the ESPecialist, vol. 26, nº 2 (183-204) 2005 ISSN 0102-7077

A INFLUÊNCIA DO TAMANHO DO CORPUS DE REFERÊNCIA NA OBTENÇÃO DE PALAVRAS-CHAVE USANDO O PROGRAMA COMPUTACIONAL WORDSMITH TOOLS*

The Influence of Reference Corpus Size on Wordsmith Tools Keywords Extraction

Tony BERBER SARDINHA[1] (PUCSP)

Abstract

A KeyWords analysis (using WordSmith Tools) enables the discovery of lexical items which reveal the main lexical sets in a text or corpus. Such an analysis requires that a reference corpus be compared to the corpus the researcher intends to describe (the study corpus). This paper presents a mathematical method for finding out the influence of reference corpus size on the number of key words extracted by the program. The results reveal that a reference corpus that is at least five times as large as the study corpus allows for drawing an amount of key words that is statistically equivalent to larger reference corpora, thus suggesting five times (as larger as the study corpora) as the minimum order of magnitude for reference corpora.

Key-words: *WordSmith Tools; KeyWords; Corpus Linguistics; reference corpus size.*

Resumo

Uma análise de palavras-chave por meio de WordSmith Tools permite a descoberta de itens lexicais de maior saliência, que normalmente revelam os principais conjuntos lexicais de um texto ou corpus. Tal análise requer um corpus de referência para ser comparado ao que o pesquisador deseja analisar (o corpus de estudo). Este trabalho apresenta um cálculo para determinação da influência do tamanho do corpus na quantidade de palavras-chave extraída pelo programa

[1] Agradeço ao CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) o apoio mediante a bolsa Produtividade em Pesquisa número 350455/2003-1.

KeyWords. Os resultados indicam que um corpus de referência cujo tamanho seja cinco vezes maior do que o do corpus de estudo permite retirar uma quantidade de palavras-chave estatisticamente equivalente àquela de corpora maiores, o que sugere cinco vezes (maior do que o corpus de estudo) como sendo a ordem de magnitude mínima de um corpus de referência.

Palavras-chave: *WordSmith Tools; palavras-chave; Lingüística de Corpus; tamanho do corpus de referência.*

1. Introdução

O programa de computador WordSmith Tools (Scott, 1998) tem se tornado uma referência para pesquisadores que utilizam programas computacionais para analisar textos. Há vários estudos que se utilizam do programa para a análise de dados (Bargiella-Chiappini e Nickerson, 1999; Batista, 1998; Bonamin, 1999; Conde, 2002; Dutra, 2002; Freitas, 1997; Fuzetti, 2003; Lima-Lopes, 1999; Lopes, 2000; Ramos, 1997; Santos, 1999; Silva, 1999).

Há muitas razões para essa preferência, entre elas o fato de ser um programa desenvolvido para o ambiente Windows, o ambiente operacional dominante no mundo de hoje e a sua disponibilização via Internet. O *software* consiste de um conjunto de diferentes programas com 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].

[2] Devido ao escopo deste trabalho, não é possível explicar em detalhes o funcionamento do programa.

2. Procedimento de extração de palavras-chave

Uma das razões prováveis para o sucesso de WordSmith Tools é 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 (corpus que se pretende descrever) com uma lista de palavras de um corpus de referência (corpus de controle). O resultado do contraste é uma lista de palavras-chave, ou palavras cujas freqüências são estatisticamente diferentes no corpus de estudo e no corpus de referência. As palavras-chave obtidas desse modo têm se mostrado muito úteis na investigação de aspectos textuais importantes, como a temática *(aboutness)*, o estilo e a organização retórica (Batista, 1998; Bonamin, 1999; Conde, 2002; Dutra, 2002; Freitas, 1997; Fuzetti, 2003; Lima-Lopes, 1999; Lopes, 2000; 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 freqüência de palavras. 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. Funciona como termo de comparação para a análise e fornece 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. 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 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 esse item na lista de palavras do corpus de referência;
(3) Se o item constar do corpus de referência, vá para o passo a seguir; se não, passe para o passo 7;
(4) Compare as freqüências por meio 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 essa palavra para uma nova lista, e chame-a de lista de palavras-chave;
(6) Repita esse 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 freqüência 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 análise típica, geralmente se referem à temática desse corpus e, por isso, são intrínsecas a várias características inerentes à textualidade do(s) texto(s) que o compõem. Nesses termos, as palavras-chave são específicas daquele corpus de estudo e, por essa razão, são intimamente ligadas à textualidade (Collins e Scott, 1997).

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 esses elementos. Por exemplo, ao se comparar um corpus de estudo de artigos 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 e um de estudo de gêneros distintos tendem a não excluir tais palavras "genéricas". Por isso, um corpus de referência geral, que inclua vários gêneros, é tido como a escolha não-marcada para estudos de palavras-chave.

Embora a natureza dos corpora de estudo e referência tenda a influenciar os resultados, essas influências podem ser antecipadas. Ou seja, é possível especular que palavras-chave serão encontradas em um determinado corpus de estudo, com base no conhecimento dos textos-fonte (por meio da sua leitura, por exemplo). Do mesmo modo, é possível prever a influência da escolha de um determinado corpus de referência 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 da natureza dos corpora sobre a chavicidade das palavras não significa que se tenha a capacidade de prever com exatidão quais palavras-chave serão obtidas. Portanto, essas previsões de caráter geral acerca do conjunto de palavras-chave poderão 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 freqüências 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 em busca de palavras-chave pode se colocar em relação à extensão dos corpora são:

- Quantas palavras-chave podem ser obtidas de um corpus de estudo x comparado a um corpus de referência y?
- Qual a diferença que se pode esperar em relação ao 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?
- Quando se usa um corpus de referência de extensão x em vez de um de extensão y?

A resposta a essas perguntas é importante porque, em primeiro lugar, se soubermos a influência de um corpus de uma certa dimensão na chavicidade das palavras, é possível planejar 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 desperdicem recursos, coletando-se dados além do que seria teoricamente necessário. O pesquisador poderia então saber o impacto de um corpus de tamanho x sobre os resultados de sua pesquisa e poderia também planejar sua coleta de dados conscientemente. Na prática, algo diferente tem acontecido: 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 adequado.

Em segundo lugar, há a questão da 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? Conhecendo-se a influência do tamanho dos corpora nos resultados, é possível dizer-se qual a quantidade de palavras-chave que teoricamente deixaram de ser incluídas na análise. Em casos extremos, se essas palavras forem de número apreciável, poderiam influenciar, ou até mesmo mudar, os resultados da pesquisa, o que, por sua vez, poderia colocar em xeque os resultados da pesquisa.

4. Questões de pesquisa

A resposta das questões levantadas na seção anterior 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. A influência do aumento do tamanho do corpus de referência (se houver) 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 que pode ser prevista matematicamente?

5. Metodologia

Os corpora de estudo usados nesta investigação são cinco, a saber:

- Corpus de cartas de pedido de emprego, proveniente do Banco de Dados do Projeto DIRECT[3], 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'.

[3] O Projeto DIRECT destina-se à análise da linguagem profissional e do ambiente de trabalho. Para mais informações, vide o site http://lael.pucsp.br/direct.

Os cinco corpora de estudo totalizam cerca de 162 mil palavras, assim distribuídas, conforme indica a Tabela 1:

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	

Tabela 1: Distribuição do total de palavras nos cinco corpora de estudo

As razões da escolha desses 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 desse 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 esses 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 já estavam digitalizados e disponíveis para o pesquisador e, assim, não precisaram ser coletados.

O material para os corpora de referência foi retirado do jornal britânico *The Guardian*. A razão dessa 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 do *WordSmith Tools,* Mike Scott, coloca à disposição uma lista com mais de 95 milhões de palavras, retiradas das edições do mesmo jornal ao longo de quatro anos. Para este estudo, os textos foram coletados aleatoriamente entre o material publicado pelo jornal no ano de 1994.

Devido ao fato de que as perguntas de pesquisa enfocadas aqui centraram-se na questão da influência do tamanho do corpus de referência, foram 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 (um número

de vezes maior do que o tamanho do corpus de estudo). As 18 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 2 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 2: Tamanho dos corpora de referência por ordem de magnitude

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 3: Ajustes do programa KeyWords utilizados na pesquisa

A extração de palavras-chave foi feita com base nesses ajustes, para cada um dos corpora de referência explicitados acima. A seguir, mostramos os resultados das análises.

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 4 abaixo 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 esse corpus com o corpus de referência de magnitude 2x foi 279; portanto, essas 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 4: Total de palavras-chave para cada corpus de referência (Legenda: Magn = ordem de magnitude; p.chave = 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 à medida que crescem os corpora de referência. Para todos os corpora, o total de palavras-chave obtido 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%, referente à literatura de Sci-Fi comparada ao corpus de referência de magnitude 2x. Nessa 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. Nesse 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 produz 50 vezes mais palavras-chave do que um corpus duas vezes maior.

> (2) A influência do aumento do tamanho do corpus de referência (se houver) 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 essa pergunta, é necessário primeiramente observar a distribuição dos totais de palavras-chave de cada corpus para cada magnitude de referência. O gráfico da Figura 1 mostra essa distribuição.

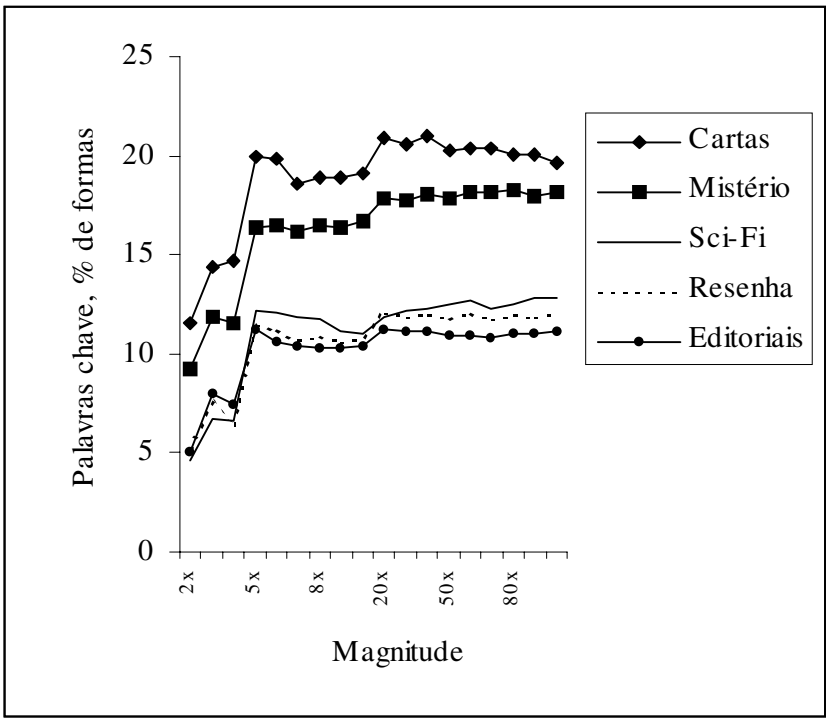

Figura 1: Distribuição de palavras-chave por magnitude de referência do corpus

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 apresentados antes. 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 são significativas, foi necessário submeter os resultados a uma prova estatística chamada Análise de Variância (ANOVA). Os resultados da prova indicaram variação significativa ($F = 267{,}98$ $p<.0001$) 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 a variação existente entre os totais de palavras-chave é significativa entre todas as magnitudes ou somente entre algumas. Para responder a essa questão, foi necessário executar mais uma prova estatística associada à Análise de Variância: o Teste F Múltiplo de REGWF (sigla proveniente de Ryan-Einot-Gabriel-Welsch). Os resultados indicaram três agrupamentos básicos no espectro de palavras-chave: magnitudes 2, 3, e 5. Essas são exatamente as marcas discernidas no exame do gráfico de distribuição de palavras-chave.

O valor crítico, portanto, é 5 (cinco). Um corpus de referência cinco vezes maior do que o de estudo permite extrair um número maior de palavras-chave do que corpora de referência menores. Isso significa que os resultados de uma análise, feita com um corpus de referência menor do que cinco vezes o tamanho do corpus de estudo, poderiam ser alterados, já que corpora menores do que esse tamanho tendem a gerar 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, $3 \approx 4$, $5 \approx 6$, $5 \approx 7$, … $5 \approx 100$). Assim, um corpus de referência que é *quatro* vezes maior do que

o corpus de estudo gera 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, há uma tendência de cerca de 7% do vocabulário do corpus ser palavras-chave em um corpus de estudo, se o seu corpus de referência tiver o dobro do seu tamanho. Com um corpus de referência que é 3 ou 4 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 do corpus de estudo tenderá a se constituir de palavras-chave.

> (3) Há uma tendência observável de variação do número de palavras chave que pode ser prevista matematicamente?

Por fim, para se responder a essa 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 os resultados anteriores, propôs-se a equação da Figura 2 como modelo para a análise de regressão:

$$\text{Total de palavras-chave} = \alpha + \beta_1 \times \text{Itens do corpus de estudo} + \beta_2 \times \text{Formas do corpus de referência} + \beta_3 \times \text{Itens do corpus de referência}$$[4]

Figura 2: Equação para análise de regressão

As variáveis da equação (em letras gregas) precisam ser substituídas pelos valores da 'estimativa do parâmetro' apresentados na tabela a seguir:

[4] Embora a fórmula empregue sinais de adição entre os termos, o resultado da equação não é necessariamente maior que 95,8 (a) 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.

Variável	Estimativa do parâmetro
α Interseção	95,890582
β_1 Itens do corpus de estudo	0,011432
β_2 Formas do corpus de referência	0,009217
β_3 Itens do corpus de referência	-0,000105

Tabela 5: Parâmetros da Análise de Regressão

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)

Figura 3: Modelo para previsão do total de palavras-chave

Com a equação acima, 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 × (11.761) + 0,009217 × (5.543) – 0,000105 × (23.522) = 279

Ou seja, a fórmula prediz que haverá 279 palavras-chave para um corpus com essas dimensões. O total observado de palavras-chave é exatamente igual e esse é o melhor resultado obtido nas previsões. Entretanto, a grande maioria dos resultados previstos divergiu em relação aos observados. Por exemplo, na magnitude 70 do corpus de resenhas, observaram-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:

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

O número efetivo de palavras-chave foi de 914, o que dá uma diferença de 34 palavras a menos na previsão.

O gráfico da Figura 4 mostra a plotagem de todos os valores previstos em contraposição aos valores reais.

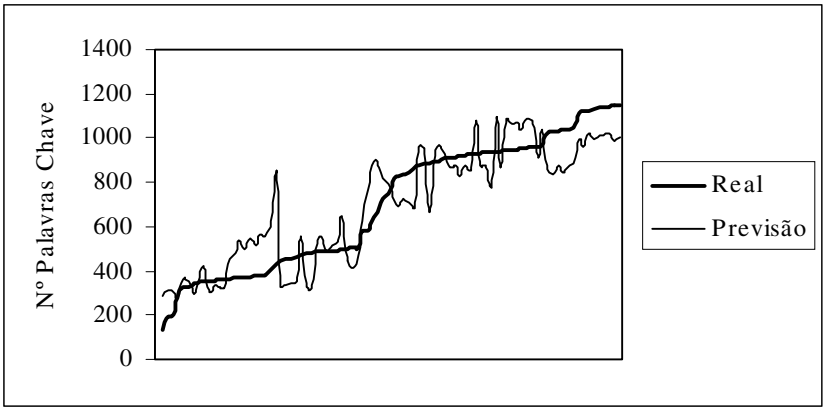

Figura 4: Valores reais e previstos pela equação

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 anteriormente, na Figura 2, serve para se estimar o total de palavras, sabendo-se o tamanho do corpus de estudo e do de referência. Como a equação utilizada trabalha com valores significativos, ela permite prever de modo estatisticamente significante a quantidade de palavras-chave resultantes da análise. A partir dessa equação, o analista pode, por exemplo, ter uma idéia da quantidade de palavras-chave que obteria caso os corpora de estudo e de referência fossem de dimensões diferentes, e pode também 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. Três 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 18 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 geravam necessariamente 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 essas dimensões geram significativamente mais palavras-chave do que corpora de tamanhos menores. Um corpus de referência que é o dobro do tamanho do corpus de estudo gera 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 quíntuplo, 14% das palavras do corpus de estudo são chave.

- Tendência de aumento do total de palavras-chave, 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 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.

Esses achados são potencialmente relevantes 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 esse achado, um pesquisador não necessita, necessariamente, coletar ou procurar um corpus de referência maior do que esse 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 haveria

num corpus de estudo quando comparado a um corpus de referência determinado. De posse disso, o pesquisador pode estimar quantas palavras-chave obteria caso seu corpus de referência fosse maior. Se a diferença for 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, esse 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 dessa variação. Por exemplo, seria pertinente saber quantas palavras-chave diferentes aparecem como fruto da comparação com os diversos tamanhos de corpus de referência[5]. 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 'número de palavras-chave varie bastante em função de extensão do texto' (Mike Scott, comunicação pessoal), o que poderia influenciar as quantidades de palavras-chave obtidas no estudo. Essas e outras questões podem e devem ser respondidas em outros estudos, a fim de que se conheça melhor o procedimento de análise por palavras-chave.

O presente estudo tenta preencher, ao menos em parte, uma lacuna importante no conhecimento relativo à aplicação do procedimento de palavras-chave. Dessa 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

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

pesquisadores, uma técnica poderosa e reveladora de análise lexical, genérica e textual.

Agradecimentos

Agradeço aos pareceristas anônimos a leitura cuidadosa deste trabalho e os comentários pertinentes.

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

Referências Bibliográficas

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. (http://lael.pucsp.br/lael)

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.) 1999 *The languages of business - an international perspective*. Edinburgh University Press.

BARGIELA-CHIAPPINI, F. & NICKERSON, C. 1999 Business writing as social action. IN: F. BARGIELA-CHIAPINI & C. NICKERSON (orgs.) *Genres, media and discourse.* Longman.

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, LAEL, PUC/SP. (http://lael.pucsp.br/lael)

DUTRA, P.B. 2002 *Explorando a Lingüística de Corpus e letras de música na produção de atividades pedagógicas. Dissertação de Mestrado, LAEL, PUC/SP. (http://lael.pucsp.br/lael)*

FREITAS, A.C. de. 1997 *América Mágica, Grã-Bretanha Real e Brasil Tropical: um estudo lexical de panfletos de hotéis*. Tese de Doutorado, LAEL, PUCSP. (http://lael.pucsp.br/lael)

FUZETTI, H. 2003 *Padrões léxico-gramaticais na linguagem de crianças em uma escola americana no Brasil*. Dissertação de Mestrado, LAEL, PUC/SP. (http://lael.pucsp.br/lael)

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, LAEL, PUC/SP. (http://lael.pucsp.br/lael)

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

SANTOS, V.B.M.P. dos. 1999 *Padrões interpessoais no gênero de cartas de negociação*. Dissertação de Mestrado, LAEL, PUC/SP. (http://lael.pucsp.br/lael)

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

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 in both the Postgraduate Program in Applied Linguistics and the Linguistics Department, Catholic University of São Paulo, and a researcher with the Brazilian National Research Council (CNPq). His main interest is in the area of Corpus Linguistic. tony4@uol.com.br