the ESPecialist, vol. 27, nº 1 (1-19) 2006 ISSN 0102-7077

O BANCO DE PALAVRAS-CHAVE COMO INSTRUMENTO DE IDENTIFICAÇÃO DE PALAVRAS-CHAVE EXCLUSIVAS NO PROGRAMA WORDSMITH TOOLS KEYWORD*

The Keyword Bank as a Tool for Finding Exclusive Keywords in WordSmith Tools

Tony BERBER SARDINHA (PUC-SP)

Abstract

KeyWords is a very useful program for computer text analysis found in WordSmith Tools. A problem with KeyWords, though, is the large number of keywords returned by the program, which can be at least 500. This paper proposes a procedure for making reductions in lists of keywords based on the concept of exclusive keywords. These are words that are key in the study corpus only, in comparison to lots of others. This procedure draws on the existence of a keyword bank, which is a collection of keywords from several corpora. When contrasted to a study corpus, the keyword bank brings up keywords that are found in the study corpus only, leaving out those that are key in other corpora. This enables the researcher to focus on words that are most typical of his/her own corpus. The analysis reported here, carried out with a large multi-register keyword bank, suggests that the keyword bank achieved its goal, by allowing for a 77% reduction in the total keywords, and by selecting keywords that are most representative of the study corpus in question.

Keywords: *WordSmith Tools; KeyWords; keywords; corpora.*

Resumo

O programa KeyWords, parte da suíte WordSmith Tools, é uma ferramenta das mais úteis na análise textual por computador. Um problema para o analista de palavras-chave é, entretanto, a quantidade de palavras que o programa retorna, inicialmente 500, mas normalmente

* Uma versão deste trabalho apareceu como DIRECT Paper 39. O autor agradece ao CNPq o apoio financeiro (350455/2003), aos pareceristas as sugestões, e à Leila Barbara e aos membros do projeto DIRECT os comentários durante uma apresentação oral deste trabalho.

1500. O presente trabalho propõe um procedimento para feitura de recortes em listas de palavras-chave baseado no conceito de palavras-chave exclusivas. Essas palavras são aquelas que são chave apenas no corpus de estudo em questão, em comparação com outros, sendo, portanto, em menor número do que o total de palavras-chave retornado pelo programa. O procedimento baseia-se na aplicação de um banco de palavras-chave, o qual, quando comparado às palavras-chave do corpus de estudo, deixa entrever quais palavras-chave são exclusivas. A extração de palavras-chave exclusivas por meio do banco de palavras-chave parece ser quantitativa e qualitativamente eficaz. O banco permitiu uma redução de 77% das palavras-chave obtidas.

Palavras-chave: *WordSmith Tools; KeyWords; palavras-chave; corpora.*

1. Introdução

Uma das ferramentas mais úteis da suíte WordSmith Tools é o programa KeyWords. Esse programa permite que o usuário compare uma lista de palavras de seu corpus de estudo (ou mais listas de vários corpora) com um corpus de referência. Desse modo, o usuário obtém uma seleção dos itens lexicais de seu corpus de estudo que são estatisticamente mais distintivos. Palavras-chave são, portanto, aquelas cujas freqüências são estatisticamente diferentes no corpus de estudo em relação ao corpus de referência.

Uma análise por KeyWords exige dois elementos básicos:

(a) um corpus de estudo, representado em uma lista de freqüência de palavras. O corpus de estudo é aquele que se pretende descrever. A ferramenta KeyWords aceita a análise simultânea de mais de um corpus de estudo.

(b) um corpus de referência, também formatado como uma lista de freqüência de palavras. Também é conhecido como 'corpus de controle', e funciona como termo de comparação para a análise. A sua função é a de fornecer uma norma com

> a qual se fará a comparação das freqüências do corpus de estudo. A comparação é feita através de uma prova estatística selecionada pelo usuário (qui-quadrado ou log-likelihood). As palavras cujas freqüências no corpus de estudo forem significativamente maiores segundo o resultado da prova estatística são consideradas chave, e passam a compor uma listagem específica de palavras-chave.

O resultado apresentado pelo programa KeyWords varia de acordo com o tamanho e a composição dos corpora de estudo e de referência, e com o nível de significância escolhido pelo usuário. Em outras palavras, as listas de palavras resultantes serão maiores ou menores, e de composição diferente, de acordo com os ajustes e seleções de corpora efetuados pelo usuário.

A variação no tamanho das listas de palavras-chave é em geral problemática não tanto porque a quantidade de palavras-chave em um corpus de estudo é relativa a esses ajustes, mas principalmente porque a quantidade de palavras-chave retornada ultrapassa a capacidade de análise detalhada de que um analista humano dispõe. Embora o computador possa selecionar os itens-chave de um corpus de milhões de palavras em segundos, um analista pode levar dias ou meses (até anos) para dar conta da descrição detalhada dessas palavras. As limitações da capacidade de análise e interpretação humanas devem, portanto, ser levadas em conta numa análise de palavras-chave.

O fato de a quantidade de palavras-chave ser em geral maior do que a capacidade de interpretação do analista não se caracteriza como um defeito do programa KeyWords. O programa é uma ferramenta, isto é, um instrumento para se dar cabo de uma tarefa; ele não foi desenhado para fazer uma análise inteira, até porque seria impossível, visto que a análise exige interpretação, e os computadores são incapazes de compreender a linguagem e, portanto, de interpretá-la.

A seleção de uma parte das palavras-chave para análise mais detalhada é um procedimento inevitável em boa parte das análises baseadas em palavras-chave, principalmente aquelas que se destinam à descrição de padrões lexicais típicos de um gênero, registro ou variedade lingüística (Barbara e Scott, 1999; Batista, 1998; Bonamin, 1999; Collins

e Scott, 1997; Conde, 2002; Dutra, 2002; Freitas, 1997; Fuzetti, 2003; Ide, 1989; Lima-Lopes, 1999; Lopes, 2000; Pressley, 1976; Ramos, 1997; Santos, 1996; Silva, 1999). Nesses estudos, o que se busca é a localização de um conjunto de itens lexicais que sejam caracterizadores do gênero em questão. O problema é que o conceito de léxico caracterizador é subjetivo, pois depende da interpretação humana na interação entre texto e leitor, ou entre falante e ouvinte. Desse modo, a localização do léxico caracterizador também é subjetiva. Há, portanto, um contraste entre a objetividade da localização das palavras-chave, a qual obedece a princípios estatísticos regulares, e a subjetividade do recorte do léxico caracterizador a partir dessas palavras-chave, o qual é subjetivo e variável.

Uma aproximação mais objetiva do conceito de léxico caracterizador é o de léxico-chave exclusivo, o qual é composto de palavras-chave não encontradas em outros corpora de estudo. Esse léxico pode ser entendido como os *hapax legomena-chave* de um corpus. O léxico exclusivo pode, ao contrário do caracterizador, ser encontrado objetivamente (e, por conseguinte, por meios computadorizados) através da comparação entre listas de palavras-chave de corpora de estudos diferentes.

Para se localizar o léxico-chave exclusivo, deve-se fazer a comparação do léxico-chave (positivo) de pelo menos dois corpora. Assim, os itens-chave que só ocorrerem no corpus 'x' são exclusivos daquele corpus; e os itens-chave só encontrados no corpus 'y' são exclusivos do corpus 'y', e assim por diante. O contraponto do léxico-chave exclusivo é o léxico-chave *compartilhado*, ou seja, aquelas palavras-chave que aparecem em mais de uma lista. Esse léxico recebe o nome, no programa KeyWords, de *chave-chave* (pelo critério mínimo de freqüência igual a 2). Desse modo, o léxico-chave exclusivo é o oposto de léxico-chave-chave.

Em resumo, os conceitos principais discutidos até este ponto são:

(a) Léxico caracterizador: itens lexicais mais característicos, relevantes, representativos, singulares, típicos, ou definidores de um corpus de estudo.

(b) Léxico-chave: itens lexicais estatística e comparativamente mais típicos de um corpus de estudo.

(c) Léxico-chave exclusivo: itens lexicais-chave não compartilhados por outros corpora de estudo, em uma dada comparação.

Para identificação do léxico-chave exclusivo, é necessário dispor-se de um conjunto de listas de palavras-chave de vários corpora diferentes. Os corpora devem ser em grande número, pois a exclusividade lexical não é um traço inerente a nenhum item. Pelo contrário, a exclusividade é relativa aos corpora com os quais se faz a comparação. Tecnicamente, quanto maior o número de corpora comparados, menor a chance de um item ser exclusivo, visto que há mais chance desse item se repetir nos outros corpora. Por isso, é essencial que se tome uma posição conservadora a respeito da exclusividade e se faça uso da maior quantidade possível de corpora na comparação das listas de palavras-chave.

Um instrumento necessário para a identificação das palavras-chave exclusivas é, portanto, um conjunto de corpora de estudo. Este trabalho pretende apresentar e aplicar o *banco de palavras-chave*, uma coletânea de listas de palavras-chave extraídas de um conjunto de corpora de estudos com as características desejadas citadas acima. Pretende-se mostrar que o banco de palavras-chave é um instrumento valioso na identificação objetiva de um subconjunto de palavras-chave, ao permitir a localização de palavras-chave exclusivas. Desse modo, os procedimentos descritos aqui podem ser de valia para os analistas de palavras-chave que se vêem às voltas com a difícil missão de fazer um recorte no léxico-chave do seu corpus de estudo.

2. Composição do banco de palavras-chave

O banco de palavras-chave é composto por 40 corpora diferentes, totalizando 2,7 milhões de palavras, distribuídas da seguinte forma:

	Corpus	Itens	Formas
1	Apresentação de caso no tribunal	9927	1640
2	Artigos de enciclopédia	226107	19062
3	Artigos de pesquisa acadêmicos	621512	27368
4	Artigos de revista especializada	232046	20316
5	Aulas radiofônicas	11327	1764
6	Biografias	90717	11902
7	Cartas de pedido de emprego	12089	2412
8	Cartas de proposta	22292	3586
9	Cartas-convite	714	326
10	Cartas-resposta	7622	1859
11	Circulares	2613	947
12	Cobertura jornalística ao vivo	13978	2575
13	Conversa face a face	361096	15473
14	Conversa telefônica	62974	4843
15	Debates parlamentares	14918	2324
16	Discurso falado político	5020	1257
17	Documentário jornalístico	4789	1015
18	Editais de licitação	16178	1903
19	Editoriais de jornal	54626	8582
20	Ficção	235095	17808
21	Folhetos de escola	17154	3365
22	Folhetos de hotéis	64984	6157
23	Folhetos de negócio	15179	3494
24	Folhetos governamentais	4251	1227
25	Folhetos turísticos	115103	12185
26	Interrogatório jurídico	4991	723
27	Livros acadêmicos	64255	8062
28	Manuais técnicos	4377	907
29	Narração esportiva radiofônica	25791	2918

	Corpus	Itens	Formas
30	Notícias	89674	11781
31	Palestras acadêmicas	29598	2639
32	Palestras em jantar	5141	1047
33	Palestras universitárias	5012	1330
34	Redações escolares	25062	3149
35	Relatórios anuais de negócio	168972	8570
36	Relatórios do Supremo Tribunal	4321	1038
37	Relatórios governamentais	42083	4946
38	Resenhas jornalísticas	35741	7746
39	Reuniões de negócio	12648	1762
40	Sermão religioso	5071	1288
	Total	2745048	71163

Tabela 1: Composição do banco de palavras-chave

Os corpora foram retirados do banco de textos do projeto DIRECT, de coleções próprias do autor e de outros corpora (London Lund, LOB e Brown).

3. Procedimentos para se usar o banco

O banco de palavras-chave encontra-se armazenado, juntamente com o banco de dados do projeto DIRECT (Para um Desempenho Mais Eficiente na Comunicação Internacional), no LAEL (Programa de Estudos Pós-Graduados em Lingüística Aplicada e Estudos da Linguagem) da Pontifícia Universidade Católica de São Paulo. No momento, o banco é de uso exclusivo dos membros do projeto. Embora o banco seja de palavras-chave, ele é mantido no formato WordList. A razão disso é que para se encontrar as palavras-chave exclusivas usa-se o programa WordList, e não o KeyWords. No WordList, aciona-se o procedimento 'Consistency Detailed', no menu 'Comparison', do

WordList. O banco encontra-se disponível em dois formatos: em uma só lista de palavras-chave (arquivo 'bancopc.lst') e em listas de palavras individuais (40 arquivos .lst diferentes, um para cada corpus). O motivo da necessidade desses dois formatos é explicado a seguir.

Para encontrar as palavras-chave exclusivas por intermédio do banco de palavras-chave, o usuário deve:

(1) Extrair as palavras-chave de seu corpus de estudo; recomenda-se empregar o mesmo corpus de referência utilizado na produção do banco de palavras-chave (ou dos outros corpora de estudo, caso não se use o banco). No banco, empregou-se o corpus do jornal inglês *The Guardian*, com mais de 95 milhões de palavras, referentes às edições completas diárias de 1991 a 1994. As vantagens desse corpus são que ele se tornou uma referência padrão no estudo de palavras-chave, e é acessível ao público em geral através da Internet (no site www.liv.ac.uk/~ms2928). Também é recomendável usar os mesmos ajustes do programa KeyWords daqueles usados no banco (vide Tabela 2 abaixo);

(2) Salvar a lista de palavras-chave como WordList (clicando no botão 'W' na barra de ferramentas do programa KeyWords);

(3) Certificar-se de que um corpus com as características genéricas do seu corpus de estudo *não faz parte* do banco de palavras-chave.
• Se não fizer, o analista pode seguir para o passo 4 a seguir;
• Se fizer, o analista deve eliminar do banco de palavras-chave o corpus similar ao seu. Para tanto, basta clicar em 'Merge Word Lists' no menu 'File' do programa WordList, selecionar as listas individuais correspondentes (entre os 40 arquivos .lst de cada corpus individual), e salvar a lista resultante.

(4) Fazer a comparação do tipo 'Consistency (detailed)', clicando no menu 'Comparison' do WordList, selecionando os arquivos correspondentes à lista de seu corpus de estudo

e à lista do banco de palavras-chave (completo ou sem o corpus similar ao do estudo; vide passo 3 acima). A tela de resultados da operação 'Consistency (detailed)' mostra, da esquerda para a direita, as seguintes colunas:
• os itens lexicais (na coluna 'Word');
• o total de arquivos em que o item aparece (na coluna 'Total');
• a freqüência do item no banco de palavras-chave (na coluna 'Bancopc');
• a freqüência do item no corpus de estudo (numa coluna com o nome do arquivo correspondente; essa coluna pode aparecer antes da anterior devido ao arranjo por ordem alfabética pré-definido no programa).

(5) Identificar as palavras-chave exclusivas. Estas serão as que tiverem freqüência 1 na coluna 'Total'. Devido ao fato de o programa WordList não permitir a classificação por mais de uma coluna, é necessário apagar-se incrementalmente as palavras que (a) aparecem em ambos os arquivos (isto é, no corpus de estudo e no banco de palavras-chave), e (b) não aparecem no corpus de estudo. Desse modo, as palavras restantes serão aquelas que aparecem somente no corpus de estudo, que são exatamente as palavras-chave exclusivas. Para tanto, deve-se seguir os passos abaixo:

• Clicar no botão 're-sort' na barra de tarefas;
• Selecionar 'Total', e clicar OK;
• Selecionar na lista todos os itens que possuírem freqüência 2 em 'Total';
• Apagá-los da lista pressionando a tecla 'delete';
• Eliminá-los clicando no botão 'Zap' na barra de tarefas;
• Clicar novamente no botão 're-sort';
• Selecionar o arquivo correspondente ao corpus de estudo, e clicar em OK;
• Selecionar na lista todos os itens que possuírem freqüência 0 na coluna relativa ao corpus de estudo;
• Apagá-los da lista pressionando a tecla 'delete';
• Eliminá-los clicando no botão 'Zap' na barra de tarefas.

(6) Salvar a lista restante.

A lista resultante é a que contém as palavras-chave exclusivas do corpus de estudo.

4. Exemplo

Para ilustrar os procedimentos acima, serão apresentados a seguir os resultados da aplicação desses procedimentos na identificação do léxico exclusivo de um corpus de conversas face a face. Como esse corpus já possui um similar no banco de palavras-chave, a decisão mais cautelosa é a de eliminar o corpus correspondente do banco de palavras-chave. Para tanto, deve-se seguir as recomendações do item 3 dos procedimentos acima. Uma vez recompilado o banco sem o corpus de conversas face a face, deve-se seguir os passos seguintes até se obter a lista final de palavras-chave exclusivas, conforme explicado no restante dos procedimentos acima.

O corpus de referência empregado foi um formado pelas edições completas do jornal inglês *The Guardian* de 1991 a 1994, somando cerca de 95 milhões de palavras. O programa KeyWords foi ajustado segundo os valores abaixo:

Ajuste	Valor
Procedimento	Loglikelyhood
Max p. value	0.05
Max wanted	16000*
Min frequency	2

* máximo permitido

Tabela 2: Ajustes do programa KeyWords utilizados na pesquisa

Os resultados numéricos obtidos são os seguintes:

Palavras-chave	2507
Palavras-chave exclusivas	1137
Redução	55%

Segundo a tabela acima, a lista original de palavras-chave foi reduzida em mais da metade. Mas para se aferir o impacto da seleção de palavras-chave exclusivas no vocabulário-chave, deve-se comparar os itens lexicais incluídos em cada lista.

A tabela a seguir mostra as 20 primeiras palavras-chave ordenadas por freqüência no corpus de estudo. A ordenação por chavicidade, embora mais adequada em se tratando de palavras-chave, não está disponível no procedimento 'Consistency' do WordList, pois os valores de chavicidade são perdidos quando da transformação do formato de lista de palavras-chave (.kws) para lista de palavras (.lst).

Word	**Total**	**Bancopc**	**Convers**
I	18	331052	11378
AND	33	2115496	9844
YOU	22	175617	7769
M	25	9451	7211
THAT	22	877191	6334
IT	16	701199	6212
YES	9	8542	3811
WAS	8	701712	3548
BUT	8	473281	2986
WELL	16	75606	2852
THIS	26	340598	2809
KNOW	14	41284	2526
HAVE	11	445852	2480
THEY	7	339982	2398
WE	21	203117	2377
IT'S	16	62587	2328
BE	15	534556	2191
THINK	15	41435	2132
NO	8	177771	2058
SO	20	155638	1953

Tabela 3: Vinte primeiras palavras-chave ordenadas por freqüência no corpus de estudo

Nota-se que as palavras-chave listadas na tabela acima, embora sejam intuitivamente caracterizadoras da conversa face a face em inglês, são de caráter geral. 'I', 'and', e 'you' são palavras gerais da língua inglesa, que podem ocorrer em outros gêneros. De fato, essas palavras ocorreram, respectivamente, em 18, 33, e 22 gêneros do banco de palavras-chave, conforme indica a coluna 'Total'.

A tabela 4 a seguir mostra as palavras-chave exclusivas, também ordenadas por ordem de freqüência no corpus de estudo.

Word	**Total**	**Bancopc**	**Convers**
PUTTING	1	7691	55
SURELY	1	6779	51
BLOODY	1	3117	50
THINKS	1	4346	42
EXPENSIVE	1	6272	41
LECTURER	1	1551	39
AFTERWARDS	1	4157	33
DREADFUL	1	1332	32
BLOKE	1	624	31
NURSE	1	2198	30
NASTY	1	1933	26
RUMBLING	1	185	25
REHEARSAL	1	744	23
GUINNESS	1	1941	22
MILLY	1	20	21
DEPARTMENTAL	1	393	20
PIGGOTT	1	444	20
BARRY	1	3066	19
CHAPS	1	531	19
DENZIL	1	57	19

Tabela 4: Vinte primeiras palavras-chave *exclusivas* ordenadas por freqüência no corpus de estudo

A lista acima, ao contrário da anterior, foi expurgada de itens gerais como 'I', 'and', e 'you'. As três palavras mais freqüentes são

agora ‘putting’, ‘surely’ e ‘bloody’, as quais são mais raras na língua em geral e intuitivamente mais restritivas em relação aos gêneros em que podem ocorrer. De fato, o banco de palavras-chave indica que essas palavras só ocorreram como chave no corpus de conversas face a face, o que é uma forte indicação de sua natureza específica em termos de gêneros discursivos.

As palavras constantes na relação de palavras-chave exclusivas parecem indicar os tópicos da interação. Na lista de palavras-chave, esses itens tornavam-se menos proeminentes devido à presença dos demais itens lexicais. Não se possuem elementos ainda para se julgar se efetivamente o banco de palavras-chave possui uma tendência de promover uma seleção dos elementos ideacionais relativos ao tópico (campo) em detrimento dos elementos interpessoais e textuais, os quais também contribuem para a caracterização genérica. Para se afirmar se essa tendência existe, seria necessário examinar a composição das listas de palavras-chave dos demais corpora, em sua totalidade. Lembre-se de que as palavras listadas na tabela acima correspondem a 1,75% (20 das 1137) das palavras-chave exclusivas do corpus de conversas. Os demais 98,25% podem incluir itens relativos à organização interpessoal e textual da conversação.

Os resultados dessa breve comparação indicam que (a) o banco de palavras-chave permite uma redução efetiva da quantidade de palavras-chave existentes no corpus de estudo, e (b) as palavras exclusivas remanescentes são aparentemente mais caracterizadoras do gênero representado pelo corpus de estudo. É possível concluir-se, portanto, que o conceito de palavras-chave exclusivas operacionalizado através dos procedimentos descritos aqui, e posto em prática através do banco de palavras-chave, não é um meio ineficaz de seleção de palavras-chave.

5. Eficácia do banco como instrumento redutor

Uma das dúvidas que resta é a respeito da significância da redução da quantidade de palavras-chave. A ilustração acima apontou que a redução da lista de palavras-chave foi da ordem de 55%, mas esse valor se refere a somente um corpus de estudo. Para se saber se taxas

tão altas de redução se mantêm em outros corpora, é preciso efetuar-se um estudo de larga escala, e testar os resultados estatisticamente, para se ter certeza de que a redução observada é maior do que o esperado por acaso. Essa verificação estatística é pertinente, pois se uma redução acentuada como a ocorrida com o corpus de conversas face a face não for sentida com outros corpora, para todos os efeitos os valores pré- e pós-redução serão similares, o que indicaria que o procedimento é supérfluo.

A testagem em larga escala foi empreendida do seguinte modo: os corpora do banco de palavras-chave foram extraídos, um de cada vez, do banco de palavras-chave, e depois comparados, também um a um, com o corpus de estudo restante. Mais especificamente, os passos seguidos foram os seguintes:

(a) Selecionou-se um dos corpora do banco de palavras-chave;

(b) Esse corpus tornou-se o corpus de estudo provisório;

(c) Agruparam-se os corpora restantes;

(d) Esses corpora tornaram-se o banco de palavras-chave provisório;

(e) Fez-se a extração das palavras-chave exclusivas seguindo-se os procedimentos descritos acima.

Para cada corpus, anotaram-se os totais de palavras-chave e de palavras-chave exclusivas. Os resultados aparecem na tabela a seguir.

Corpus	**Palavras-chave**	**Palavras-chave exclusivas**	**Redução %**
Apresentação de caso no tribunal	421	62	85.3
Artigos de enciclopédia	5479	2296	58.1
Artigos de pesquisa acadêmicos	8668	4934	43.1
Artigos de revista especializada	4382	1414	67.7
Aulas radiofônicas	543	80	85.3
Biografias	2150	668	68.9
Cartas de pedido de emprego	628	95	84.9

Corpus	Palavras-chave	Palavras-chave exclusivas	Redução %
Cartas de proposta	852	138	83.8
Cartas-convite	60	0	100.0
Cartas-resposta	582	99	83.0
Circulares	190	21	88.9
Cobertura jornalística ao vivo	714	184	74.2
Conversa face a face	2507	1136	54.7
Conversa telefônica	1046	262	75.0
Debates parlamentares	515	83	83.9
Discurso falado político	268	42	84.3
Documentário jornalístico	292	38	87.0
Editais de licitação	714	88	87.7
Editoriais de jornal	1385	331	76.1
Ficção	4428	2291	48.3
Folhetos de escola	844	152	82.0
Folhetos de hotéis	2202	679	69.2
Folhetos de negócio	1101	247	77.6
Folhetos governamentais	406	35	91.4
Folhetos turísticos	3918	1562	60.1
Interrogatório jurídico	235	13	94.5
Livros acadêmicos	2009	365	81.8
Manuais técnicos	353	53	85.0
Narração esportiva radiofônica	819	307	62.5
Notícias	2242	798	64.4
Palestras acadêmicas	1145	338	70.5
Palestras em jantar	251	25	90.0
Palestras universitárias	282	32	88.7
Redações escolares	725	197	72.8
Relatórios anuais de negócio	2854	1076	62.3
Relatórios do Supremo Tribunal	273	41	85.0
Relatórios governamentais	1410	152	89.2
Resenhas jornalísticas	1323	443	66.5
Reuniões de negócio	384	44	88.5
Sermão religioso	284	28	90.1

Tabela 5: Percentuais de redução de quantidade de palavras-chave por corpus

Em todos os corpora, nota-se um grau de redução similar ou maior do que aquele apresentado no corpus de conversas face a face (da ordem de 55%). A média de redução é superior ao alcançado com o corpus de conversas, situando-se em torno de 77%. Isso significa que, em média, as listas passaram de 1472 palavras-chave para 521 palavras-chave exclusivas.

Para se saber se esses números são de fato significativos, recorreu-se ao Teste T de Student, através do qual se fez a comparação das médias de palavras-chave originais e exclusivas. Os resultados aparecem abaixo.

Palavras	N*	Média*	Desvio padrão	T	GL	P
Chave	39	1508.307692	1769.141781	3.0399	76.0	0.0032
Exclusivas	39	534.589744	933.630089			

Tabela 7: Teste-T de significância para os valores de redução de palavras-chave

* Os dados referentes ao corpus de cartas-convite foram eliminados, daí a diferença no total de corpora e nas médias em relação ao apresentado nas tabelas anteriores.

Os resultados do teste T indicam que as diferenças entre os totais de palavras-chave e exclusivas são estatisticamente significantes (p=0.0032). Pode-se concluir, portanto, que o banco de palavras-chave contribui significativamente para a redução do total de palavras-chave.

6. Considerações finais

O programa KeyWords é uma ferramenta das mais úteis na análise textual por computador. Através dela é possível se ter acesso, em instantes, ao conjunto de vocabulário mais típico ou definidor do corpus de estudo, identificados por meio da comparação das freqüências do vocabulário do corpus de estudo com o de um corpus de referência. A análise computadorizada não prescinde, entretanto, da análise humana, detalhada, das palavras-chave. Nesse ponto surge, em geral, um problema para o analista: a quantidade de palavras-chave de um corpus de estudo

é em geral grande demais, girando em torno de cerca de 1500, segundo os números obtidos neste estudo, mas podendo chegar a mais de 8 mil! Por isso, é inevitável que o analista lance mão de um recurso para efetuar um recorte no vocabulário-chave, para que a análise detalhada, manual e interpretativa, seja exeqüível.

O presente trabalho propôs um procedimento para feitura de recortes em listas de palavras-chave baseado no conceito de palavras-chave exclusivas. Essas palavras são aquelas que são chave apenas no corpus de estudo em questão, em comparação com outros. O procedimento baseia-se na aplicação de um banco de palavras-chave, o qual, quando comparado às palavras-chave do corpus de estudo, deixa entrever quais palavras-chave são exclusivas.

A extração de palavras-chave exclusivas por meio do banco de palavras-chave parece ser quantitativa e qualitativamente eficaz. O banco permitiu uma redução de 77% das palavras-chave obtidas. A redução é praticável na vasta maioria dos corpora de estudo; apenas um dos 40 corpora empregados não apresentou palavras-chave exclusivas. As palavras retornadas em si também aparentam ser mais caracterizadoras do que as palavras-chave originais. Os itens lexicais exclusivos elencados parecem ser menos gerais e mais exclusivos do gênero em questão. Em outras palavras, o procedimento de escolha de palavras-chave exclusivas parece filtrar as palavras-chave mais comuns.

Recebido em: 01/2004; Aceito em: 06/2004.

Referências Bibliográficas

Barbara, L. & Scott, M. 1999 Homing on a genre: invitations for bids. IN: F. Bargiela-Chiapini & C. Nickerson (orgs.) *Writing business: genres, media and discourse*. Longman.

Batista, M.E. 1998 *E-mails na troca de informação numa multinacional: o gênero e as escolhas léxico-gramaticais*. Dissertação de Mestrado. LAEL, PUC/SP.

Bonamin, M.C. 1999 *Análise organizacional e léxico-gramatical de duas seções de revistas de informática, em inglês*. Dissertação de Mestrado. LAEL, PUC/SP. (http://lael.pucsp.br/lael)

COLLINS, H. & SCOTT, M. 1997 Lexical landscaping in business meetings. IN: F. BARGIELA-CHIAPPINI & S. HARRIS (orgs.) *The languages of business - an international perspective*. Edinburgh University Press.

CONDE, H. 2002 Escolhas léxico-gramaticais em composições de alunos avançados de inglês originários de instituições de ensino bilíngües e monolíngües - um estudo multidimensional baseado em corpus. Dissertação de mestrado inédita. LAEL, PUC/SP. Disponível online em http://lael.pucsp.br/lael-inf/def_teses.html

DUTRA, P.B. 2002 Explorando a Lingüística de Corpus e letras de música na produção de atividades pedagógicas. Dissertação de mestrado inédita. LAEL, PUC/SP. Disponível online em http://lael.pucsp.br/lael-inf/def_teses.html

FREITAS, A.C. de 1997 *América mágica, Grã-Bretanha real e Brasil tropical: um estudo lexical de panfletos de hotéis*. Tese de doutorado. LAEL, PUC/SP. (http://lael.pucsp.br/lael)

FUZETTI, H. 2003. Padrões léxico-gramaticais na linguagem de crianças em uma escola americana no Brasil. Dissertação de mestrado inédita. LAEL, PUC/SP. Disponível online em http://lael.pucsp.br/lael-inf/def_teses.html

IDE, N. 1989 A statistical measure of theme and structure. *Computers and the Humanities*, **23**: 277-283.

LIMA-LOPES, R.E. 1999 Padrões colocacionais dos participantes em cartas de negócios em língua inglesa. Trabalho final de módulo de Lingüística de Corpus. LAEL, PUC/SP.

LOPES, M.C. 2000 Homepages institucionais em português e suas versões em inglês: um estudo baseado em corpus sobre aspectos lexicais e discursivos. Dissertação de mestrado inédita. LAEL, PUC/SP. Disponível online em http://lael.pucsp.br/lael-inf/def_teses.html

PRESSLEY, G.M. 1976 Mental imagery helps eight-year-olds remember what they read. *Journal of Educational Psychology*, **68**: 355-359.

RAMOS, R.G. 1997 *Projeção de imagem através de escolhas lingüísticas: um estudo no contexto empresarial*. Tese de Doutorado. LAEL, PUC/SP.

SANTOS, V.B.M.P. dos 1996 *Padrões interpessoais no gênero de cartas de negociação*. Dissertação de mestrado. LAEL, PUC/SP. (http://lael.pucsp.br/lael)

SILVA, M.S.F. da 1999 *Análise lexical de folhetos de propagandas de escolas de línguas e as representações de ensino*. Dissertação de mestrado. LAEL, PUC/SP. (http://lael.pucsp.br/lael)

Tony Berber Sardinha is Associate Professor at the Graduate Program in Applied Linguistics and the Linguistics Department, Catholic University of São Paulo, as well as a researcher with the Brazilian National Research Council (CNPq). His main interests are in the areas of Corpus Linguistics and Metaphor. tony4@uol.com.br