Novembro de 2000), São Paulo : ICMC/USP, pp. 105-113, 2000. Disponível em: <http://www.linguateca.pt/documentos/SantosPROPOR2000.pdf>. Acessado em 10 de setembro de 2009.

OKSEFJELL, Signe & SANTOS, Diana. Breve panorâmica dos recursos de português mencionados na Web. In: Lima, V. L. S. de (ed.). *III Encontro para o Processamento Computacional do Português Escrito e Falado* (PROPOR'98), Porto Alegre, RS, 3-4 de Novembro de 1998, p. 38-47.

PINHEIRO, G. M & ALUÍSIO, S.M. *Córpus Nilc: descrição e análise crítica com vistas ao projeto Lacio-Web.* NILC-TR-03-03, fevereiro 2003.

ROCHA, P. A & SANTOS, D. CETEM Público: Um corpus de grandes dimensões de linguagem jornalística portuguesa. In: NUNES, M. G. V. (ed.). *V Encontro para o processamento computacional da língua portuguesa escrita e falada* (PROPOR 2000), São Paulo, ICMC/USP, pp. 131-140, 19-22 de Novembro de 2000. Disponível em: <http:// http://www.linguateca.pt/documentos/RochaSantosPROPOR2000.pdf>. Acesso em: 13 de setembro de 2009.

SANTOS, Diana & COSTA, Luís. *A Linguateca e o projecto Processamento Computacional do português. Terminómetro - A terminologia em Portugal e nos países de língua portuguesa em África*, nº 7, 2005, p. 63-69. Disponível em: <http:///www.linguateca.pt/documentos/SantosCostaTerminometro2005.pdf>. Acesso em: 13 de

setembro de 2009.
SANTOS, Diana & BICK, Eckhard. Providing internet access to portuguese corpora: the AC/DC project. In: GAVRILIDOU, M; CARAYANNIS, G; MARKANTONATOU, S; PIPERIDIS, S; STAINHAUER, G. (eds.). *Proceedings of the Second International Conference on Language Resources and Evaluation* (LREC 2000), Atenas, Grécia, 31 de Maio a 2 de Junho de 2000, p. 205-210. Disponível em: <http://www.linguateca.pt/documentos/SantosBickLREC2000.pdf>. Acesso em: 14 de setembro de 2009.
SARDINHA, Tony Berber. *Linguística de Corpus*. Barueri, SP: Editora Manole, 2004.