# A influência do tamanho do corpus de referência na obtenção de palavras chave[1]

Tony Berber Sardinha
LAEL, PUCSP
Programa de Estudos Pós-Graduados em Linguística Aplicada e Estudos da Linguagem
Pontifícia Universidade Católica de São Paulo
Rua Monte Alegre 984
05014-001 São Paulo, SP, Brazil
tony4@uol.com.br
http://sites.uol.com.br/tony4/homepage.html

DIRECT Papers 38
ISSN 1413-442x

1999

## 1. Introdução

O programa de computador 'WordSmith Tools' (Scott, 1998) tem se tornado uma referência para estudiosos da linguagem por meio de computador. Há vários estudos publicados (Barbara e Scott, 1999; Batista, 1998; Bonamin, 1999; Freitas, 1997; Ramos, 1997; Santos, 1999; Silva, 1999) ou em curso (Conde, 1999; Dutra, 1999; Fuzetti, 1999; Lima-Lopes, 1999; Lopes, 1998) que se utilizam do programa para a análise de dados.

Há várias razões para esta preferência. A primeira é a facilidade de uso; trata-se de um programa escrito para o ambiente Windows, o ambiente operacional dominante no mundo de hoje, o que significa dizer que a maioria dos interessados já terão alguma familiaridade com WordSmith Tools mesmo antes de se utilizarem do programa propriamente dito. A segunda razão é decorrência da primeira: devido ao fato de rodar num ambiente gráfico como Windows, o programa oferece uma facilidade maior na utilização de seus recursos disponíveis, o que por sua vez propicia um aprendizado mais rápido e intuitivo de suas várias funções. A terceira razão é a facilidade de obtenção: o programa é distribuído por uma grande editora internacional (Oxford University Press), o que facilita sua penetração em vários cidades do mundo e em pontos de venda de alta visualização, como congressos e encontros acadêmicos. Além disso, o programa é também disponibilizado via Internet, o que significa que o usuário não precisa comprá-lo numa loja ou por correio, bastando baixá-lo da rede e encomendar a sua senha pagando com cartão de crédito. A quarta razão do sucesso de WordSmith Tools é sua versatilidade. O software consiste na verdade de uma 'suíte' de diferentes programas, que se destinam a várias aplicações, que compreendem o pré-processamento, a organização de dados, e a análise propriamente dita de corpora ou textos isolados. O programa oferece ferramentas para a consecução de tarefas essenciais, como listas de palavras (através do programa WordList) e de concordâncias (por meio do Concord).

## 2. Palavras chave

Uma das razões para o sucesso de WordSmith Tools é talvez a ferramenta KeyWords, a qual se destina à comparação de listas de palavras. KeyWords contrasta uma lista de palavras (ou mais de uma) de um corpus de estudo com uma lista de palavras de um

[1] O autor agradece a Mike Scott pelos comentários sobre uma versão anterior do trabalho.

corpus de referência. O resultado do contraste é uma lista de palavras chave, ou palavras cujas frequências são estatisticamente diferentes no corpus de estudo e no corpus de referência. As palavras chaves obtidas deste modo têm se mostrado muito úteis na investigação de aspectos textuais importantes, como a temática ('aboutness'), estilo, e organização retórica (Batista, 1998; Bonamin, 1999; Conde, 1999; Dutra, 1999; Freitas, 1997; Fuzetti, 1999; Lima-Lopes, 1999; Lopes, 1998; Ramos, 1997; Santos, 1999; Silva, 1999).

Os componentes principais de uma análise de palavras chave são, portanto, dois:

(a) um corpus de estudo, representado em uma lista de frequência 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 frequência 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 frequências do corpus de estudo. A comparação é feita através de uma prova estatística selecionada pelo usuário (qui-quadrado ou log-likelihood). As palavras cujas frequências 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 procedimento para extração de palavras chave pode ser resumido segundo o algoritmo abaixo:

(1) Selecione o primeiro item na lista de palavras do corpus de estudo;
(2) Procure por este item na lista de palavras do corpus de referência;
(3) Se o item constar do corpus de referência, vá para o passo a seguir, senão passe para o passo 7;
(4) Compare as frequências através de uma prova estatística escolhida pelo usuário (log-likelihood é 'default', mas qui-quadrado também está disponível);
(5) Se o resultado da comparação for estatisticamente significante (segundo o nível de significância definido pelo usuário), copie esta palavra para uma nova lista, e chame-a de lista de palavras chave;
(6) Repita este procedimento até o último item da lista de palavras do corpus de estudo;
(7) Se um item constante da lista de palavras do corpus de estudo não aparecer na lista de palavras do corpus de referência, assuma frequência 0 para o item no corpus de referência;
(8) Execute os passos 4, 5, e 6.

## 3. Variação no conjunto de palavras chave

As listas de palavras chave obtidas segundo o algoritmo descrito acima variam de acordo com alguns parâmetros:

- A natureza do corpus de estudo;
- A natureza do corpus de referência;
- O tamanho do corpus de estudo;
- O tamanho do corpus de referência;
- O nível de significância para se atingir a chavicidade ('keyness').

Os dois primeiros parâmetros referem-se ao conteúdo dos corpora a serem comparados. Em relação ao corpus de estudo, as palavras chave encontradas numa comparação típica em geral se referem à temática do corpus de estudo, e por isso são intrínsecas a várias características inerentes à textualidade dos textos (ou texto) que compõem o corpus de estudo. Em outras palavras, as palavras chave são específicas daquele corpus de estudo, e, desse modo, elas são intimamente ligadas à textualidade (Collins e Scott, 1997). Por isso, a variação no conjunto de palavras chave encontradas ao se comparar, por exemplo, textos semelhantes com um mesmo corpus de referência é natural, pois os textos-fonte exibem inevitavelmente uma composição específica da frequência vocabular que emana, entre outras coisas, da temática, do fraseado ('wording'), da organização genérica, do estilo dos autores, etc.

Em relação ao corpus de referência, as palavras chave obtidas tendem a ser influenciadas do seguinte modo: um corpus de características genéricas semelhantes ao corpus de estudo tende a 'filtrar', ou seja, eliminar, os elementos genéricos (i.e. relativos a um mesmo gênero) em comum, resultando em uma lista de palavras chave que não inclui estes elementos. Alguns traços linguísticos que podem vir a ser filtrados são, entre outras, marcadores discursivos privilegiados, escolhas lexicais típicas, e formas verbais flexionadas em comum. Por exemplo, se se comparar um corpus de estudo de artigos de pesquisa acadêmicos de medicina com um corpus de referência do mesmo tipo, pode se esperar que palavras como 'resultados', 'análise', 'sugerem' não se tornem chave. Já um corpus de referência de um gênero distinto do de estudo tende a não excluir tais palavras 'genéricas'. Por isso, um corpus de referência geral, que inclua vários gêneros, é tida como a escolha não-marcada para estudos de palavras chave.

Embora a natureza dos corpora de estudo e referência tendam a influenciar os resultados de forma distintas, estas influências podem ser antecipadas. Ou seja, é possível especular qual tipo de palavras chave serão obtidas a partir da escolha de um determinado corpus de estudo, com base no conhecimento dos textos-fonte, obtido por meio da leitura dos mesmos, por exemplo. Do mesmo modo, é possível prever que tipo de influência a escolha de um determinado corpus de referência terá sobre os resultados, com base no conhecimento das características genéricas dos corpora a serem comparados. É preciso salientar, entretanto, que a habilidade de antecipar o tipo de influência que a natureza dos corpora terá sobre a chavicidade das palavras não significa que se tenha a capacidade de prever com exatidão quais palavras chave serão obtidas. O que se quer dizer aqui é que é possível fazer-se previsões de caráter geral acerca do conjunto de palavras chave, previsões estas que podem ou não se confirmar.

Da mesma forma, a escolha do nível de significância tem influência conhecida: quanto menor o valor, menor o número de palavras chave resultantes. Em outras palavras, um nível de significância menor exige uma diferença mais acentuada entre as frequências para que se atinja a chavicidade.

Por outro lado, a influência do tamanho do corpus de estudo e de referência nos resultados é bem menos previsível. Algumas perguntas que um pesquisador de palavras chave pode se colocar em relação à extensão dos corpora são:

- Quantas palavras chave pode se esperar obter de um corpus de estudo x comparado a um corpus de referência y?

- Qual a diferença que se pode esperar no número de palavras chave a serem obtidas quando se usa como corpus de estudo um corpus de extensão x em vez de um de extensão y?
- Qual a diferença que se pode esperar no número de palavras chave a serem obtidas quando se usa um corpus de estudo de extensão x em vez de um de extensão y?
- E quando se usa um corpus de referência de extensão x em vez de um de extensão y?

A resposta a perguntas deste tipo é importante por pelo menos dois motivos principais. O primeiro refere-se ao planejamento da pesquisa. Sabendo-se a influência de um corpus de uma certa dimensão na chavicidade das palavras, é possível planejar qual o tamanho ideal dos corpora. E tendo-se conhecimento do tamanho ideal dos corpora, torna-se possível planejar a pesquisa de modo que não se desperdice recursos coletando-se dados além do que seria teoricamente necessário. O pesquisador, poderia, então, saber qual o impacto que um corpus de tamanho x teria nos resultados de sua pesquisa, e planejar sua coleta de dados conscientemente. Na realidade, o que tem acontecido na prática é diferente: o pesquisador coleta uma certa quantidade de dados de acordo com suas possibilidades, efetua a análise, mas não sabe se sua coleta foi além ou aquém do que seria teoricamente mais adequado.

O segundo motivo relaciona-se à confiabilidade dos resultados. Que diferença haveria, em termos do total de palavras chave, se um pesquisador tivesse optado por corpora maiores do que os que efetivamente empregou na sua análise? Sabendo-se a influência do tamanho dos corpora nos resultados, é possível dizer-se qual a quantidade palavras chave que teoricamente deixaram de ser incluídas na análise. Em casos extremos, se estas palavras forem de número apreciável, poderiam influenciar, ou até mesmo mudar, os resultados da pesquisa, o que por sua vez, poderia colocar em cheque os resultados da pesquisa.

## 4. Questões

A resposta a estas questões virá do exame empírico de resultados de análise com corpora de estudo e de referência de vários tamanhos, em busca de tendências estatisticamente robustas de variação no número de palavras chave. O presente trabalho enfocará um dos lados das questões: a do tamanho do corpus de referência. As perguntas que se colocam, são, portanto, as seguintes:

1. Quantas palavras chave são obtidas a partir de um mesmo corpus de estudo, quando este é comparado a corpora de referência de tamanhos variados?
2. Se houver, a influência do aumento do tamanho do corpus de referência é sempre constante ou há pontos em que o tamanho do corpus de referência deixa de influir na variação do número de palavras chave?
3. Há uma tendência observável de variação do número de palavras chave?
4. Esta tendência pode ser prevista matematicamente?

## 5. Metodologia

Os corpora de estudo usados nesta investigação são cinco, nomeadamente:

- Corpus de cartas de pedido de emprego, proveniente do Banco de Dados do Projeto DIRECT, doravante ‘cartas’;

- Corpus de editoriais jornalísticos, referente ao sub-corpus ‘B’ do corpus Brown, doravante ‘editoriais’;
- Corpus de resenhas jornalísticas, referente ao sub-corpus ‘C’ do corpus Brown, doravante ‘resenhas’;
- Corpus de ficção de mistério (romance, contos), referente ao sub-corpus ‘L’ do corpus Brown, doravante ‘mistério’;
- Corpus de ficção científica (romance, contos), referente ao sub-corpus ‘M’ do corpus Brown, doravante ‘sci-fi’.

Os cinco corpora de estudo totalizam cerca de 162 mil palavras, assim distribuídas:

| Corpus | Itens (‘tokens’) | Formas (‘types’) |
|---|---|---|
| Cartas | 11.761 | 2.415 |
| Editoriais | 54.626 | 8.582 |
| Resenhas | 35.741 | 7.746 |
| Mistério | 48.298 | 6.281 |
| Sci-Fi | 12.081 | 2.982 |
| Total | 162.507 | |

As razões da escolha destes corpora foram duas. A primeira é de ordem funcional. Todos os corpora foram utilizados em pesquisa prévia, e considerados ‘representativos’ dos gêneros dos quais se compõem, servindo deste modo como fonte de ‘insights’ acerca da linguagem. Além disso, por terem sido usados em pesquisa prévia reconhecida, os dados já foram validados quanto à sua constituição e lisura. Assim, ao escolher estes corpora, evitou-se fazer um exercício espúrio com dados de pouca validade fora do contexto desta pesquisa. A segunda razão é de ordem prática. Todos os dados estavam disponíveis para o pesquisador, não necessitando ser coletados.

O material para os corpora de referência foi retirado do jornal britânico ‘The Guardian’. A razão desta escolha é que o jornal tem sido uma fonte padrão de material para constituição de referência no estudo de palavras chave, tanto assim que o autor de WordSmith Tools (Mike Scott) coloca à disposição uma lista de palavras de mais de 95 milhões de palavras retiradas de quatro anos de publicação do mesmo jornal. Para este estudo, foi escolhido o ano de 1994, e os textos foram retirados aleatoriamente do material publicado pelo jornal naquele ano.

Devido ao fato de as perguntas de pesquisa enfocadas aqui centraram-se na questão da influência do tamanho do corpus de referência, foram na verdade retirados vários corpora de referência do ‘The Guardian’ de 1994. Para cada corpus de estudo, foram criados 18 corpora de referência, cujos tamanhos correspondiam a uma ordem de magnitude (i.e. um número de vezes maior do que o tamanho do corpus de estudo). As ordens de magnitude escolhidas foram as seguintes: 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, e 100. Por exemplo, o corpus de estudo de cartas possui 11.761 itens; o seu corpus de referência de magnitude 2x compreende então 23.552 itens, ou seja, 11.761 multiplicado por 2, o de 3x possui 35.283 (11.761 x 3), o de 4x  47.044, e assim por diante, até o de magnitude 100x, que dispõe de 1.176.100 ocorrências. A tabela a seguir mostra o tamanho de todos os corpora de referência usados no estudo:

| | | Magnitude do corpus de referência | | | | | |
|---|---|---|---|---|---|---|---|
| | | 2x | 3x | 4x | 5x | 6x | 7x |
| Cartas | Itens | 23.522 | 35.283 | 47.044 | 58.805 | 70.566 | 82.327 |
| | Formas | 5.543 | 7.409 | 8.863 | 10.161 | 11.163 | 12.249 |
| Editoriais | Itens | 109.252 | 163.878 | 218.504 | 273.130 | 327.756 | 382.382 |
| | Formas | 14.973 | 18.378 | 21.746 | 24.118 | 26.537 | 28.382 |
| Resenhas | Itens | 71.482 | 107.223 | 142.964 | 178.705 | 214.446 | 250.187 |
| | Formas | 11.000 | 14.331 | 17.758 | 19.490 | 21.559 | 23.402 |
| Mistério | Itens | 96.596 | 144.894 | 193.192 | 241.490 | 289.788 | 338.086 |
| | Formas | 13.880 | 17.636 | 20.285 | 22.861 | 24.925 | 26.928 |
| Sci-Fi | Itens | 24.162 | 36.243 | 48.324 | 60.405 | 72.486 | 84.567 |
| | Formas | 5.644 | 7.550 | 9.032 | 10.325 | 11.318 | 12.422 |
| | | Magnitude do corpus de referência | | | | | |
| | | 8x | 9x | 10x | 20x | 30x | 40x |
| Cartas | Itens | 94.088 | 105.849 | 117.610 | 235.220 | 352.830 | 470.440 |
| | Formas | 13.095 | 13.896 | 14.879 | 22.650 | 27.763 | 31.471 |
| Editoriais | Itens | 437.008 | 491.634 | 546.260 | 1092.520 | 1.638.780 | 2.185.040 |
| | Formas | 30.292 | 31.825 | 33.672 | 47.305 | 57.325 | 65.237 |
| Resenhas | Itens | 285.928 | 321.669 | 357.410 | 714.820 | 1.072.230 | 1.429.640 |
| | Formas | 24.940 | 26.524 | 27.812 | 38.610 | 47.081 | 53.695 |
| Mistério | Itens | 386.384 | 434.682 | 482.980 | 965.960 | 1.448.940 | 1.931.920 |
| | Formas | 28.563 | 30.084 | 31.669 | 44.755 | 53.867 | 61.531 |
| Sci-Fi | Itens | 96.648 | 108.729 | 120.810 | 241.620 | 362.430 | 483.240 |
| | Formas | 13.305 | 14.209 | 15.156 | 22.918 | 28.144 | 32.010 |
| | | Magnitude do corpus de referência | | | | | |
| | | 50x | 60x | 70x | 80x | 90x | 100x |
| Cartas | Itens | 588.050 | 705.660 | 823.270 | 940.880 | 1.058.490 | 1.176.100 |
| | Formas | 35.083 | 38.560 | 42.421 | 44.607 | 47.061 | 48.902 |
| Editoriais | Itens | 2.731.300 | 3.277.560 | 3.823.820 | 4.370.080 | 4.916.340 | 5.462.600 |
| | Formas | 71.680 | 77.397 | 82.743 | 87.902 | 92.884 | 97.121 |
| Resenhas | Itens | 1.787.050 | 2.144.460 | 2.501.870 | 2.859.280 | 3.216.690 | 3.574.100 |
| | Formas | 59.690 | 64.753 | 69.242 | 73.167 | 76.945 | 80.574 |
| Mistério | Itens | 2.414.900 | 2.897.880 | 3.380.860 | 3.863.840 | 4.346.820 | 4.829.800 |
| | Formas | 68.117 | 73.623 | 78.508 | 83.076 | 87.578 | 92.157 |
| Sci-Fi | Itens | 604.050 | 724.860 | 845.670 | 966.480 | 1.087.290 | 1.208.100 |
| | Formas | 35.460 | 38.959 | 42.822 | 45.101 | 47.474 | 49.617 |

Tabela 1: Palavras chave obtidas com os diversos corpora de referência

De posse dos corpora de estudo e de seus respectivos corpora de referência, foram extraídas as palavras chave *positivas* de cada um dos cinco corpora em comparação a cada um de seus 18 corpora de referência. Os ajustes do programa KeyWords empregados foram os seguintes:

| Ajuste | Valor |
|---|---|
| Procedimento | loglikelyhood |
| Max p. value | 0.01 |

| | |
|---|---|
| Max wanted | 16000* |
| Min frequency | 2 |

* máximo permitido

Tabela 2: Ajustes do programa KeyWords utilizados na pesquisa

## 6. Resultados

A seguir serão apresentados os resultados referentes a cada uma das três questões de pesquisa elencadas acima.

(1) Quantas palavras chave são obtidas a partir de um mesmo corpus de estudo, quando este é comparado a corpora de referência de tamanhos variados?

A tabela a seguir detalha o número de palavras chave obtidas a partir da comparação de cada corpus de estudo com seus dezoito corpora de referência respectivos. Devido ao fato de os corpora de estudo serem de tamanhos diferentes, os totais de palavras chave serão também apresentados em porcentagens do total de formas (palavras diferentes) do corpus de estudo. Por exemplo, o corpus de cartas possui 2.415 formas; as palavras chave obtidas comparando-se o este corpus com o corpus de referência de magnitude 2x foi 279; portanto, estas 279 palavras chave correspondem a 11.6% do total de formas do corpus de cartas.

| Mag. | Cartas | | Editoriais | | Resenhas | | Mistério | | Sci-Fi | |
|---|---|---|---|---|---|---|---|---|---|---|
| | P.Chave | % | P.Chave | % | P.Chave | % | P.Chave | % | P.Chave | % |
| 2x | 279 | 11,6 | 433 | 5,0 | 401 | 5,2 | 583 | 9,3 | 137 | 4,6 |
| 3x | 347 | 14,4 | 686 | 8,0 | 582 | 7,5 | 748 | 11,9 | 202 | 6,8 |
| 4x | 354 | 14,7 | 637 | 7,4 | 496 | 6,4 | 728 | 11,6 | 196 | 6,6 |
| 5x | 481 | 19,9 | 963 | 11,2 | 889 | 11,5 | 1027 | 16,4 | 363 | 12,2 |
| 6x | 480 | 19,9 | 910 | 10,6 | 872 | 11,3 | 1035 | 16,5 | 361 | 12,1 |
| 7x | 450 | 18,6 | 892 | 10,4 | 829 | 10,7 | 1018 | 16,2 | 355 | 11,9 |
| 8x | 457 | 18,9 | 887 | 10,3 | 846 | 10,9 | 1037 | 16,5 | 350 | 11,7 |
| 9x | 457 | 18,9 | 880 | 10,3 | 822 | 10,6 | 1031 | 16,4 | 332 | 11,1 |
| 10x | 462 | 19,1 | 896 | 10,4 | 837 | 10,8 | 1050 | 16,7 | 330 | 11,1 |
| 20x | 506 | 21,0 | 967 | 11,3 | 935 | 12,1 | 1119 | 17,8 | 353 | 11,8 |
| 30x | 497 | 20,6 | 960 | 11,2 | 919 | 11,9 | 1116 | 17,8 | 364 | 12,2 |
| 40x | 507 | 21,0 | 953 | 11,1 | 926 | 12,0 | 1135 | 18,1 | 367 | 12,3 |
| 50x | 490 | 20,3 | 936 | 10,9 | 914 | 11,8 | 1123 | 17,9 | 373 | 12,5 |
| 60x | 492 | 20,4 | 942 | 11,0 | 933 | 12,0 | 1141 | 18,2 | 378 | 12,7 |
| 70x | 492 | 20,4 | 928 | 10,8 | 914 | 11,8 | 1140 | 18,1 | 368 | 12,3 |
| 80x | 485 | 20,1 | 948 | 11,0 | 929 | 12,0 | 1145 | 18,2 | 374 | 12,5 |
| 90x | 485 | 20,1 | 943 | 11,0 | 922 | 11,9 | 1130 | 18,0 | 383 | 12,8 |
| 100x | 475 | 19,7 | 952 | 11,1 | 939 | 12,1 | 1143 | 18,2 | 382 | 12,8 |

Tabela 3: Total de palavras chave para cada corpus de referência
(Legenda: Magn = ordem de magnitude; p.chaves = palavras chave; % = porcentagem do total de formas do corpus de estudo).

Os resultados indicam três pontos importantes. O primeiro é um crescimento da quantidade de palavras chave que acompanha o crescimento dos corpora de referência. Para todos os corpora, o total de palavras chave obtidas com os corpora de 100x é maior do que com os corpora de 2x.

O segundo ponto é a não linearidade do crescimento do total de palavras chave. Por exemplo, o total de palavras chave para a magnitude 2x do corpus de cartas é 279, para a magnitude 3x 347, e para a magnitude 100x a contagem é de 475. Se o crescimento fosse linear e progressivo em relação à magnitude 2x, os totais seriam 418 (3x) e 13.950 (100x). Obviamente, o total de 13.950 nunca poderia ser alcançado visto que o total máximo de palavras chave passível de ser obtido é 2.415, correspondente ao total de formas do corpus de cartas. O mesmo acontece com os outros corpora. Claramente, portanto, o total de palavras chave não pode crescer linearmente em conjunto com o tamanho do corpus de referência, visto que o total máximo de palavras chave é limitado pelo número de formas no corpus de estudo, enquanto o tamanho do corpus de referência teoricamente não o é (ou seja, é possível usar-se um corpus de referência de tamanho que excederia em muito o tamanho do vocabulário do corpus de estudo).

O terceiro ponto de relevância é uma variação grande entre os corpora em relação à quantidade relativa de palavras chave entre os níveis de magnitude. O menor número é 4.6%, referentes à literatura de Sci-Fi comparada ao corpus de referência de magnitude 2x. Nesta mesma magnitude, os valores para os outros corpora são 5% (editoriais), 5.2% (resenhas), 9.3% (mistérios), e 11.6% (cartas). O maior número é 21%, correspondente ao corpus de cartas comparado ao corpus de referência 40 vezes maior. Neste mesmo patamar de tamanho, os demais valores são 11.1% (editoriais), 12% (resenhas), 12.3% (sci-fi), e 18.1% (mistérios).

Em conclusão, um corpus de referência maior do que o dobro do corpus de estudo tende a revelar mais palavras chave do que um corpus de apenas duas vezes o tamanho do corpus de estudo. Entretanto, o aumento de palavras chave não é progressivo e linear em relação ao aumento do tamanho do corpus de referência. Um corpus cem vezes maior não retorna 50 vezes mais palavras chave do que um corpus duas vezes maior.

(2) Se houver, a influência do aumento do tamanho do corpus de referência é sempre constante ou há pontos em que o tamanho do corpus de referência deixa de influir na variação do número de palavras chave?

Para se responder a esta pergunta, é necessário primeiramente observar a distribuição dos totais de palavras chave de cada corpus para cada magnitude de referência. O gráfico abaixo mostra esta distribuição.

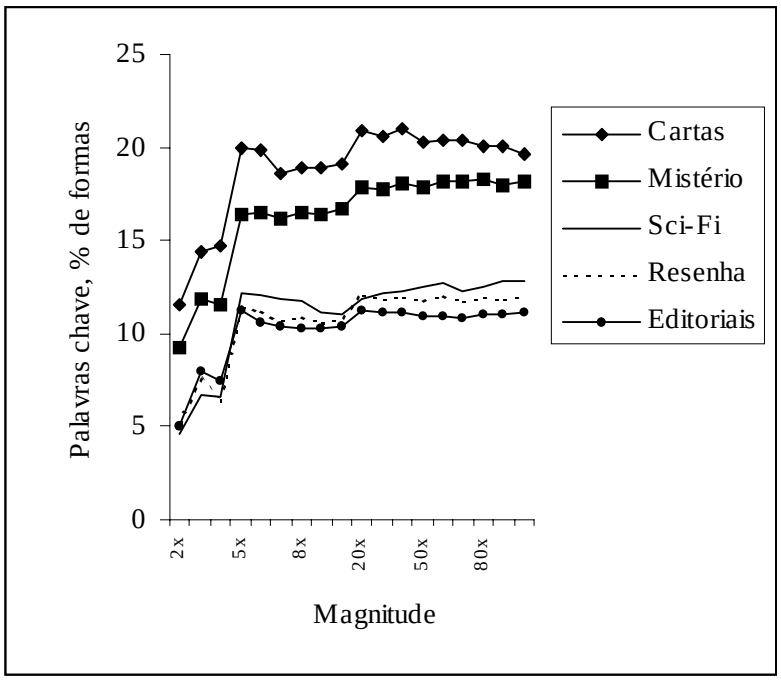


Figura 1: Distribuição de palavras chave

Percebe-se que a distribuição dos totais de palavras chave demonstra uma certa regularidade. Em todos os corpora, o total de palavras chave sobe de 2x para 3x, desce ou se estabiliza em 4x, sobe novamente em 5x e depois praticamente se estabiliza. Para confirmar, basta checar os números da Tabela 3. No corpus de cartas, as porcentagens de palavras chave, para 2x, 3x, 4x, 5x, e 6x são respectivamente 11.6, 14.4, 14.7, 19.9, e 19.9. De fato, então, há um aumento considerável de 2x para 3x (11.6 a 14.4), um aumento desprezível de 3x para 4x (14.4 a 14.7), um crescimento vertiginoso de 4x a 5x (14.7 a 19.9), e uma estabilização de 5x para 6x (19.9 a 19.9). Uma flutuação similar acontece com sci-fi, por exemplo: de 4.6 para 6.8 em 2x (aumento grande), de 6.8 para 6.6 em 3x (redução), de 6.6 para 6.8 em 4x (aumento desprezível), de 6.8 para 12.2 em 5x (aumento), e 12.2 para 12.1 (praticamente uma estabilização).

Parece haver uma influência do aumento do tamanho do corpus de referência, mas para se saber ao certo quais diferenças entre magnitudes eram significativas, é necessário submeter os resultados a uma prova estatística chamada Análise de Variância (ANOVA). Esta prova compara a variação dentro de cada corpus entre os vários níveis de magnitude. Os resultados são apresentados na tabela a seguir.

| Fonte | Gl | SS | F | p |
|---|---|---|---|---|
| Magnitude | 21 | 1540.80873556 | 267.98 | < 0.0001 |
| Erro | 68 | 18.61846000 | | |
| Total Corrigido | 89 | 1559.42719556 | | |

Tabela 4: Resultados da Análise de Variância

O valor de F(21,68)=267.98 é significante (p<0.0001), o que indica que o total de variância atribuído ao tamanho do corpus de referência para cada corpus de estudo (representado pelo total de SS na linha 'magnitude') é maior do que a variância atribuível ao acaso (o total de SS na linha 'erro'). Há, portanto, variação significativa entre os totais de palavras chave para cada ordem de magnitude. Em outras palavras, o tamanho do corpus de referência influi na quantidade de palavras chave obtidas.

Resta saber agora se esta variação que existe entre os totais de palavras chave é significativa entre todas as magnitudes ou somente entre algumas. Para responder à esta questão, é necessário executar-se mais uma prova estatística associada à Análise de Variância, o Teste F Múltiplo de REGWF (sigla proveniente de Ryan-Einot-Gabriel-Welsch). Os resultados aparecem na tabela a seguir, ordenados em ordem decrescente pela porcentagem média (entre os cinco corpora) de cada ordem de magnitude.

| Agrupamentos | | | | % Média | Magnitude |
|---|---|---|---|---|---|
| | | A | | 14,8840 | 40 |
| | | A | | 14,8480 | 60 |
| | | A | | 14,7900 | 20 |
| | | A | | 14,7780 | 100 |
| | | A | | 14,7780 | 80 |
| | | A | | 14,7600 | 90 |
| | | A | | 14,7220 | 30 |
| B | | A | | 14,6940 | 70 |
| B | | A | C | 14,6780 | 50 |
| B | D | A | C | 14,2280 | 5 |
| B | D | A | C | 14,0660 | 6 |
| B | D | | C | 13,6860 | 8 |
| | D | | C | 13,6340 | 10 |
| | D | | | 13,5660 | 7 |
| | D | | | 13,4640 | 9 |
| | | E | | 9,100 | 3 |
| | | E | | 9,3280 | 4 |
| | | F | | 7,1300 | 2 |

Tabela 5: Resultados do Teste F Múltiplo de REGWF

O teste REGWF apresenta os resultados em termos de agrupamentos identificados por uma letra do alfabeto. Os agrupamentos são formados pelas diferentes ordens de magnitude (2x a 100x). As magnitudes que possuem a mesma letra do alfabeto pertencem a um mesmo agrupamento. O fato mais importante é que *as magnitudes de um mesmo agrupamento não são significativamente diferentes entre si*. Assim, segundo a tabela, as magnitudes 40, 60, 20, 100, 80, 90, 30, 70, 50, 5, e 6 formam o agrupamento A. Este agrupamento tem médias de palavras chave que vão de 14.066% a 14.884%. Isto significa que as quantidades de palavras chave relativas ao cinco corpora nestas magnitudes não são estatisticamente significativas. Do mesmo modo, as magnitudes 70, 50, 5, 6, e 8 formam o agrupamento B, o que significa dizer que os totais de palavras chave obtidas com estas ordens de magnitude também não são estatisticamente

diferentes (médias de 13.686% a 14.694%) Note que os agrupamentos não são auto-excludentes. Assim, os agrupamentos A, B, C, e D formam um conjunto que vai das ordens de magnitude 5 a 100, relativos a, em média, 13.464% a 14.884% de palavras chave. Há outros dois agrupamentos, desta vez independentes. O agrupamento E é formado pelas magnitudes 3 e 4 (9.328% a 9.71%), e o agrupamento F é constituído somente pela magnitude 2 (7.13% de média).

Há, portanto, duas divisões básicas no espectro de palavras chave, nomeadamente relativas às magnitudes 2, 3, e 5. Estas são exatamente as marcas discernidas no exame do gráfico de distribuição de palavras chave (veja Figura 1).

O valor crítico, portanto, é cinco. Um corpus de referência cinco vezes maior que o de estudo permite extrair um número maior de palavras chave corpora de referência menores, e um número parecido de palavras chave que corpora maiores. Isto significa que os resultados de uma análise feita com um corpus de referência menor que cinco vezes o tamanho do corpus de estudo poderiam ser alterados, já que corpora menor que este tamanho tendem a retornar menor quantidade de palavras chave, o que influenciaria a interpretação dos resultados.

Em suma, os resultados indicam que o tamanho do corpus de referência influencia o número de palavras chave obtidas, mas a influência não é constante: há pontos em que o tamanho do corpus de referência é irrelevante. Mais especificamente, corpora de referência duas, três e cinco vezes maior do que o de estudo tendem a propiciar números *maiores* de palavras chave (isto é $2 < 3 < 5$); corpora de referência de outros tamanhos não (ou seja, $3 \approx 4$, $5 \approx 6$, $5 \approx 7$, … $5 \approx 100$). Deste modo, um corpus de referência que é *quatro* vezes maior do que corpus de estudo retorna um número *parecido* de palavras chave do que um corpus de referência que é apenas o *dobro* do tamanho do de estudo. Entretanto, um corpus de referência que é *quatro* vezes maior tende a fornecer *menos* palavras chave do que um corpus de referência *cinco* vezes maior que o de estudo.

Numericamente, com um corpus de referência que é o dobro do de estudo, há uma tendência de cerca de 7% do vocabulário do corpus ser palavras chave; com um corpus de referência que é 3 ou quatro vezes maior, por volta de 9% das palavras do corpus de estudo tenderão a ser chave; e com um corpus de referência 5 vezes maior ou mais, por volta de 14% do vocabulário tenderá a se constituir de palavras chave.

(3) Há uma tendência observável de variação do número de palavras chave?

Para se responder a esta questão, é necessário observar-se os comentários feitos acerca da Tabela 3 novamente. Conforme dito antes, não há uma relação entre aumento da ordem de magnitude e aumento da quantidade de palavras chave retornadas. Há um aumento, mas este não é linear. Além disto, há uma variação grande entre os totais de palavras chave, de 137 (Sci-Fi, 2x) a 1145 (Mistério, 80x).

O primeiro passo para se tentar observar uma tendência nestes dados é ver sua distribuição visualmente. O gráfico abaixo representa os totais de palavras chave, ordenados ascendentemente, obtidos em todas as 90 comparações realizadas (5 corpora de estudo vezes 18 corpora de referência).

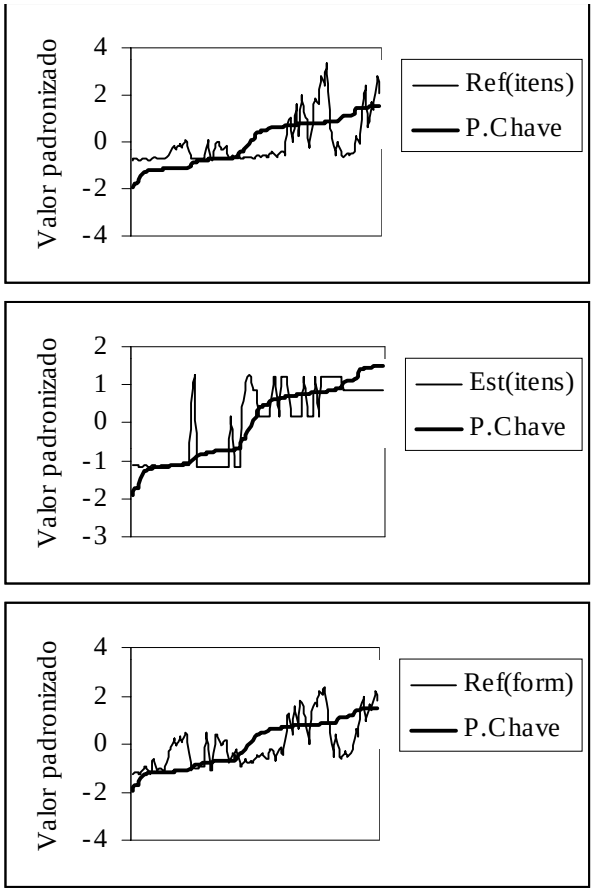


Figura 2: Distribuição dos totais de palavras chave, ordenados ascendentemente

Percebe-se que há uma visível tendência de alta no número de palavras chave. Agora é necessário saber quais são as possíveis influências nesta alta. Para tanto, inicialmente deve-se mapear os valores dos corpora de referência e de estudo sobre a linha de palavras chave. Os gráficos da Figura 3 mostram a distribuição sobreposta das palavras chave e dos corpora de referência e de estudos.

Figura 3 Distribuição comparativa de palavras chave, (itens e formas dos) corpora de referência, e (itens dos) corpora de estudo (Legenda: Ref(itens): Itens dos corpora de referência; Ref(form): formas dos corpora de referência; Est(itens): Itens dos corpora de estudo.)

Percebe-se nos gráficos que a tendência de alta no número de palavras chave é acompanhada por uma alta também no número de itens e formas dos corpora de referência, e no número de itens nos corpora de estudo. Em outras palavras, corpora maiores tendem a retornar mais palavras chave. Note que este achado não contradiz o anterior, segundo o qual um corpus de magnitude maior que cinco retorna quantidades estatisticamente semelhantes de palavras chave; tratava-se, então, da percentagem de palavras chave em relação *a cada corpus de estudo*. Agora trata-se do *total de palavras chave*. Combinando-se os dois achados, o que se conclui é que corpora maiores retornam quantidades *absolutas* maiores de palavras chaves, mas as quantidades *relativas* de palavras chave não crescem linearmente.

Em resumo, um pesquisador que se utilize de um corpus de estudo maior, e de um corpus de referência maior, obterá mais palavras chave do que um outro que se utilizar

de corpora de estudo e referência menores. Comparativamente, contudo, um aumento de seu corpus de referência não ocasionará, necessariamente, um aumento significativo de palavras chave.

(4) Esta tendência pode ser prevista matematicamente?

Por fim, para se responder a esta pergunta, é necessário submeter os dados a uma análise de regressão, que é uma técnica estatística que permite derivar um modelo matemático que explica a variação de uma determinada variável em função de outra(s). Tendo em vista as observações resultantes do exame visual dos gráficos anteriores, propôs-se a seguinte equação como modelo para a análise de regressão:

> Total de palavras chave = $\alpha + \beta_1 \times$ Itens do corpus de estudo + $\beta_2 \times$ Formas do corpus de referência + $\beta_3 \times$ Itens do corpus de referência[2]

Os resultados da análise de regressão são apresentados a seguir.

| Fonte | gl | Soma de quadrados | Quadrado médio | F | P |
|---|---|---|---|---|---|
| Modelo | 3 | 6222008,1811 | 2074002.727 | 123,117 | < 0,0001 |
| Erro | 86 | 1448739,9189 | 16845.81301 | | |
| Total | 89 | 7670748,1 | | | |
| $R^2$ | | 0,8111 | | | |
| $R^2$ Ajustado | | 0,8045 | | | |

Tabela 6: Resultados da Análise de Regressão

Segundo a tabela acima, o valor de F(3,86)=123,117, que é significativo ($p < 0,0001$). Portanto, o modelo explica a variação do total de palavras chave. O valor de $R^2$ Ajustado = 0,8045 significa que o modelo explica 80,45% da variação do total de palavras chave.

Os valores dos parâmetros estimados pela análise são elencados na tabela a seguir.

| Variável | gl | Estimativa do parâmetro | Erro padrão | T para parâmetro = 0 | P |
|---|---|---|---|---|---|
| Interseção | 1 | 95,890582 | 43,50547290 | 2,204 | 0,0302 |
| Itens do corpus de estudo | 1 | 0,011432 | 0,00085903 | 13,308 | 0,0001 |
| Formas do corpus de referência | 1 | 0,009217 | 0,00190581 | 4,837 | 0,0001 |
| Itens do corpus de referência | 1 | -0,000105 | 0,00003481 | -3,022 | 0,0033 |

Tabela 7: Parâmetros da Análise de Regressão

[2] Embora a fórmula empregue sinais de adição entre os termos, o resultado da equação não é necessariamente maior que 95,8 ($\alpha$) visto que o resultado será ditado pelo valor dos parâmetros, que podem ser negativos. Sendo negativos, o resultado final pode ser menor que 95,8.

Todos os valores são significativos ($p < 0,05$), o que indica que todos contribuem para o modelo. Transferindo-se os valores para a equação proposta inicialmente, obtém-se:

Palavras chave= 95,890582 + 0,011432 × (itens do corpus estudo) + 0,009217 × (formas do corpus de referência) – 0,000105 × (itens do corpus de referência)

Equação 1: Modelo para previsão do total de palavras chave

Com esta equação, é possível prever-se com 80% de exatidão o total de palavras chave resultantes da comparação de um corpus de estudo com um corpus de referência de dimensões conhecidas. Por exemplo, tomando-se os valores da magnitude 2 para o corpus de cartas, tem-se:

Itens do corpus de estudo: 11.761
Formas do corpus de referência: 5.543
Itens do corpus de referência: 23.522

os quais transferidos para a fórmula resultam em:

$$95,890582 + 0,011432 \times (11.761) + 0,009217 \times (5.543) - 0,000105 \times (23.522) = 279$$

Ou seja, a fórmula prediz que haverá 279 palavras chave para um corpus com estas dimensões. O total observado de palavras chave é exatamente igual. Este foi o melhor resultado obtido nas previsões. Porém, a grande maioria dos resultados previstos divergiu em relação aos observados. Por exemplo, na magnitude 70 do corpus de resenhas, observou-se os seguintes valores:

Itens do corpus de estudo: 35.741
Formas do corpus de referência: 69.242
Itens do corpus de referência: 2.501.870

Aplicando-os à fórmula, obtém-se:

$$95,890582 + 0,011432 \times (35.741) + 0,009217 \times (69.242) - 0,000105 \times (2.501.870) = 880$$

Na realidade, o número de palavras chave real foi de 914, o que dá uma diferença de 34 palavras a menos na previsão.

O gráfico a seguir mostra a plotagem de todos os valores previstos em contraposição aos valores reais.

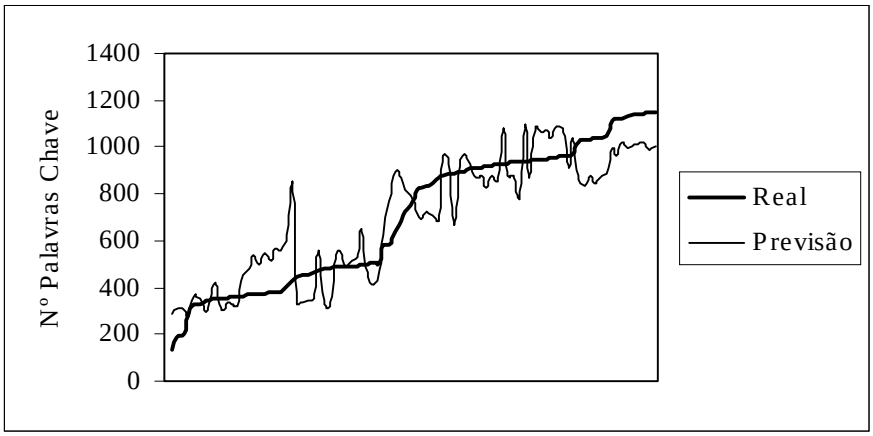


Figura 4: Valores reais e previstos pela equação

Como se pode notar, a linha dos valores previstos segue a tendência de alta dos valores reais.

Em resumo, a tendência de crescimento do total de palavras chave pode ser prevista matematicamente. A fórmula descrita acima serve para se estimar o total de palavras sabendo-se o tamanho do corpus de estudo e do de referência. Ela permite prever-se de modo estatisticamente significante a quantidade de palavras chave resultantes da análise. A partir desta equação, o analista pode, por exemplo, ter uma idéia da quantidade de palavras chave que obteria caso seu corpus de estudo e de referência fossem de dimensões diferentes, e ter condições de refletir sobre o impacto que uma mudança do tamanho do corpus de referência teria em sua pesquisa.

## 7. Conclusão

O tamanho do corpus de referência é um dos cinco elementos que podem influenciar o resultado de uma análise por palavras chave, no tocante à quantidade de palavras chave que podem ser obtidas. Ao contrário da natureza dos textos do corpus de estudo e de referência, os efeitos do tamanho do corpus de referência ainda não podiam ser previstos de antemão. Este estudo propôs-se a verificar a influência da mudança do tamanho de um corpus de referência geral na quantidade de palavras chave de cinco corpora de estudos diferentes. Quatro perguntas de pesquisa foram colocadas (vide seção 4 acima), e a partir delas os achados principais foram os seguintes:

- Variação na parcela de palavras chave do total do corpus de estudo. Os resultados indicaram uma variação grande entre o número de palavras chave obtidas com os dezoito tamanhos de corpora de referência, nos cinco corpora de estudos empregados. Não havia uma relação direta visível entre tamanho do corpus de referência e chavicidade, isto é, não havia uma consistência na parcela de palavras do corpus de estudo que se tornavam chave de acordo com a mudança no tamanho do corpus de referência que se aplicava na análise. Notou-se, contudo, que corpora de referência maiores tendiam a produzir mais palavras chave, mas não progressivamente; isto é, corpora maiores não necessariamente retornavam mais

palavras do que qualquer outro menor.

- Diferença significativa entre os diversos tamanhos dos corpora de referência. Os tamanhos críticos de corpora de referência são 2, 3 e 5 vezes o tamanho do corpus de estudo. Corpora de referência com estas dimensões retornam significativamente mais palavras chave do que corpora de tamanhos menores. Um corpus de referência que é o dobro do tamanho do corpus de estudo retornou cerca de 7% das palavras (do corpus de estudo) como chave; com um corpus de referência que é o triplo, 9%; e com um corpus de referência que é o quintuplo, 14% das palavras do corpus de estudo eram chave.

- Tendência de aumento absoluto do total de palavras chave. Embora a porcentagem de palavras chave de cada corpus de estudo não crescesse constantemente com o aumento do corpus de referência, quanto maior o corpus de estudo, mais palavras chave pareciam ser retornadas.

- Tendência previsível matematicamente. A tendência de alta do total de palavras chave foi prevista através de uma fórmula matemática que incorpora a relação entre o aumento dos corpora de estudo e referência. A fórmula permite estimar-se de antemão, de modo estatisticamente significativo, a quantidade de palavras chave obtidas quando se sabe o tamanho dos corpora de estudo e referência empregados na análise.

Estes achados são potencialmente relevantes a questões relacionadas ao planejamento da pesquisa e ao julgamento da confiabilidade dos resultados. Quanto ao planejamento, o achado mais importante é aquele referente ao valor crítico de cinco vezes o tamanho do corpus de estudo. Segundo este achado, um pesquisador não necessita, necessariamente, coletar ou procurar um corpus de referência maior do que este valor, pois a quantidade de palavras chave a serem obtidas seria igualável a quantidades obtidas com corpora maiores. Em relação à confiabilidade dos resultados a partir do ponto de vista do impacto que uma quantidade de palavras chave diferente teria na interpretação dos resultados, o achado mais importante apresentado aqui é aquele que concerne a possibilidade de previsão do número de palavras chave. Usando-se a fórmula proposta aqui, é possível saber quantas palavras chave haveriam num corpus de estudo quando comparado a um corpus de referência determinado. De posse disto, o pesquisador pode estimar quantas palavras chave obteria caso seu corpus de referência fosse maior. Caso a diferença seja expressiva, é possível questionar-se a confiabilidade dos resultados, já que a existência de mais palavras chave em potencial poderia mudar o teor dos resultados apresentados na análise.

Obviamente, este estudo não responde a várias questões. Uma delas é o efeito do tamanho do corpus de estudo. Não se sabe ainda como corpora de estudos de tamanhos variados se comportam quando comparados a um mesmo corpus de referência. Seria importante saber qual o ganho ou perda de palavras chave conforme o tamanho do corpus de estudo. Com exceção da fórmula de previsão de palavras chave, este efeito não foi levado em conta nos resultados obtidos aqui. Uma outra questão que o estudo não responde é quanto a diferenças no teor das palavras chave obtidas nas várias comparações efetuadas. O estudo apresentado aqui se deteve nos aspectos quantitativos da variação do conjunto de palavras chave, mas seria importante também levar em conta os aspectos qualitativos desta variação. Por exemplo, seria pertinente saber quantas

palavras chave diferentes aparecem como fruto da comparação com os diversos tamanhos de corpus de referência[3]. Finalmente, o fato de os textos extraídos do corpus Brown não serem completos pode ter influenciado os resultados. Essa questão não foi investigada no presente estudo, embora o 'numero de palavras-chave varia bastante em função de extensão do texto' (Mike Scott, comunicação pessoal), o que poderia influenciar as quantidades de palavras-chave obtidas no estudo. Estas e outras questões podem, e devem, ser respondidas em outros estudos, a fim de que se saiba cada vez sobre o procedimento de análise por palavras chave.

O presente estudo vem a contribuir positivamente para preencher, em parte pelo menos, uma lacuna importante no conhecimento relativo à aplicação do procedimento de palavras chave. Desta forma, vem a colaborar para um maior entendimento e aproveitamento do potencial do programa KeyWords de WordSmith Tools, que disponibiliza a um número crescente de pesquisadores uma técnica poderosa e reveladora de análise lexical, genérica, e textual.

**Referências**


BATISTA, M. E. (1998). *E-Mails na troca de informação numa multinacional: o gênero e as escolhas léxico-gramaticais.* Tese de Mestrado, PUC/SP, LAEL, São Paulo.

BONAMIN, M. C. (1999). *Análise organizacional e léxico-gramatical de duas seções de revistas de informática, em inglês.* Tese de MA, LAEL, PUCSP. São Paulo. (http://cogeae.pucsp.br/~pos/mdoutor/lal.htm)

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

CONDE, H. (1999). Aspectos culturais da escrita de alunos de uma escola americana em São Paulo -- Uma perspectiva baseada em corpus. Projeto de mestrado. LAEL, PUCSP.

DUTRA, P. B. (1999). Análise léxico-gramatical baseada em corpus da música pop contemporânea. Projeto de mestrado. LAEL, PUCSP.

FREITAS, A. C. de. (1997). *América mágica, Grã-Bretanha real e Brasil tropical: um estudo lexical de panfletos de hotéis.* Tese de MA, LAEL, PUCSP. São Paulo. (http://cogeae.pucsp.br/~pos/mdoutor/lal.htm)

FUZETTI, H. (1999). A interação oral entre crianças numa escola americana -- Uma abordagem baseada em corpus. Projeto de mestrado. LAEL, PUCSP.

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

LOPES, M. C. (1998). Homepages institucionais em português e suas versões para o inglês: globalização, discurso e cultura. Projeto de mestrado. LAEL, PUCSP.


[3] Seria possível fazer isso apurando-se a consistência das listas.

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

SANTOS, V. B. M. P. dos. (1999). *Padrões interpessoais no gênero de cartas de negociação*. Tese de MA, LAEL, PUCSP. São Paulo. (http://cogeae.pucsp.br/~pos/mdoutor/lal.htm)

SCOTT, M. (1998). *WordSmith Tools Version 3*. Oxford: Oxford University Press.

SILVA, M. S. F. da. (1999). *Análise lexical de folhetos de propagandas de escolas de línguas e as representações de ensino*. Tese de MA, LAEL, PUCSP. São Paulo. (http://cogeae.pucsp.br/~pos/mdoutor/lal.htm)