ESTUDO DE FERRAMENTA COMPUTACIONAL DE ANÁLISE DE *CORPORA* APLICADA À TERMINOLOGIA: ANTCONC

ESTUDIO DE HERRAMIENTA COMPUTACIONAL DE ANÁLISIS DE *CORPORA* APLICADA A LA TERMINOLOGÍA: ANTCONC

Ana Rachel Salgado[1]

RESUMO: O presente artigo teve origem no trabalho final da disciplina de Tecnologias Linguísticas I, do doutorado em Linguística Aplicada (UNISINOS), que teve por objetivo aplicar os princípios da Linguística de *Corpus* relativos à compilação e anotação de *corpora*. Para isso, foi elaborado um *corpus* piloto de artigos científicos da área de Psiquiatria, usando por base a Revista de Psiquiatria do Rio Grande do Sul, disponível *online*. O *corpus* foi previamente analisado utilizando a ferramenta Antconc, a fim de testar suas funcionalidades e verificar sua adequação para a tarefa de reconhecimento terminológico, para posterior aplicação na análise de *corpora* que embasa a tese da autora.

PALAVRAS-CHAVE: linguística de *corpus*, ferramentas de análise de *corpora*, tecnologias linguísticas, compilação de *corpora*, análise de *corpora*.

RESUMEN: Este artículo se originó del trabajo final de la asignatura de Tecnologías Lingüísticas I, del doctorado en Lingüística Aplicada (UNISINOS), que tuvo por objetivo aplicar los principios de la Lingüística de *Corpus* relativos a la compilación y anotación de *corpora*. Para ello, se elaboró un *corpus* piloto de artículos científicos del área de Psiquiatría, usando por base la Revista de Psiquiatria do Rio Grande do Sul, disponible en línea. El *corpus* fue previamente analizado utilizando la herramienta Antconc, a fin de probar sus funcionalidades y verificar su adecuación para la tarea de reconocimiento terminológico, para posterior aplicación en el análisis de *corpora* que embasa la tesis de la autora.

PALABRAS-CLAVE: lingüística de *corpus*, herramientas de análisis de *corpora*, tecnologías lingüísticas, compilación de *corpora*, análisis de *corpora*.

INTRODUÇÃO

O presente artigo teve por objetivos aplicar os princípios da Linguística de *Corpus* relativos à compilação e anotação de *corpora* e testar uma das ferramentas computacionais de análise de *corpora* disponíveis, a fim de verificar sua eficiência não só quanto à *interface*, mas principalmente quanto aos resultados da extração de candidatos a termos.

Tendo em vista tais objetivos, para a realização dos testes foi escolhida a ferramenta AntConc, por se tratar de um *software* livre. Além disso, o AntConc tem a vantagem de ser um arquivo bastante leve (apenas 4Mb) e dispensa a necessidade de instalação – o que,

1 Doutoranda em Linguística Aplicada pela Universidade do Vale do Rio dos Sinos (UNISINOS) – São Leopoldo, RS, Brasil. Mestre em Estudos Linguísticos pela Universidade Federal do Rio Grande do Sul (UFRGS). Bacharel em Letras, ênfase Tradução Português-Espanhol, pela UFRGS. Tradutora e revisora. E-mail: ar.salgado@terra.com.br.

parece-nos, torna seu uso mais fácil mesmo para usuários que estejam começando seus estudos em Linguística de *Corpus*. Tais características o tornam interessante também para o uso em aula, pois é possível rodá-lo em qualquer computador, já que há versões para Windows, Mac e Linux e, pelas características mencionadas anteriormente, é um programa que não exige demais do sistema.

Além da escolha da ferramenta, foi necessário também compilar um *corpus* piloto para a realização dos testes. A partir de minha experiência de trabalho como tradutora de artigos de psiquiatria e psicanálise, e visando contemplar também o tema do projeto de tese[2], foram selecionados 19 artigos da Revista de Psiquiatria do Rio Grande do Sul para compor o *corpus* piloto.

Para a compilação do *corpus*, foram seguidas as etapas propostas por Aluísio e Almeida (2006, p. 159-160) e, para a análise prévia, a metodologia da Linguística de *Corpus* (BIBER, 1993; BERBER SARDINHA, 2000; SINCLAIR, 2005).

Antes de passar ao detalhamento da compilação do *corpus* e da ferramenta, cabe destacar aqui que, para o presente artigo, não havia o objetivo de gerar uma lista de candidatos a termo propriamente dita, mas sim de colocar em prática os conhecimentos teóricos relativos à compilação de *corpora*, bem como de testar uma ferramenta de análise, a fim de verificar sua adequação (ou não) ao posterior uso na pesquisa para a tese.

1 COMPILAÇÃO DO *CORPUS*

De acordo com Aluísio e Almeida (2006, p. 159-160), a compilação de um *corpus* possui três etapas principais, quais sejam:

> 1) o projeto do *corpus*, que inclui a seleção dos textos e os cuidados com os requisitos que foram discutidos na seção anterior [autenticidade, representatividade, balanceamento, diversidade[3]]; 2) compilação (ou captura), manipulação, nomeação dos arquivos de texto e pedidos de permissão de uso e 3) anotação.

Seguindo tais etapas, no *site* Scielo.br (http://www.scielo.br), foi selecionada a *Revista de Psiquiatria do Rio Grande do Sul* (RPRS), uma publicação da Sociedade de Psiquiatria do Rio Grande do Sul. A escolha desta revista, em particular, ocorreu em função de meu trabalho

2 Na época, o projeto de tese estava voltado para o reconhecimento terminológico em artigos de psiquiatria e psicanálise. Posteriormente, foi feito um recorte e, agora, a tese enfoca apenas o reconhecimento terminológico em artigos de psicanálise.

3 Inserção nossa, conforme critérios expostos por Aluísio e Almeida (2006, p. 158-159).

como tradutora para revistas das áreas de psiquiatria e psicanálise, além da facilidade de acesso aos artigos, pois há várias publicações *on-line* reconhecidas na área de psiquiatria. Inicialmente, o objetivo era pesquisar termos da área de psicanálise, mas não foram encontradas, em português, publicações de livre acesso *on-line* reconhecidas nesta área de especialidade. Outro critério de escolha foi o fato de o conteúdo da RPRS estar licenciado por uma licença Creative Commons (http://creativecommons.org/licenses/by-nc/3.0/deed.pt_BR), ou seja, pode ser livremente copiado, distribuído e retransmitido, desde que mediante atribuição clara da autoria/licença e de forma não comercial.

Na página da RPRS estão disponíveis revistas dos anos de 2003 a 2011 (volumes 25 a 33), havendo uma média de três números anuais e um suplemento. Para o presente estudo, foram escolhidas as três revistas do ano de 2010. Destas revistas, inicialmente foram selecionados apenas os textos escritos em português[4], publicados na seção "Artigos Originais". Entretanto, após uma breve análise de textos publicados em outras seções, e levando em consideração o critério de balanceamento proposto por Sinclair (2005) e a observação de Aluísio e Almeida (2006, p. 173), pareceu interessante incluir todos, em função da variedade de estilos (artigo original, artigo de revisão, editorial, carta ao editor, etc.) – o que pode trazer uma maior riqueza para a pesquisa quando for tratada a questão dos termos em contexto.

Assim, chegou-se a um total de 19 textos, distribuídos da seguinte forma:

Nº Revista	Artigo Especial	Artigo Original	Artigo de Revisão	Editorial	Relato de Caso	Resenha
01	-	03	01	01	-	01
02	-	02	01	01	-	01
03	01	03	01	01	01	01

Tabela . **Número de textos publicados em cada seção por número da RPRS.**

Realizada a seleção dos textos, passou-se à etapa de compilação propriamente dita. Os textos foram copiados e salvos em formato *plain text* (.txt), havendo sido excluídos os seguintes elementos: títulos traduzidos, resumos (bem como suas traduções), tabelas, quadros, figuras, algarismos, agradecimentos, declarações de conflitos de interesses e referências bibliográficas. Para a nomeação dos arquivos, foi usado o critério a seguir:

[4] A revista conta, também, com textos publicados em inglês e textos traduzidos do inglês, os quais não foram incluídos em nossa pesquisa por não se enquadrarem no critério de autenticidade, conforme detalhado em Sardinha (2000, p. 338-339).

Sigla da revista	Ano publicação	Número	Tipo de Texto	Número Texto[5]
RPRS	2010	01	AO	01

Tabela . **Esquema de nomeação dos arquivos do *corpus*.**

Para a nomeação dos arquivos, o único critério de classificação utilizado para tipo de texto foi a seção da revista onde o texto foi publicado. Assim, para fins de arquivamento e posterior recuperação de informação, temos as seguintes siglas: artigo especial (AE), artigo original (AO), artigo de revisão (AR), editorial (ED), relato de caso (RC) e resenha (RES).

Os textos selecionados perfizeram um total de aproximadamente 41.000 *tokens* (número total de palavras do *corpus*). Consideramos que, para o presente trabalho, essa amostra cumpra com o requisito de representatividade (BIBER, 1993; BERBER SARDINHA, 2000; SINCLAIR, 2005), uma vez que se trata de um estudo-piloto com o objetivo de testar as funcionalidades das ferramentas de análise de *corpus*.

A etapa seguinte foi a anotação do *corpus*. Para o presente trabalho, foi realizada apenas a edição manual dos cabeçalhos, contendo as seguintes informações: código do arquivo (conforme esquema de nomeação de arquivos exposto anteriormente), título, autor(es), referência (fonte, volume, número, local e ano da publicação), *link* e número total de palavras do texto. A seguir, um exemplo de cabeçalho utilizado:

```
<head>
<name>RPRS-2010-01-ED</name>
<title>A falácia da adequação da cobertura dos Centros de Atenção Psicossocial no estado do Rio Grande do Sul: comentário</title>
<author>Fernando Lejderman</author>
<ref>Rev. psiquiatr. Rio Gd. Sul, v. 32, n. 1, Porto Alegre, 2010</ref>
<link>http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0101-81082010000100001&lng=pt&nrm=iso&tlng=pt</link>
<ntoken>678</ntoken>
</head>
```

Cumpridas as três etapas de compilação do *corpus*, foi realizada a análise prévia deste, utilizando a ferramenta e AntConc. Para um melhor resultado na geração de listas de palavras, foi utilizada uma lista de *stopwords* (palavras gramaticais e outras palavras muito frequentes

5 Esta informação só foi utilizada quando havia mais de um texto do mesmo tipo, por exemplo, no caso dos artigos originais.

que o programa deve ignorar). Essa lista foi baixada do blog "Text Mining", disponível em http://miningtext.blogspot.com/2008/11/listas-de-stopwords-stoplist-portugues.html.

2 A FERRAMENTA ANTCONC

O AntConc é um *freeware*, desenvolvido por Lawrence Anthony e disponível para *download* em http://www.antlab.sci.waseda.ac.jp/software.html em versões para Windows, Mac e Linux. Após baixar o arquivo, que tem aproximadamente 4 Mb, não é necessário instalá-lo – basta dar um clique duplo no ícone que já aparecerá a tela inicial do programa (Figura 1).

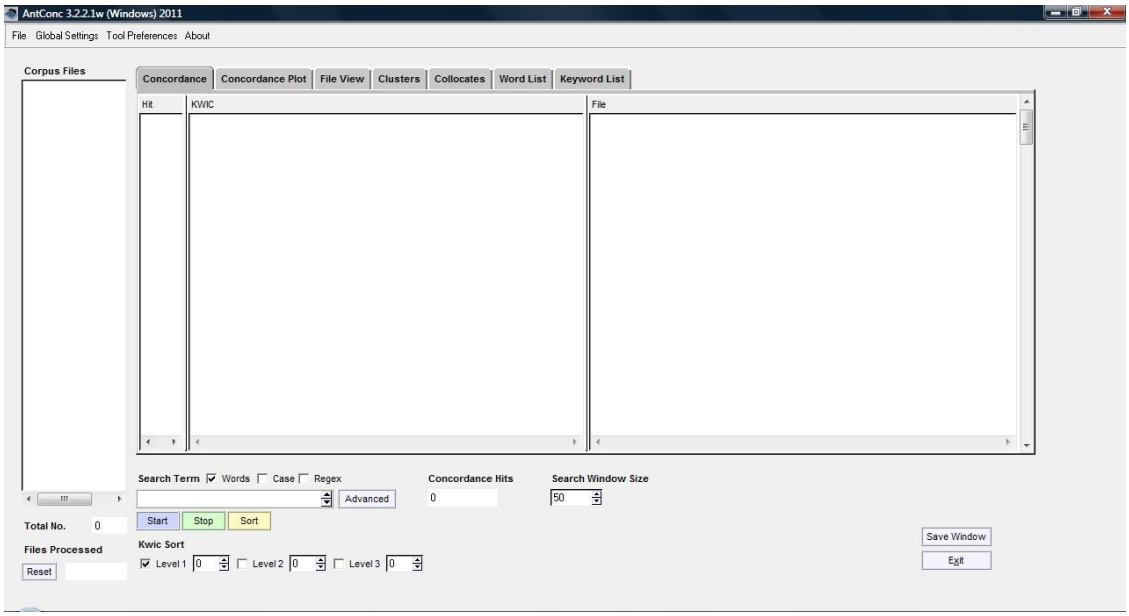

Figura . **Tela inicial do AntConc.**

A interface do programa é bastante simples e, em uma mesma janela, é possível navegar por diferentes opções de análise. Para iniciar o trabalho, é necessário carregar os textos do *corpus*, através do menu *File – Open Dir*, que abre uma janela de navegação por pastas como as do Windows Explorer. Selecionada a pasta, basta clicar em OK que os textos são automaticamente carregados. Os nomes dos arquivos aparecerão no quadro ***Corpus Files***, à esquerda da tela (ver Figura 1).

Para a inclusão de uma *stoplist*, é necessário acessar o menu *Tool Preferences*, opção *Word List* e, no campo *Word List Range Options*, selecionar *Use a stoplist listed below*. A inclusão pode ser feita inserindo palavra por palavra no campo *Add Word* ou inserindo um

arquivo no campo *Add Words from File* (opção aqui utilizada). Depois de selecionado o arquivo, clicar em *Apply*.

É importante lembrar de marcar, no campo *Other Options*, a caixa *Treat all data as lowercase* – caso contrário, o programa irá diferenciar entre maiúsculas e minúsculas, o que causará problemas de exaustividade na geração da lista de palavras (por exemplo, *depressão* e *Depressão* seriam entendidas como palavras diferentes e gerariam duas entradas na lista). A tela *Tool Preferences* pode ser vista na Figura 2:

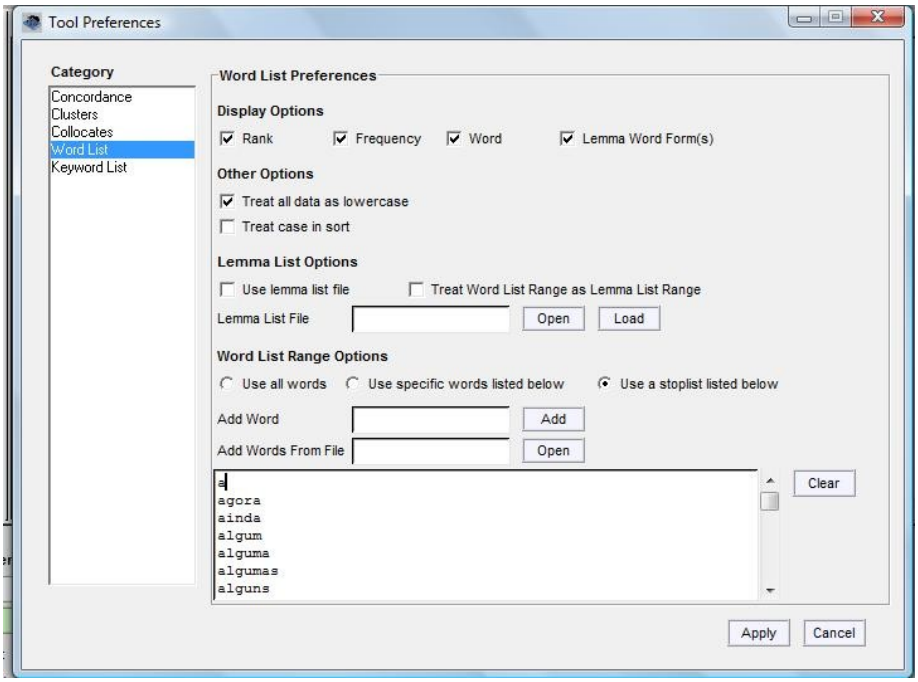

Figura . **Tela *Tool Preferences* do AntConc.**

Concluída essa etapa inicial, basta clicar na guia desejada e começar o trabalho. Começamos pela guia *Word List*. Para gerar a lista de palavras, basta clicar no botão *Start*. A partir daí, o processo todo é muito simples. Para visualizar linhas de concordância, por exemplo, basta selecionar um termo e clicar sobre ele – o programa vai pular diretamente para a aba de concordância. Nela, é possível ver, além do termo em contexto, o arquivo de origem à direita da tela. A figura a seguir mostra as linhas de concordância apresentadas para o termo *depressão*:

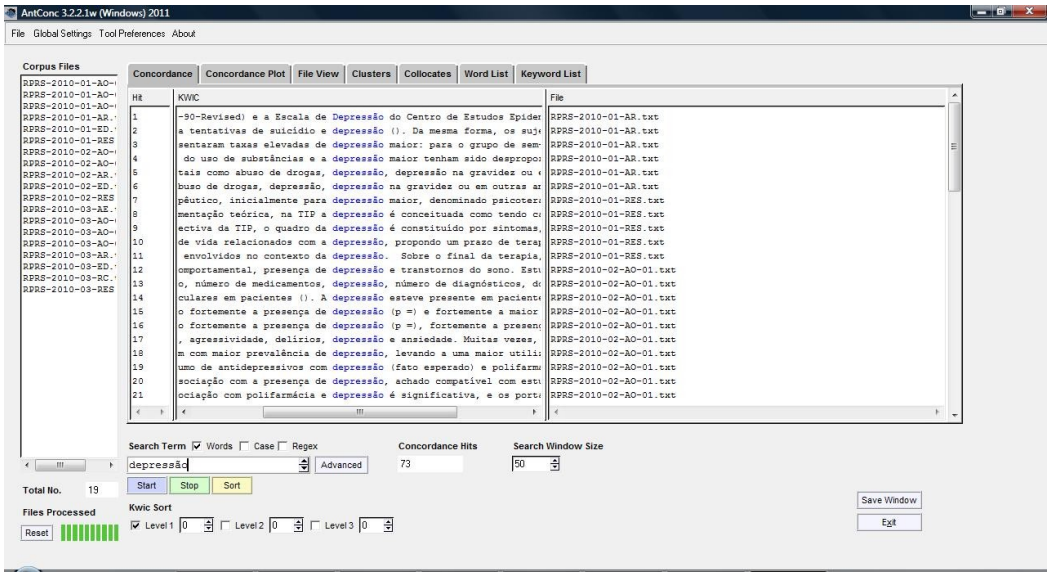

Figura . **Linhas de concordância para o termo *depressão*.**

Caso o pesquisador queira ampliar o contexto, basta clicar no termo (destacado em azul), e será direcionado para a aba *File View*, em que é possível ver onde o termo ocorre dentro do texto. Os termos aparecem destacados (em azul) e, no topo, é possível ver quantas ocorrências há naquele texto.

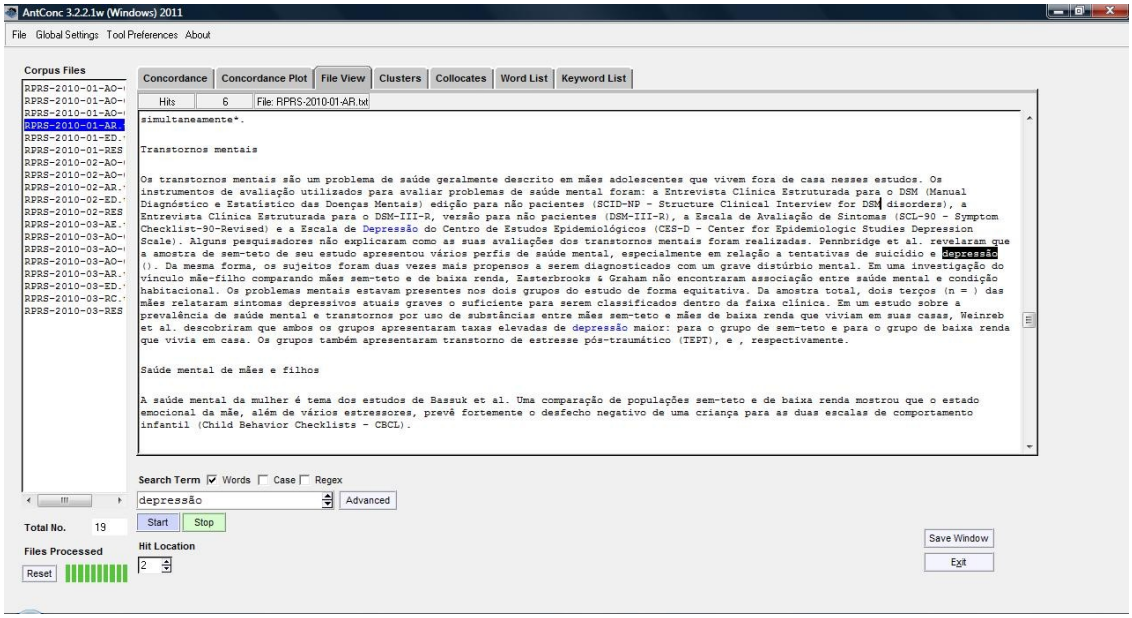

Figura . **Resultado apresentado na aba *File View*.**

Além destas funcionalidades, o AntConc também dispõe de gerador de N-gramas (aba *Clusters*) e de colocados (aba *Collocates*), recursos bastante úteis na pesquisa linguística. Os resultados obtidos nas abas de lista de palavras, concordanciador, *clusters* e colocados podem ser exportados para arquivos em formato .txt.

3 DISCUSSÃO

Após a testagem da ferramenta, foi possível verificar que o AntConc apresenta uma série de vantagens:

- o tamanho pequeno do arquivo, que permite um *download* rápido (mesmo com conexões à internet de baixa velocidade) e não ocupa muito espaço em disco;
- o fato de não haver necessidade de instalação e licenciamento;
- ser um *freeware* com versões para diferentes sistemas operacionais;
- a facilidade de uso, que permite acessar várias funcionalidades em uma mesma interface, com apenas um clique.

Apesar das vantagens apresentadas pelo AntConc, antes de escolher o analisador que será usado na tese pretendemos também testar o ambiente e-Termos (http://www.etermos.cnptia.embrapa.br/), a fim de verificar o que é mais adequado aos propósitos do trabalho. Além disso, parece-nos que seja necessário testar melhor algumas das funcionalidades das ferramentas utilizadas neste trabalho.

Com relação à análise do *corpus*, a ferramenta testada gerou uma lista de palavra muito grande (mais de 5.000 *types* ou palavras diferentes), trazendo uma grande quantidade de material indesejado, o que sugere a necessidade de:

- rever a etapa de limpeza dos arquivos, pois talvez seja necessária e exclusão de outros elementos (muitos nomes de autores citados no meio do texto não foram excluídos, por exemplo);
- revisar a lista de *stopwords* e incluir novas palavras, com base nas listas geradas;
- selecionar um *corpus* de referência para a geração de lista de palavras-chave.

Entretanto, acreditamos que mesmo com todo esse trabalho, a etapa de seleção manual dos candidatos a termo não será eliminada, pois a máquina se baseia em critérios de frequência, o que nem sempre nos traz aquilo que buscamos.

Em função do volume de sujeira gerado nas listas de palavras, não foi realizada uma análise mais detalhada do *corpus*, pois isso demandaria algum tempo na limpeza e seleção

manual dos candidatos a termo – etapas que serão realizadas no decorrer da pesquisa mas que, para o presente estudo-piloto, não nos pareceram pertinentes tendo em vista que o objetivo era a análise da ferramenta de análise de *corpora*.

CONSIDERAÇÕES FINAIS

As ferramentas de análise de análise de *corpora* têm um papel bastante importante na pesquisa linguística atualmente. No entanto, tais ferramentas têm por base um critério tão somente quantitativo, o que, para uma pesquisa de reconhecimento terminológico, pode não ser interessante. Isso acontece porque alguns termos podem ter um baixo número de ocorrências em um *corpus* sem que, por isso, sejam menos importantes enquanto representações de conceitos-chave de uma determinada área de conhecimento.

Dessa forma, por mais que a ferramenta possa ser programada para eliminar elementos que não interessem ao pesquisador – uso de listas de *stopwords*, uso de *corpora* de referência – o trabalho de seleção manual de termos ainda está longe de ser eliminado. O programa pode fazer o trabalho da geração de listas de candidatos a termo, o que realmente ajuda muito na pesquisa. No entanto, a seleção daquilo que realmente poderá constar em um glossário continuará sendo feita pelo pesquisador/terminólogo em conjunto com o especialista da área.

REFERÊNCIAS BIBLIOGRÁFICAS

ALUÍSIO, S.M.; ALMEIDA, G.M.B. O que é e como se constrói um *corpus*? Lições aprendidas na compilação de vários *corpora* para a pesquisa lingüística. In: **Calidoscópio**. V4, n. 3, setembro/dezembro, 2006. Disponível em <http://www.unisinos.br/publicacoes_cientificas/images/stories/pdfs_calidoscopio/vol4n3/art04_aluisio.pdf>. Acesso em 20 mai 2011.

ANTHONY, L. **Lawrence Anthony Website (AntConc).** Disponível em <http://www.antlab.sci.waseda.ac.jp/index.html>. Acesso em 23 mai 2011.
_____. **Arquivo de ajuda do AntConc**.

BERBER SARDINHA, T. Lingüística de Corpus: histórico e problemática. In: **DELTA** [online]. Vol. 16, n. 2, 2000, p. 323-367. Disponível em <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-44502000000200005&lng=en&nrm=iso>. Acesso em 20 mai 2011.

BIBER, D. Representativeness in corpus design. In: **Literary and Linguistic Computing,** 8, 1993, 243-257.

LOH, S. (blogger). **Text mining por Stanley Loh.** Blog. Disponível em <http://miningtext.blogspot.com/>. Acesso em 06 jun 2011.

LEXICAL ANALYSES SOFTWARE. **WordSmith Tools Website**. Disponível em: <http://www.lexically.net/wordsmith/>. Acesso em 26 mai 2011.

_____. **Arquivo de ajuda do WordSmith Tools**.

SINCLAIR, J. Corpus and Text - Basic Principles. In: WYNNE, M. (ed.) **Developing Linguistic Corpora: a Guide to Good Practice**. Oxford: Oxbow Books, 2005, p. 1-16. Disponível em <http://ota.ahds.ac.uk/documents/creating/dlc/chapter1.htm>. Acesso em 20 mai 2011.