Como Encontrar as Palavras-chave mais Importantes de um Corpus com WordSmith Tools*

(How to Find the most Important Keywords in a Corpus with WordSmith Tools)

Tony Berber-Sardinha**
(PUC-SP)

Abstract: *One of the most sensitive issues surrounding a keywords analysis with WordSmith Tools is the selection of a subset of words in a corpus that deserve being looked at in greater detail. This selection is normally needed because the size of the key word list can reach several hundred, up to 1,500 or more. One way to extract a selection consists of the pulling out 'exclusive key words'. This key lexis is made up of keywords that only in a single corpus only, in comparison with a bank of keyword lists. Nevertheless, comparing several keyword lists together is a demanding task, which most users of WordSmith Tools are not expected to cope with. An alternative would be the application of a general cut-off point, established through previous uses of the keyword bank. Such a cut-off point would indicate the section of a keyword list where it would be more likely to find exclusive keywords, with a certain degree of likelihood. The results obtained here suggest that the area corresponding to the top 31% to 53% of a keyword list are more likely to contain exclusive keywords.*
Key-words: *Corpora; key words; WordSmith Tools; key lexis.*

Resumo: *Um dos procedimentos mais delicados envolvidos numa análise de corpus via palavras-chave com WordSmith Tools KeyWords é a seleção de um sub-conjunto de palavras para serem investigadas em detalhe. A seleção se faz necessária, via de regra, porque o tamanho do léxico chave de um corpus de estudo é em geral muito grande, em geral em torno de 1500 palavras ou até mais. Uma maneira de fazer esse recorte consiste na extração de palavras-chave exclusivas. O léxico chave exclusivo é composto das palavras-chave que ocorrem somente no corpus de estudo em questão em comparação*

---

* Uma versão prévia deste artigo apareceu no DIRECT Paper 41, http://lael.pucsp.br/direct .

** O autor agradece os auxílios fornecidos pelo CNPq (350455/03-1) e pela CAPES (397/04-0).

com palavras-chave de outros corpora de estudo. Contudo, comparar a lista de palavras-chave com várias outras é um procedimento custoso e complicado, que não pode ser exigido da maioria dos usuários de WordSmith Tools KeyWords. Uma alternativa para este cenário seria a aplicação de um ponto de corte generalizado baseado em tendências de retorno de palavras-chave observadas através da aplicação do banco de palavras-chave existente. Tal ponto de corte indicaria a região da lista de palavras-chave na qual há maior probabilidade de ocorrência do léxico chave exclusivo. Os resultados obtidos aqui indicam um ponto de corte entre 31% a 53% das palavras da lista, a partir da primeira de uma lista ordenada por chavicidade.
Palavras-chave: *Corpora; Palavras-chave; WordSmith Tools; léxico chave.*

1. Introdução

O procedimento de extração de palavras-chave de um corpus de estudo através do programa KeyWords da suíte WordSmith Tools (Scott 1998) tem se mostrado muito útil como auxílio na análise de aspectos discursivos de uma variedade de tipos de texto. A ferramenta KeyWords contrasta uma lista de palavras (ou mais de uma) de um corpus de estudo com uma lista de palavras de um corpus de referência, produzindo uma terceira lista contendo somente as palavras-chave do corpus de estudo. Palavras-chave são aquelas cujas frequências são estatisticamente diferentes no corpus de estudo e no corpus de referência.

Um dos procedimentos mais delicados envolvidos numa análise de corpus via palavras-chave com WordSmith Tools KeyWords é a seleção de um sub-conjunto de palavras para serem investigadas em detalhe. A seleção se faz necessária, via de regra, porque o tamanho do léxico chave de um corpus de estudo é em geral muito grande. Berber Sardinha (1999) estimou a quantidade média de palavras-chave numa coletânea de 40 corpora de estudos variados como sendo da ordem de 1472, com alguns corpora ultrapassando a marca de 3000 palavras chave. Devido a esta quantidade grande de palavras chave, o analista se vê forçado a fazer um recorte de seus dados para estudo mais detalhado.

Sabendo disso, o próprio programa estabelece um teto máximo geral, de 500 palavras, como default. Ou seja, para qualquer análise, são mostradas apenas as primeiras 500 palavras-chave. Mas mesmo essas 500 palavras já formam um número muito alto para um analista de corpus

informatizado. Até porque a intenção de usar uma ferramenta computacional de análise é justamente que ela identifique um subconjunto dos dados que mereça maior atenção. Porém, se esse número permanece além da capacidade de análise humana, então a ferramenta perde seu valor. Embora o valor de 500 palavras possa ser alterado pelo usuário, a questão ainda permanece sobre qual seria um número razoável que garantisse a representatividade das palavras-chave. Em outras palavras, seria importante podermos aplicar um ponto de corte criterioso a listas de palavras-chave.

2. Léxico chave exclusivo e o ponto de corte generalizado

Uma maneira de fazer um recorte na lista de palavras-chave consiste na extração de palavras-chave exclusivas (Berber Sardinha 1999). *O léxico chave exclusivo* é composto das palavras-chave que ocorrem somente no corpus de estudo em questão em comparação com palavras-chave de outros corpora de estudo. Um pré-requisito para a localização das palavras-chave exclusivas é que os corpus de estudo para comparação sejam em grande número e de composição variada, visto que a quantidade de palavras-chave exclusivas é super-estimada quando da comparação com poucos corpora. É mais seguro que se faça a comparação com muitos corpora de estudo para que se tenha um léxico chave exclusivo mais robusto. Além disso, os resultados são mais eficientes, já que quanto maior o número de corpora de estudo envolvidos na comparação, menor é o conjunto de palavras-chave exclusivas resultantes. Vale lembrar que por palavra-chave exclusiva entendemos apenas a forma ortográfica e não o seu sentido, visto que uma mesma forma ortográfica pode possuir sentidos diversos. Assim sendo, mesmo que uma palavra-chave ortográfica não seja exclusiva, seu sentido pode ser, o que significaria dizer que se trata, de fato, de uma palavra-chave cuja forma não é exclusiva mas o sentido o é. Esse é um assunto bastante complexo que abarca questões de polissemia, estando o seu aprofundamento além do escopo do presente trabalho.

Berber Sardinha (1999) apresenta uma implementação do procedimento de localização de palavras-chave exclusivas do inglês através da comparação de corpora de estudo com um *banco de palavras chave*. O banco é formado por 40 corpora de estudo diferentes, de fala e escrita, e devido à sua diversidade, é adequado para servir como referente para a filtragem do léxico chave. Com a aplicação do banco de palavras chave, a quantidade

média de palavras-chave por corpus de estudo foi de 535, muito menor do que as 1472 palavras-chave em média retornadas pelos corpora de estudo originalmente.

Um problema com a utilização de um banco de palavras-chave para extração do léxico chave exclusivo é que um banco variado é difícil de construir. Poucos pesquisadores dispõem de tempo e condições materiais para a criação de um banco próprio. Outro problema é que o único banco de palavras-chave pronto, conhecido, é de acesso restrito a um grupo de pesquisa (mais especificamente, o projeto DIRECT, da PUCSP). Com todas estas restrições, torna-se difícil o aproveitamento das vantagens de um banco de palavras-chave pela maioria dos pesquisadores da área.

Uma alternativa para este cenário seria a identificação de um *ponto de corte generalizado* baseado em tendências de retorno de palavras-chave observadas através da aplicação do banco de palavras-chave existente. Um ponto de corte generalizado seria um índice em porcentagem que indicasse a parcela de palavras-chave que se deveria manter de modo que se tenha a probabilidade de inclusão das palavras-chave exclusivas de um corpus de estudo. Em outras palavras, *o ponto de corte indicaria a região da lista de palavras-chave na qual há maior probabilidade de ocorrência do léxico chave exclusivo.* A probabilidade é entendida como uma tendência estatística empiricamente atestada. A aplicação do ponto de corte sugerido seria então um procedimento válido, e a decisão de aplicá-lo não seria arbitrária, visto que estaria respaldada por pesquisa empírica prévia.

O objetivo do presente trabalho é justamente buscar um ponto de corte com as condições descritas acima. Para tanto, serão investigadas as quantidade de palavras-chave exclusivas presentes nos 40 corpora do banco de palavras-chave exclusivas. A motivação do trabalho é que o ponto de corte sugerido seja de utilidade para todos os analistas de palavras chave, e mais especificamente, para aqueles pesquisadores que não dispõem de acesso a um banco de palavras-chave para verificação *in loco* das palavras-chave exclusivas do seu corpus de estudo.

3. Metodologia

Os corpora (em inglês) empregados na pesquisa foram os constantes no banco de palavras chave, os quais são os seguintes:

Tabela 1 – Composição dos corpora de estudo

		Itens	Formas
1	Livros acadêmicos	64255	8062
2	Palestras acadêmicas	29598	2639
3	Relatórios anuais de negócio	168972	8570
4	Editais de licitação	16178	1903
5	Biografias	90717	11902
6	Folhetos de negócio	15179	3494
7	Folhetos governamentais	4251	1227
8	Folhetos de hotéis	64984	6157
9	Folhetos de escola	17154	3365
10	Folhetos turísticos	115103	12185
11	Reuniões de negócio	12648	1762
12	Circulares	2613	947
13	Conversa face-a-face	361096	15473
14	Conversa telefônica	62974	4843
15	Relatórios do Supremo Tribunal	4321	1038
16	Interrogatório jurídico	4991	723
17	Palestras em jantar	5141	1047
18	Editoriais de jornal	54626	8582
19	Artigos de enciclopédia	226107	19062
20	Ficção	235095	17808
21	Relatórios governamentais	42083	4946
22	Cartas-convite	714	326
23	Cartas de pedido de emprego	12089	2412
24	Notícias	89674	11781
25	Resenhas jornalísticas	35741	7746
26	Debates parlamentares	14918	2324
27	Discurso falado político	5020	1257
28	Apresentação de caso no tribunal	9927	1640
29	Cartas de proposta	22292	3586
30	Cobertura jornalística ao vivo	13978	2575
31	Documentário jornalístico	4789	1015
32	Aulas radiofônicas	11327	1764
33	Narração esportiva radiofônica	25791	2918
34	Sermão religioso	5071	1288
35	Cartas-resposta	7622	1859
36	Artigos de pesquisa acadêmicos	621512	27368
37	Artigos de revista especializada	232046	20316
38	Redações escolares	25062	3149
39	Manuais técnicos	4377	907
40	Palestras universitárias	5012	1330
	Total	2745048	71163

A intenção ao se selecionar esta ampla gama de corpora de estudo foi a de se ter uma medida a mais conservadora o possível da exclusividade das palavras chave. O estatuto de exclusividade imputado a um item lexical não é absoluto. Nenhuma palavra-chave é exclusiva por natureza; a exclusividade é relativa às comparações efetuadas entre os corpora de estudo. Quanto maior o número de comparações, menor é a quantidade de palavras-chave exclusivas, pois há mais probabilidade de uma mesma palavra ser chave em vários corpora.

Os procedimentos adotados no estudo foram os seguintes:

(1) Extração das palavras-chave para cada um dos 40 corpora de estudo tendo como corpus de referência as edições completas dos anos de 1991 a 1994 do jornal inglês 'The Guardian', que totalizam 95 milhões de palavras. Este corpus é tipo como o corpus de referência padrão para estudos de palavras chave, tendo sido usado em vários estudos diferentes. As listas foram salvas ordenadas por *chavicidade* ('keyness'). Os ajustes do programa KeyWords para extração das palavras foram os seguintes:

Tabela 2 – Ajustes do programa KeyWords utilizados na pesquisa (na versão 3 do programa)

Ajuste	Valor
Procedimento	Log-likelihood
Max p. value	0.05
Max wanted	16000*
Min frequency	2

* máximo permitido

(2) Agrupamento dos níveis de significância. Os níveis de significância retornados pelo programa são contínuos, variando de 0,05 (adotado neste estudo) a 0,0000001 (mínimo possível). Entretanto, antes da rodagem do programa, no ato de ajuste do nível de significância desejado para a análise, os níveis de significância aparecem agrupados. Por isso, para todos os efeitos, o usuário tem à sua disposição um leque de *patamares* de significância, dentro dos quais há uma grande variação em termos de casas decimais. Assim, para refletir estes patamares de escolha, cada palavra-chave foi colocada em um dos seguintes níveis de significância agrupados:

Tabela 3 – Níveis de significância agrupados

De	A	Nível agrupado
0,05	0,009999999	0,05
0,01	0,000999999	0,01
0,001	0,000099999	0,001
0,0001	0,000009999	0,0001
0,00001	0,000000999	0,00001
0,000001	0,000000099	0,000001
0,0000001	< 0,0000001	0,0000001

(3) Identificação de palavras-chave exclusivas. Foi feita uma comparação entre as listas de palavras-chave de cada corpus de estudo através de uma rotina executada no programa SAS para análises estatísticas. As palavras-chave repetidas foram identificadas e apagadas. As palavras restantes foram consideradas *exclusivas* pois somente ocorreram em um dos corpora de estudo.

(4) Atribuição de uma posição ordinal para cada palavra-chave exclusiva, posição esta que era *idêntica* à posição ordinal na lista de palavras-chave *original* classificada por chavicidade. Por exemplo, suponha que no corpus 'x' a palavra 'y' é chave e possui a posição de número 10 na lista ordenada por chavicidade. Agora suponha que esta palavra, após a comparação, é identificada como exclusiva, ao contrário das anteriores da lista (as palavras-chave de número 1 a 9). Embora seja a primeira palavra-chave exclusiva da lista, a palavra 'x' continua mantendo seu número de ordem original 10, e não 1. A manutenção do número original é importante porque é o número fornecido pelo programa KeyWords ao pesquisador, e é justamente com base neste número inicial que o ponto de corte generalizado será calculado. Se assim não fosse, o ponto de corte seria inútil, pois pressuporia a existência de uma lista de palavras exclusivas.

(5) Identificação da posição na lista de palavras-chave na qual ocorria a *última* palavra-chave exclusiva. Por última entende-se a palavra cujo número de ordem era o mais alto entre as palavras-chave exclusivas. Por exemplo, se no corpus 'x' a palavra 'y' é exclusiva e é a 5ª palavra da lista (em ordem de chavicidade), e a palavra 'z' é também exclusiva, mas é a 6ª palavra da lista, então a palavra 'z' é considerada a última entre as exclusivas, pois ocorre na 6ª posição (versus a 5ª posição da outra palavra).

(6) Localização dos pontos de corte para cada corpus de estudo. Os pontos de corte foram considerados como sendo a posição correspondente à última palavra-chave exclusiva de cada corpus. Tomando o exemplo anterior, o ponto de corte do corpus em questão seria equivalente à 6ª palavra chave. Os valores dos pontos de corte foram calculados em *porcentagem* do total de palavras-chave originais. Assim, se no corpus de estudo citado no exemplo anterior houvesse 10 palavras chave, o ponto de corte seria na posição 60%, visto que a 6ª posição corresponde a 60% de uma lista de 10 palavras.

4. Resultados

Os pontos de corte identificados nos vários corpora são elencados na Tabela 4. Em todos os corpora nota-se um aumento do ponto de corte com o aumento do valor da significância. Por exemplo, no corpus de 'apresentação de caso no tribunal', o ponto de corte passa de 24,7% para 100%, e no de artigos de enciclopédia, de 32,45% para 99,98%.

Tabela 4 – Pontos de corte (em %) dos vários corpora, por níveis de significância agrupados

Corpus	0,0000001	0,000001	0,00001	0,0001	0,001	0,01	0,05
Apresentação de caso no tribunal	24.70	25.42	32.78	39.67	50.36	76.48	100.00
Artigos de enciclopédia	32.45	40.14	44.70	54.12	65.69	83.43	99.98
Artigos de pesquisa acadêmicos	49.38	49.38	65.90	70.51	78.65	89.59	100.00
Artigos de revista especializada	21.50	26.77	31.20	40.23	52.67	74.81	100.00
Aulas radiofônicas	36.65	39.23	43.83	52.67	64.64	83.06	99.82
Biografias	21.40	22.65	27.49	35.49	47.63	71.07	99.95
Cartas de pedido de emprego	25.00	28.66	33.12	42.52	52.87	73.25	96.98
Cartas de proposta	26.76	28.29	34.27	42.49	54.81	76.06	98.24
Cartas-resposta	23.88	34.17	35.74	42.96	57.73	79.73	93.99
Circulares	22.63	22.63	28.42	38.95	41.05	65.79	100.00
Cobertura jornalística ao vivo	32.07	32.19	40.62	50.84	64.85	82.21	98.18
Conversa face-a-face	33.39	35.42	46.47	52.73	63.10	80.77	100.00
Conversa telefônica	41.01	43.31	48.18	54.59	64.91	82.22	99.71
Debates parlamentares	23.30	26.60	31.65	39.42	55.34	74.76	98.45
Discurso falado político	18.28	21.27	25.75	30.60	51.49	69.78	88.43
Documentário jornalístico	28.77	51.15	38.01	45.21	58.90	77.74	93.49
Editais de licitação	39.64	42.30	49.58	58.68	70.45	86.27	98.04
Editoriais de jornal	16.75	19.13	23.61	30.04	43.61	68.16	99.71
Ficção	27.94	32.86	37.04	45.37	58.31	78.95	100.00

Folhetos de escola	36.61	38.98	41.23	51.18	63.15	79.15	99.65
Folhetos de hotéis	50.64	53.09	58.31	65.99	75.52	89.42	99.96
Folhetos de negócio	36.78	39.42	45.96	55.13	67.21	85.20	96.46
Folhetos governamentais	26.35	29.56	35.96	50.00	60.34	64.34	90.39
Folhetos turísticos	44.54	46.78	52.17	60.11	71.57	86.29	100.00
Interrogatório jurídico	30.64	30.64	44.68	44.68	52.77	75.74	96.60
Livros acadêmicos	28.12	30.46	34.79	44.15	57.79	77.45	99.50
Manuais técnicos	34.84	40.51	43.63	52.69	65.44	81.30	100.00
Narração esportiva radiofônica	45.54	48.11	52.99	60.81	70.21	84.98	99.39
Notícias	26.58	28.06	34.39	42.77	54.82	75.42	99.51
Palestras acadêmicas	45.85	48.21	52.05	60.35	71.70	86.72	99.91
Palestras em jantar	20.32	30.68	30.68	40.64	53.39	70.12	97.21
Palestras universitárias	14.89	18.09	21.63	32.27	49.65	70.57	98.58
Redações escolares	30.62	32.83	38.48	43.45	56.00	76.00	99.03
Relatórios anuais de negócio	52.84	54.77	59.08	66.47	74.35	87.74	99.93
Relatórios do Supremo Tribunal	20.51	25.64	28.94	40.66	57.14	73.26	98.90
Relatórios governamentais	33.19	34.75	41.13	48.58	61.06	80.50	98.87
Resenhas jornalísticas	17.76	20.11	25.17	33.26	46.64	73.32	99.62
Reuniões de negócio	32.55	34.90	38.80	49.22	61.46	77.86	96.62
Sermão religioso	19.37	28.35	28.52	38.38	53.17	64.08	88.45

A fim de se saber se as diferenças entre os níveis de significância eram estatisticamente significantes, foi feita uma Análise de Variância tendo-se como variáveis os valores dos pontos de corte distribuídos entre os níveis de significância agrupada e os corpora. O teste foi feito pelo procedimento General Linear Models disponível no programa SAS. O modelo testado foi ponto de corte = significância agrupada * corpus. Os resultados aparecem na Tabela 5 a seguir.

Tabela 5 – Resultados da Análise de Variância

Fonte	Gl	SS	F	p
Significância Agrupada	44	156798,44	214,83	< 0,0001
Erro	220	3649,31		
Total Corrigido	264	160447,75		

O valor de $F(44,220)=214.83$ é significante ($p<0.0001$), o que indica que há diferença entre os pontos de corte. Para se saber quais eram os níveis que apresentavam diferenças, os dados foram subsequentemente submetidos ao Teste F Múltiplo de REGWF (sigla proveniente de Ryan-Einot-Gabriel-Welsch). Os resultados aparecem na tabela a seguir.

Tabela 6 – Resultados do Teste F Múltiplo de REGWF

Agrupamento	Média do ponto de corte	Significância agrupada
G	30,6	0,0000001
F	34,3	0,000001
E	39,2	0,00001
D	47,5	0,0001
C	59,5	0,001
B	78,1	0,01
A	98,2	0,05

O teste REGWF apresenta os resultados em termos de agrupamentos identificados por uma letra do alfabeto. Níveis de significância agrupadas que possuem a mesma letra de agrupamento não são diferentes entre si. A tabela indica que não há nenhum nível que compartilhe o mesmo agrupamento. Isto significa que todos os níveis são diferentes entre si.

Estes resultados indicam que não é possível indicar-se um ponto de corte sugerido único, independente de nível de significância. Para que se estabeleça um ponto de corte é preciso que se defina previamente um nível de significância desejado.

Note que a Tabela 6 mostra os valores médios dos pontos de corte. Mas para que sirvam como pontos de corte sugeridos, os valores devem ser expressos em valores máximos. A Tabela 7 a seguir apresenta os pontos de corte sugeridos para os vários níveis de significância.

Tabela 7 – Pontos de corte sugeridos (valores máximos)

Nível de significância	Ponto de corte (%)
0,0000001	52,8
0,000001	54,8
0,00001	65,9
0,0001	70,5
0,001	78,6
0,01	89,6
0,05	100,0

Segundo a Tabela 7, para p=0,0000001, o ponto de corte sugerido é 52,8% da lista de palavra chave. Na prática, isto significa que mantendo-se as 53% primeiras palavras-chave da lista (ordenada por chavicidade), seriam mantidas todas as palavras-chave exclusivas dos 40 corpora investigados neste estudo. Com p=0,000001, o ponto de corte é um pouco mais alto, 54,8%. Assim, mantendo-se as primeiras 55% das palavras-chave de todos os corpora, conseguir-se-ia incluir todas as palavras-chave exclusivas. Os pontos de corte sobem continuamente, chegando a 100% com p=0,05. Isso implica em dizer que não há ponto de corte sugerido para este nível. O analista teria de incluir todas as palavras-chave obtidas. É preciso ressaltar que esses valores são relativos a esses corpora e que podem ser diferentes com outros corpora. Embora tenhamos tentado abarcar uma ampla gama de gêneros, a variação existente na língua ainda é muito maior do que a representada no corpus. Sendo assim, é prudente utilizar esses pontos de corte com cautela, levando-se em conta que o corpus em questão pode divergir em conteúdo, vocabulário e tantas outras dimensões dos analisados nesta pesquisa.

5. Discussão

Os resultados indicam que a idéia de um ponto de corte único para todos os níveis de significância disponíveis no programa KeyWords deve ser revista. Os valores dos pontos de corte diferem significantemente entre si entre os valores de p agrupados. Valores menos expressivos de significância (isto é, mais próximos de 0,05) exigem pontos de corte maiores.

O conceito de ponto de corte sugerido empregado baseia-se nos valores máximos. Mas conforme mostrou a Tabela 4, os valores variam muito entre os corpora. Assim, se quiséssemos estabelecer um valor único para todos os corpora, teríamos de adotar o valor mais alto entre todos os corpora, que seria de 53% (para p=0,0000001, relativo a relatórios anuais de negócios). Esse valor é muito distante do valor exigido pelo corpus de palestras universitárias, que é de apenas 14,89%, o que significa de que com apenas 15% das (primeiras) palavras-chave deste corpus ter-se-ia mantido todas as suas palavras-chave exclusivas.

O valor médio aritmético é um valor menos extremo de ponto de corte, e pode ser uma alternativa ao ponto de corte entendido como valor

máximo. Seguindo-se estes valores, inclui-se a maioria das palavras-chave exclusivas do corpus de estudo. O valor médio para p=0.0000001 é 30,62. Assim, com 31% das palavras-chave é possível manter-se a maioria do léxico chave exclusivo dos 40 corpora. A Tabela 8 a seguir mostra os valores médio e mínimo dos vários níveis de significância agrupada.

Tabela 8 – Valores médio e mínimo em % dos pontos de corte

	Médio	Mínimo
0,0000001	30,6	14,9
0,000001	34,3	18,1
0,00001	39,2	21,6
0,0001	47,5	30,0
0,001	59,5	41,1
0,01	78,1	64,1
0,05	98,2	88,4

A aplicação do valor médio do ponto de corte deve ser feita com cuidado. O analista deve ter sempre em mente que está na verdade diminuindo a probabilidade de sua filtragem incluir as palavras-chave exclusivas. Já a aplicação do valor mínimo é inaceitável, pois significa que apenas um dos corpora possuía este ponto de corte. Assim, se o analista adotar o valor mínimo estará eliminando a maioria das palavras exclusivas da maioria dos corpora.

Em resumo, os resultados obtidos aqui indicaram uma diferença significativa entre os pontos de corte obtidos com níveis de significância diferentes. Mais especificamente, os pontos de corte do nível 0,05 eram significativamente maiores do que os do nível seguinte (0,01), e assim sucessivamente. Isto significa na prática que o ponto de corte generalizado deveria corresponder àquele do valor máximo, de 0,05. Mas aceitar este valor seria contraproducente, pois os pontos de corte deste nível estão muito próximos dos 100%, o que significa praticamente nenhum ponto de corte. Além disso, o programa KeyWords dá liberdade aos analistas de escolher o nível de significância desejada, e assim é natural que a escolha do nível de significância seja uma variável inerente à localização das palavras-chave e, por conseguinte, das palavras-chave exclusivas também. Dessa forma, é mais lógico que se pense em um ponto de corte relativo ao nível de signi-

ficância que se deseje. Como o intuito é reduzir a lista de palavras-chave de modo que se mantenha aquelas palavras mais estatisticamente características do corpus de estudo, a sugestão que se apresenta mais apropriada é a de se utilizar como ponto de corte generalizado o valor referente ao nível de significância mais restritivo, qual seja, 0,0000001.

O valor de significância de 0,0000001, deve-se frisar, é relativo somente ao desenho de pesquisa empregado neste trabalho. Ele não significa o valor que o analista deva colocar em prática na sua análise. Na prática, ele indica um ponto de corte generalizado mais eficiente, já que:

(a) As palavras-chave exclusivas com este nível de significância são aquelas cujas frequências no corpus de estudo são mais altas, comparativamente, ao corpus de referência.

(b) O ponto de corte relativo a este nível é o mais eficiente, pois reduz em maior quantidade a lista de palavras chave.

6. Considerações finais

Este trabalho buscou trazer uma contribuição à tarefa de se fazer um recorte na lista de palavras-chave retornadas pelo programa KeyWords. Um problema encontrado nesta tarefa é que não há critérios objetivos para a escolha das palavras chave. Uma alternativa sugerida anteriormente é comparar-se a lista de palavras do corpus de estudo em questão com um banco de palavras chave, a fim de se descobrir quais palavras-chave são exclusivas daquele corpus de estudo. Embora pertinente, esta alternativa não é viável para a maioria dos pesquisadores, pois esbarra no problema da disponibilidade restrita do único banco de palavras-chave abrangente de que se tem notícia.

Uma outra possibilidade aventada é a observação de um ponto de corte generalizado, atestado no banco de palavras chave, que sirva de guia geral para os pesquisadores. De posse deste ponto de corte, os analistas poderiam efetuar o recorte nas suas listas de palavras com a confiança de que estariam selecionando uma parcela das palavras-chave na qual há a maior probabilidade de se encontrarem as palavras-chave exclusivas de seu corpus de estudo.

Os resultados obtidos aqui indicaram uma diferença significativa entre os pontos de corte obtidos com níveis de significância diferentes. O ponto de corte generalizado mais eficiente é aquele referente ao nível de significância mais restritivo (0,0000001), o qual correspondia a 53%. Desta forma, o analista que fizer um recorte escolhendo as 53% primeiras palavras-chave de sua lista ordenada por chavicidade terá uma probabilidade alta de estar selecionando as palavras-chave exclusivas de seu corpus. E como as palavras-chave exclusivas são um tipo de léxico caracterizador, o analista tem o respaldo do presente trabalho de que as palavras-chave selecionadas são aquelas que são provavelmente as mais caracterizadoras de seu corpus de estudo.

De modo geral, espera-se que o presente trabalho tenha contribuído de modo prático para que os analistas de palavras-chave façam uma escolha mais objetiva das palavras-chave para estudo detalhado em suas respectivas pesquisas.

E-mail: tony4@uol.com.br
Recebido em março de 2004
Aprovado em setembro de 2004

Referências Bibliográficas

Berber Sardinha, A. P. 1999. O banco de palavras-chave. *DIRECT Papers*, 39. Disponível eletronicamente em http://lael.pucsp.br/direct

Scott, M. 1998. *WordSmith Tools Version 3*. Oxford: Oxford University Press.