the ESP, São Paulo, vol. 23, nº 2 103-122

TAMANHO DE CORPUS
Corpus size

Tony BERBER SARDINHA
(LAEL, PUCSP)

Abstract

This paper addresses the question of determining the typical size of corpora used in research reported in Corpus Linguistics conferences and meetings. By surveying the corpora actually used by corpus linguistics in their research projects over a period of several years, it was possible to calculate the range of variation in corpus size in the field and estimate levels of acceptability held by the community. This approach contrasts with subjective views put forth by Corpus Linguistics practitioners on the issue of corpus size.

Key-words: *Corpus Linguistics; corpus size; large, average and small corpora.*

Resumo

O trabalho enfoca a questão da determinação do tamanho típico dos corpora usados na pesquisa relatada em conferências e encontros de Lingüística de Corpus. A partir de um levantamento dos corpora efetivamente usados em projetos de pesquisa num período de alguns anos, foi possível calcular a extensão da variação do tamanho dos corpora empregados na área e estimar graus de aceitabilidade mantidos pela comunidade. Essa abordagem contrasta com opiniões subjetivas a respeito da questão do tamanho de corpus, expressas por praticantes da Lingüística de Corpus.

Palavras-chave: *Lingüística de Corpus; tamanho de corpus; corpora grandes, médios e pequenos.*

1. Introdução

Um corpus computadorizado é um objeto valioso para a investigação da linguagem (Kennedy, 1998; McEnery e Wilson, 1996). Através dele pode ser estimada a ocorrência de uma ampla gama de traços lingüísticos, incluindo morfológicos, morfossintáticos, sintáticos, semânticos, discursivos, etc. O emprego de um corpus na pesquisa lingüística traz vários benefícios, entre eles a possibilidade da explicação de diferenças de uso de palavras, expressões, formas gramaticais e outros traços por meio da probabilidade de ocorrência em contextos específicos (Biber et al., 1998), a possibilidade de descoberta de fatos novos não disponíveis pela intuição ou eliciação (Sinclair, 1991) e a descrição objetiva da linguagem enquanto um sistema probabilístico (Halliday, 1992).

Uma definição de corpus que engloba vários elementos importantes é oferecida por Sánchez (1995: 8-9):

> *Um conjunto de dados lingüísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos em amplitude e profundidade, de maneira que sejam representativos da totalidade do uso lingüístico ou de algum de seus âmbitos, dispostos de tal modo que possam ser processados por computador, com a finalidade de propiciar resultados vários e úteis para a descrição e análise.*

Um elemento central dessa definição é que o corpus deva ser *representativo*. Não há abordagens formais para se estabelecer a representatividade de um corpus (Sinclair, 1996), isto é, não há nenhuma fórmula matemática amplamente aceita que informe a quantidade ou distribuição de palavras ou textos que um corpus deva ter para ser representativo. Por isso, um preceito corrente entre os praticantes da área é que para ser representativo um corpus deva ser o maior possível:

> *O corpus deve ser o maior possível dentro do que pode ser atingido com a tecnologia da época*. (Sinclair, 1996:3)

Essa recomendação é tida como uma 'salvaguarda' (Sinclair, 1991), ou seja, um paliativo ou medida de emergência que visa a oferecer alguma segurança ao pesquisador que o seu corpus seja representativo *em relação a outro*. Em outras palavras, um corpus grande não é representativo *em absoluto*, mas é mais representativo do que um outro menor. O perigo dessa salvaguarda é obviamente a circularidade que ela gera.

Mas se um corpus grande é mais representativo, quão grande seria esse corpus? Em outras palavras, o que seria um corpus grande? E um pequeno? Até o momento não há respostas objetivas a essas questões, embora se faça menção constante de corpora[1] grandes e pequenos na literatura.

2. Abordagens acerca da questão da representatividade de corpus

Até o momento, pode-se distinguir dois tipos de abordagem para a questão da representatividade:

- Estatística: fundamenta-se na aplicação de teorias estatísticas. Pode-se subdividir a abordagem estatística em três perspectivas: (1) Interna: dado um corpus preexistente que serve como amostra maior, qual o tamanho mínimo de uma amostra que mantém estáveis as características dessa amostra maior? Essa é a perspectiva seguida por Biber (1990, 1993). (2) Externa: dada uma fonte externa de referência cuja dimensão é conhecida, qual o tamanho do corpus necessário para representar majoritariamente essa fonte? Essa vertente tem sido discutida pela comunidade de lingüistas do corpus (Berber Sardinha, 1998). (3) Relativa: quanto se perderia se o corpus fosse de um tamanho x? Dados os meus recursos existentes, quais parâmetros posso utilizar para abalizar minha decisão relativa ao tamanho de corpus que posso compi-

1 Por ser mais usual na literatura, será empregada a forma 'corpora' para denotar o plural de corpus, em preferência a 'corpuses'.

lar? Uma proposta segundo essa perspectiva ainda não foi formalizada, mas está presente, por exemplo, em Sanchez e Cantos (1997a, b), os quais estimam matematicamente a quantidade do vocabulário presente em corpora de diversos tamanhos hipotéticos, e em Yang e Song (1998), os quais fazem uma previsão da quantidade de dados necessários para incluir certas características gramaticais.

- Impressionística: baseia-se em constatações derivadas da prática da criação e exploração de corpora, em geral feitas por autoridades da área. Por exemplo, Aston (1997) menciona patamares que caracterizariam um corpus pequeno (20 a 200 mil palavras) e um grande (100 milhões ou mais). Leech (1991) fala de 1 milhão de palavras como a taxa usual (*going rate*), sugerindo o que seja o patamar mínimo. Outros são mais vagos, como Sinclair (1996), mencionado acima, que postula que o corpus deva ser tão grande quanto a tecnologia permitir para a época, deixando-se subentender que a extensão de um corpus deva variar de acordo com o padrão corrente nos grandes centros de pesquisa, os quais possuem equipamentos de última geração.

No que se segue serão discutidos, detalhadamente, exemplos centrais de cada uma dessas abordagens.

2.1. Abordagem estatística

Há algumas propostas estatísticas para o estabelecimento da representatividade de um corpus (Biber, 1990, 1993; De Haan, 1992; Yang e Song, 1998). Essas abordagens partem do ponto de que um corpus é uma amostra da língua como um todo. O problema nesse contexto seria, portanto, o de aferir o tamanho mínimo dessa amostra de tal modo que ela represente a população de que se origina, ou seja, a língua. Como não se sabe o tamanho de uma língua como o português, por exemplo, não se possuem meios inequívocos para se testar a representatividade da amostra. Algumas propostas para conjuntos de

dados genéricos, baseadas em critérios puramente estatísticos, existem, tais como Sibson (1972), que sugere 60 unidades como uma amostra mínima conveniente. Daí se poderia extrapolar e considerar uma amostra com 60 textos como representativa. Mas como traduzir essa marca para a quantidade de palavras, por exemplo?

Uma das propostas mais claras a respeito de amostragem representativa de amostras lingüísticas é oferecida por Biber (1990, 1993). Em dois estudos, ele enfoca especificamente a questão do número de textos mínimos para se constituir uma amostra representativa. No primeiro, Biber (1990) comparou as freqüências de vários traços lingüísticos em amostras de vários tamanhos, e computou correlações entre as amostras pequenas e o corpus. As correlações indicaram que as amostras de 10 textos mantiveram as características dos traços em questão conforme aparecem no corpus. Portanto, uma amostra de 10 textos seria suficiente para representar o corpus. Entretanto, esses resultados são inexatos porque se baseiam em traços freqüentes, o que tende a diminuir o tamanho mínimo necessário para se atingir a representatividade.

No outro estudo, Biber (1993) enfocou uma gama de traços mais e menos freqüentes. Os resultados indicaram que o número de textos mínimo varia de acordo com o traço que se toma como base: traços mais freqüentes exigem uma quantidade de textos menor e vice-versa, ou seja, traços mais raros necessitam de um número maior de textos. Por exemplo, uma característica freqüente seriam os substantivos, já que qualquer texto tem, em geral, muitos substantivos; uma amostra de textos para representar os substantivos deveria ter, no mínimo, 60 textos (por coincidência ou não, é a mesma cifra proposta por Sibson, 1972). Isto porque uma amostra de 60 textos guardaria as características do corpus (a amostra maior). Mas uma amostra representativa de uma característica infreqüente, como orações condicionais, precisaria ter no mínimo 1190 textos. Há, portanto, uma grande variação entre os números propostos por Biber: 10 ou 60, para traços freqüentes, e mais de 1100 para infreqüentes.

2.2. Abordagem impressionística

Em seu estudo, no entanto, Aston (1997) sugere alguns números: um corpus pequeno teria entre 20 e 200 mil palavras, e um grande, 100 milhões de palavras ou mais. Assume-se que os que tenham entre 200 mil e 100 milhões sejam 'médios'. E os que tenham menos de 20 mil palavras sejam 'minúsculos', ou pior, não-corpora. Note a circularidade dessa argumentação: os corpora maiores são aqueles que têm mais palavras que os menores, e vice-versa. A circularidade deve-se ao fato de não haver normas (*standards*) aceitas pela comunidade. Mas a sugestão de Aston (1997) toca num ponto importante, qual seja, a questão da aceitabilidade pela comunidade: pode-se interpretar sua estimativa como significando que os corpora acima de 20 mil palavras são geralmente aceitos pela comunidade e, nesse caso, se transforma em um padrão de tamanho mínimo aceitável.

Há que se enfatizar que o tamanho mínimo dos corpora depende do tipo de pesquisa. Aston (1997) lembra que, para fins de aplicação na sala-de-aula ou de preparação de materiais de ensino de línguas, corpora de 100 milhões de palavras são impraticáveis. Ele lista, ainda, alguns corpora utilizados em pesquisas voltadas ao ensino e suas dimensões:

Corpus	**Palavras**
Revista Byte	1.000.000
Revista New Scientist	760.000
Palestras acadêmicas	155.000
Periódicos científicos e técnicos	114.000
Corpus Longman de aprendizes	55.000
Geologia	34.000
Economia	21.000
Mala direta	16.000
Filosofia	7.000

Tabela 1: Dimensão de diversos corpora segundo Aston (1997)

Notam-se três aspectos principais na Tabela 1. Primeiramente, há uma variação grande de tamanho entre os corpora. A diferença entre o menor e o maior corpus é de mais de 100 vezes. Em segundo lugar, os corpora maiores são aqueles compostos por textos provindos da imprensa. Este é o efeito da disponibilidade, ou seja, há mais chances de um corpus ser maior se os textos forem fáceis de coletar. E em terceiro lugar, pelo menos dois estudos utilizaram corpora abaixo daquele patamar de aceitabilidade de 20 mil palavras, proposto por Aston (1997). Isso significa que o patamar nada mais é do que a observação de uma tendência, e não uma norma empregada pela comunidade.

O problema com essas estimativas é que são baseadas em intuição e julgamento pessoal. Em outras palavras, elas não utilizam a metodologia da Lingüística de Corpus ao se furtarem do levantamento de dados reais acerca da real situação da utilização de corpora pelos lingüistas de corpus. Esses julgamentos possuem credibilidade na medida em que provêm de um membro respeitado na comunidade, mas não se fundamentam em evidências.

2.3. Uma nova abordagem: histórica

Para se saber o que seria um corpus grande, ou um corpus pequeno, seria necessário, então, adotar-se uma terceira via, partindo-se de uma perspectiva diferente: a *observação dos corpora que estão de fato sendo usados pela comunidade de pesquisadores* através de um levantamento criterioso. Tal abordagem, que chamaremos de *histórica*, busca identificar os patamares aceitos em um período de tempo pela comunidade. Em sua essência, a abordagem propõe-se a seguir o mesmo critério norteador pelo qual a própria Lingüística de Corpus se pauta: a observação de dados de uso. Só que, nesse caso, em vez de se observar dados de linguagem em um corpus, busca-se delinear o que seria um corpus aceitável perante a comunidade de usuários que perfaz a própria Lingüística de Corpus. O conceito central nessa abordagem é, portanto, o de *aceitabilidade*, ou seja, busca-se *saber* o que é aceitável, e não *prescrever* o que seria ideal.

3. Dados do levantamento

Para se colocar em prática a perspectiva histórica, fez-se um levantamento que consistiu em tomar-se nota do tamanho, em palavras, de todos os corpora utilizados nos trabalhos apresentados nas conferências mais importantes de Lingüística de Corpus dos últimos quatro anos (de 1995 a 1998). Esse período de tempo representa, até o momento, o final dos anos 90. As conferências foram as seguintes:

Evento	Ano	Cidade	Referência
ICAME 16	1995	Toronto	Percy et al., 1996
ICAME 17	1996	Estocolmo	Ljung, 1997
ICAME 18	1997	Chester	Renouf, 1998
PALC'97	1997	Lodz	Lewandowska-Tomaszczyk e Melia, 1997
TALC'98	1998	Oxford	Teaching and Language Corpora 98, 1998

Tabela 2: Conferências participantes do levantamento

Dois esclarecimentos são necessários acerca do procedimento de coleta. Primeiramente, por 'trabalhos apresentados nas conferências' entende-se aqueles que foram publicados nos anais. E em segundo lugar, por 'corpora utilizados' entende-se 'corpora empregado para pesquisa ou compilação', isto é, um corpus que tenha servido de base para investigação lingüística ou que tenha sido criado e apresentado durante a conferência. Não entraram no levantamento corpora que ainda estivessem em fase de planejamento, ou seja, que na verdade não existissem ainda. É muito importante frisar que muitos estudos se utilizaram de *combinações de corpora*; por isso, quando se falar em corpora de um determinado valor, isso não significa que seja um corpus apenas.

Houve vários problemas durante a coleta dos dados, ligados à falta de consistência e clareza na descrição dos corpora. Isso em si já é merecedor de nota, já que se esperava maior rigor no relato do objeto central da Lingüística de Corpus dentro da própria disciplina. O pri-

meiro problema é que muitos estudos simplesmente não informavam o leitor acerca do tamanho do corpus (vide as estatísticas abaixo). Algumas vezes, mencionavam apenas o nome do corpus. Nesses casos, quando o corpus era conhecido (por exemplo, London-Lund), o número de palavras poderia ser inferido. Em outros casos, entretanto, isso não foi possível porque o corpus era desconhecido (por exemplo, Chemnitz English Translation Corpus), ou porque simplesmente não se nomeava o corpus (p. ex. 'in my corpus, …'). O segundo problema é que várias vezes os autores não mediam a extensão do corpus em palavras (por exemplo, 200 sentenças do LOB, 5 textos, 15 horas de gravação).

Por fim, muitos autores apresentavam o número de palavras das várias partes do corpus ou dos corpora individualmente, mas não o total. Em alguns desses casos, as informações eram de unidades diferentes, por exemplo, '300 textos de 2000 palavras cada'.

A Tabela 3 a seguir traz as contagens referentes aos trabalhos publicados que apresentaram informações acerca da extensão dos corpora em número de palavras:

Conferência	**Trabalhos publicados**	**Trabalhos com informação de número de palavras**	
ICAME 1995	20	15	75%
ICAME 1996	24	17	71%
ICAME 1997	22	17	77%
PALC 1997	39	16	41%
TALC 1998	46	19	41%
Total	151	84	56%

Tabela 3: Trabalhos com informação acerca de número de palavras

Dos 151 trabalhos que apareceram publicados, apenas cerca da metade (56%) deram informações relativas à extensão dos corpora em número de palavras. Há, portanto, na comunidade de Lingüística de

Corpus, um descompasso entre a teoria e a prática. Nos textos principais da área, há uma ênfase constante na importância do tamanho do corpus para a sua representatividade. Já nas conferências anuais, os principais veículos da área, há uma falta de consciência da importância de se medir e relatar a extensão do corpus. Entre os que o fazem, contudo, há uma falta de padronização em relação à unidade de medida básica do corpus. Aqui também há um desencontro entre a teoria e a prática. Os textos de referência reportam rotineiramente o número de palavras dos corpora; já nas conferências e seus anais, há uma variação considerável na unidade de medida (palavras, sentenças, textos, horas) e normalmente essas unidades não são mutuamente conversíveis.

Percebe-se, contudo, que há uma diferença entre os tipos de conferência. As conferências ICAME possuem uma taxa maior de relato, em torno de 75%, enquanto as conferências PALC e TALC têm em média 41%. Isso significa que, enquanto nas conferências ICAME cerca de 3 dentre cada 4 trabalhos expuseram a extensão numérica dos corpora, nas outras conferências apenas 2 em cada 5 trabalhos o fizeram.

Como se pode explicar essa diferença? A primeira explicação seria em relação à natureza da conferência. As conferências TALC e PALC são voltadas ao ensino, e portanto congregam pesquisadores que fazem pesquisas mais informais, nas quais a aplicação do corpus é mais essencial do que a contagem de palavras. Já nas conferências ICAME, o foco das pesquisas é na descrição da língua, e daí surge naturalmente uma necessidade maior de descrever qual variedade da língua está sendo descrita e qual corpus está sendo usado como base da descrição. O detalhamento da extensão do corpus é peça central na descrição do corpus de estudo.

A segunda explicação seria relativa à natureza da seleção. As conferências ICAME não são abertas; os participantes são convidados. Isso faz com que sejam convidados membros de instituições conhecidas e que em geral possuem maior tradição em Lingüística de Corpus. De certo modo, isso contribui para que os autores tenham consciência da importância da contagem de palavras na descrição do corpus. Além disso, os grupos de pesquisa que freqüentam as conferências ICAME

têm mais acesso aos corpora mais tradicionais (Brown, LOB, London-Lund)[2], cujas contagens de palavras são conhecidas.

Finalmente, há uma certa influência da seleção para publicação nos anais. Os trabalhos para os anais do ICAME são submetidos a um processo de seleção maior do que para o TALC e PALC. A conferência TALC lança os anais durante a conferência, e, portanto, inclui relatos preliminares dos trabalhos apresentados e publica, em alguns casos, apenas um resumo do trabalho. No resumo, a omissão do relato do tamanho do corpus é menos séria do que no corpo da versão completa.

4. Tendências anuais

Nesta seção, serão apresentados os resultados do levantamento relativo a tendências observadas ano a ano, primeiramente no que concerne aos valores máximo e mínimo por ano; depois, aos valores médios.

4.1. Os corpora maiores e menores

Os valores relativos ao tamanho de cada corpus por conferência aparecem na Tabela 4 a seguir.

Segundo a Tabela 4, observa-se que o valor máximo aumentou, de 200 milhões em 1995 para 480 milhões em 1998, mais que o dobro. O valor mínimo, entretanto, não cresceu: de 50 mil em 1995, diminuiu para 17 mil em 1998. Em outras palavras, os corpora grandes tornaram-se ainda maiores, mas os menores não acompanharam essa tendência. Isso significa que há uma maior flexibilidade dentro da Lingüística de Corpus acerca do tamanho aceitável de um corpus. Note-se que o menor valor dos quatro anos é 2800 palavras, um corpus extremamente reduzido e que, no entanto, foi aceito pela comunidade como válido.

2 Vale lembrar que o ICAME é o distribuidor desses corpora em CD-ROM.

Tema	Descrição			Ensino	
Conferência	**ICAME 16**	**ICAME 17**	**ICAME 18**	**PALC**	**TALC 98**
Ano	**1995**	**1996**	**1997**	**1997**	**1998**
PALAVRAS	200.000.000	200.000.000	280.000.000	320.000.000	480.000.000
	2.280.000	4.000.000	280.000.000	100.000.000	423.000.000
	2.000.000	4.000.000	100.000.000	100.000.000	331.000.000
	1.800.000	1.200.000	54.000.000	20.000.000	259.000.000
	1.000.000	1.000.000	11.700.000	10.000.000	105.000.000
	1.000.000	1.000.000	3.300.000	3.100.000	13.000.000
	500.000	500.000	2.500.000	2.500.000	7.000.000
	120.000	270.000	2.500.000	645.000	2.400.000
	80.000	217.000	860.000	130.000	700.000
	80.000	211.000	771.000	95.000	607.000
	80.000	156.000	500.000	43.000	600.000
	56.000	154.000	400.000	40.000	400.000
	52.000	150.000	300.000	20.000	375.000
	50.000	111.000	250.000	16.000	300.000
	50.000	40.000	53.000	10.000	200.000
		12.000	50.000	10.000	120.000
		2.800	16.500		115.000
					20.000
					17.000
Máximo	200.000.000	200.000.000	280.000.000	320.000.000	480.000.000
Mínimo	50.000	2.800	16.500	10.000	17.000
Média	13.943.200	12.530.812	43.364.735	34.788.063	85.466.000
Mediana	120.000	217.000	860.000	387.500	607.000

Tabela 4: Valores relativos aos corpora pesquisados

O aumento do tamanho máximo deve-se, em boa parte, à maior utilização de textos de jornais em CD-ROM. Outro fator importante é o papel do aumento do emprego de corpora com 100 milhões de palavras, que se tornaram mais freqüentes, passando de um em 1995 para cinco em 1998. O valor de 100 milhões de palavras parece crítico na mudança da extensão dos corpora ao longo desses 4 anos, e isso se deve em parte ao BNC (British National Corpus), o qual tem essa quantidade de palavras, e cuja maior disponibilidade tem alterado o perfil dos corpora empregados pela comunidade.

A maior variação entre os tamanhos dos corpora mencionada acima reflete a realidade do mundo da pesquisa em Lingüística de Corpus; nem todos os pesquisadores possuem recursos para dispor de corpora de grandes proporções, como o BNC ou jornais em CD-ROM. Além disso, vale lembrar que, além das financeiras, há restrições burocráticas de acesso: o BNC, até pouco tempo, era comercializado somente dentro da Comunidade Européia, e o Bank of English não é vendido, permitindo o acesso somente a pesquisadores ligados ao COBUILD, ou que tenham assinatura do seu serviço *on-line*. Outras limitações impedem o acesso de maiores usuários a corpora grandes, como a necessidade de equipamentos com grande capacidade de armazenamento, e software adequado.

4.2. Os corpora médios

Há dois tipos de medida que se pode usar no cálculo da média de tamanho dos corpora investigados. A primeira, a média aritmética, mostra que o tamanho médio em palavras vem crescendo, de cerca de 14 milhões em 1995 para mais de 85 milhões em 1998. Essa média é altamente influenciada por valores extremos na distribuição, o que acontece nesse caso. Veja, por exemplo, que em 1995 o maior corpus tinha 200 milhões de palavras, e o segundo menor, apenas cerca de 2 milhões, quase cem vezes a menos. A partir de 1997, esses abismos entre os maiores valores vêm diminuindo, mas mesmo assim há pontos de grande discrepância da distribuição, como por exemplo entre o quarto e o quinto valor no ICAME de 1997 (de 54 milhões para cerca de 12 milhões) e 1998 (de 105 milhões para 13 milhões).

Por isso tudo, a mediana, uma outra estatística de tendência central, parece mais apropriada. A mediana descreve o ponto central de uma distribuição, dividindo-a em duas metades de 50%. Por exemplo, considere uma seqüência ordenada de dados como a seguinte: '1, 5, 5, 5, 100'. Nela, a mediana é o terceiro elemento, pois é o ponto central, deixando dois elementos à esquerda e dois à direita dele. A mediana é 5, portanto. Já a média dessa mesma seqüência é 23,2 (=116/5). Essa média não representa nenhum número do conjunto e é claramente afetada por 100. A mediana, ao contrário, é menos afetada por valores extremos.

O valor médio segundo a mediana aumentou também, mas não linearmente, através das cinco conferências. Ela subiu, sim, dentro de cada tipo de conferência. A mediana das conferências ICAME, dedicadas à descrição da linguagem, subiu de 120 mil para 860 mil, e no PALC e no TALC, voltados para a aplicação de corpora no ensino, passou de mais de 387 mil para 607 mil. Isso pode se dever ao fato de na conferência ICAME os corpora serem todos de inglês (embora haja estudos comparativos que utilizam corpora de outras línguas para contraste), mais abundantes e fáceis de coletar, enquanto nas outras conferências os corpora podem ser de qualquer língua, mais raros e difíceis de coletar.

4.3. Resumo

Em resumo, há uma tendência de aumento do tamanho dos corpora: os corpora tornaram-se maiores, e tendem a ter em média mais palavras. O valor mínimo aceitável, porém, não tem subido. Isto causa uma discrepância entre os corpora em utilização pela comunidade. De um lado há corpora (ou combinações de corpora) gigantescos, com quase meio bilhão de palavras, e por outro há outros que não passam de alguns milhares de palavras.

5. Tendência geral

Nesta seção, os resultados do levantamento serão interpretados em sua totalidade, sem a divisão anual feita acima. Para tanto, os valo-

res das cinco conferências foram agrupados em um conjunto único de dados com 84 casos (trabalhos apresentados). Isto auxiliará na busca de respostas a duas questões: 'Levando-se em conta os quatro anos pesquisados, qual o tamanho de um corpus médio?', e 'como eu classifico a extensão do meu corpus: pequeno, médio ou grande?'

5.1. O que é um corpus médio

A Tabela 5 a seguir traz as estatísticas descritivas do levantamento. O tamanho médio aritmético dos corpora de 1995 a 1998 é de aproximadamente 40 milhões de palavras. Contudo, conforme discutido acima, esse valor é afetado pela distribuição desigual dos valores, sugerido pela diferença de 171 mil vezes entre os tamanhos máximo e mínimo absolutos (480 milhões e 2.800). A mediana é de exatas 500 mil palavras.

Média	39.759.944
Mediana	500.000
Máximo	480.000.000
Mínimo	2.800
Diferença	171.429
50 mil ou mais	86%
100 mil ou mais	74%
1 milhão ou mais	42%
10 milhões ou mais	21%
100 milhões ou mais	15%
300 milhões ou mais	5%

Tabela 5: Estatísticas descritivas do levantamento

A resposta é, portanto, que um corpus médio na Lingüística de Corpus, conforme revelado por esse levantamento, é de 500 mil pala-

vras. Essa é a mediana, a qual, conforme dito antes, é a melhor medida de valor médio para esse caso. Então, quanto mais próximo de 500 mil, mais mediano será o corpus. Ou, dito de outro modo, se possuir mais de 500 mil palavras, estará entre 'os maiores'; menos do que isto, figurará entre 'os menores'.

5.2. Quando um corpus é grande, médio ou pequeno

Com esses comentários, pode-se passar para a segunda questão: como se classificam os corpora em pequenos, médios e grandes? A tabela acima ajuda a começar a responder essa questão.

Olhando a Tabela 5, percebe-se que 86% dos corpora pesquisados possuem tamanho superior a 50 mil palavras. Portanto, se um corpus possuir 50 mil palavras ou mais, estará entre os 86% maiores. Se possuir menos que 50 mil, estará entre os 14% menores. Se tiver 100 mil palavras, não haverá grande mudança: será um dos 74% maiores (se tiver menos do que 100 mil palavras, estará entre os 26% menores). Se possuir 1 milhão de palavras, já há uma mudança de 'lado', uma diferença perceptível, passando a estar entre os 42% maiores. Acima de 10 milhões, o seu corpus passa a estar entre os 21% maiores, ou seja, apenas 1 dentre 5 dos estudos pesquisados terá utilizado um corpus igual ou maior. Acima disso, pode-se dizer que o seu corpus está entre a 'elite' da área. Com 100 milhões, só há 15% iguais ou menores (1 em 7), e com 300 milhões, só 5% (1 em 20).

Para se ter maior precisão na classificação de corpora de diversos tamanhos, é necessário uma escala gradual. Para tanto, deve-se tomar a listagem completa dos valores e reparti-la em quantidades iguais relativas à classificação qualitativa desejada.

Assim, pode-se dividir a distribuição dos corpora em cinco faixas: pequeno, pequeno-médio, médio, médio-grande, e grande, cada uma ocupando 20% (1/5) da distribuição.

Duas divisões foram inexatas. No caso dos 20%, o ponto de corte cairia no valor 53 mil, que é ladeado por outros valores similares:

52 e 56 mil. Por isso, achou-se por bem avançar-se a linha divisória para os 22,6%, quando se passava de 56 para 80 mil palavras, uma fronteira mais legítima. E em relação aos 60%, o ponto de corte se situava no meio de uma seqüência repetida de 1 milhão, portanto não havia nenhuma fronteira ali. O ponto mais próximo era 58,3%, por isso colocou-se a fronteira entre 860 mil e 1 milhão. Desse modo, a classificação fica a seguinte:

Tamanho em palavras	**Classificação**
Menos de 80 mil	Pequeno
80 a 250 mil	Pequeno-médio
250 mil a 1 milhão	Médio
1 milhão a 10 milhões	Médio-grande
10 milhões ou mais	Grande

Tabela 6: Classificação relativa do tamanho de corpus

Graficamente, a escala seria esta:

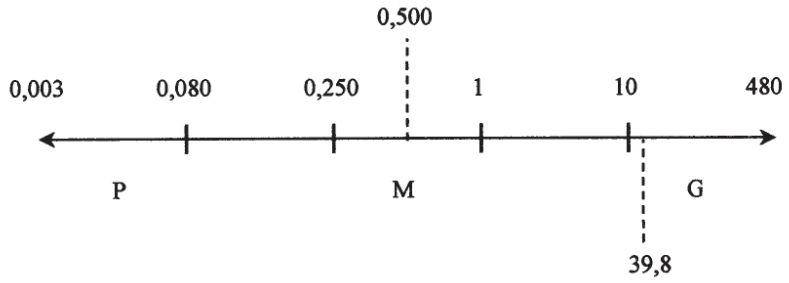

Figura 1. Escala de tamanho relativo de corpora. Os números se referem a quantias em milhões; assim, '0,500' representa meio milhão; as letras (P, M e G) referem-se a 'Pequeno', 'Médio' e Grande'. O número sobre a linha tracejada superior indica a mediana e, sob a inferior, a média aritmética.

6. Comentários finais

Em suma, propôs-se aqui uma nova perspectiva para a questão da representatividade de corpora e, mais especificamente, para a questão do que seria um corpus grande, médio e pequeno. Foi relatado um levantamento dos corpora empregados pela comunidade de lingüistas de corpus segundo estudos apresentados nas principais conferências de Lingüística de Corpus num período de quatro anos recentes. Foi proposta, como resultado, uma escala de tamanhos relativos de corpora. Segundo essa escala, um corpus pequeno seria aquele que possui menos de 80 mil palavras, um médio, menos de 1 milhão, e um grande, 10 milhões ou mais. Mais especificamente, um corpus 'médio' seria aquele que possuísse 500 mil palavras.

Essa proposta de classificação de corpora segundo seu tamanho permite avaliar objetivamente as previsões de pesquisadores baseadas na intuição. Leech (1991), por exemplo, dizia que 'the going rate' era 1 milhão de palavras. Na verdade, '*the going rate*' é menor que isso (500 mil palavras). Aston (1997) julgava que um corpus de 20 a 200 mil palavras era pequeno. O que os dados revelam, entretanto, é que um corpus de 200 mil palavras não é pequeno em comparação ao que se tem usado. Aston também fez um comentário segundo o qual um corpus grande teria 100 milhões de palavras ou mais. Essa previsão está superestimada, uma vez que um corpus grande não precisa ser tão extenso, já que os maiores corpora usados pela comunidade são mais modestos.

Levantamentos como o apresentado aqui são importantes e devem ser feitos periodicamente, para se monitorar as possíveis mudanças de padrões aceitos pela comunidade. Eles devem ser o mais abrangente possível, ampliando-se a base de dados tanto em quantidade quanto em amplitude. Os dados colhidos podem ser considerados uma amostra representativa de conferências voltadas à Lingüística de Corpus, mas não dão conta, obviamente, de toda a produção da área. Para isso, seria necessário incluirem-se dados de várias publicações, como livros e periódicos.

Recebido em: 02/2000. Aceito em: 05/2001.

Referências bibliográficas

ASTON, G. 1997. Small and large corpora in language learning. Paper presented at the PALC Conference, University of Lodz, Poland, April 1997.

BERBER SARDINHA, A.P. 1998. Size of a representative corpus. Summary of discussion on CORPORA email discussion list, 26 August 1998.

BIBER, D. 1990. Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing*, **5**: 257-269.

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

BIBER, D. et al. 1998. *Corpus linguistics - Investigating language structure and use*. Cambridge University Press.

DE HAAN, P. 1992. The optimum corpus sample size? IN: G. LEITNER (org.). *New directions in English language corpora*. De Gruyter.

HALLIDAY, M.A.K. 1992. Language as system and language as instance: The corpus as a theoretical construct. IN: J. SVARTVIK (org.). *Directions in corpus linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*. De Gruyter.

KENNEDY, G. 1998. *An introduction to corpus linguistics*. Longman.

LEECH, G. 1991. The state of the art in corpus linguistics. IN: K. AIJMER & B. ALTENBERG (orgs.) *English corpus linguistics - Studies in honour of Jan Svartvik*. Longman.

LEWANDOWSKA-TOMASZCZYK, B. & P.J. MELIA (orgs.) 1997. *PALC'97 - Practical applications in language corpora*. Lodz University Press.

LJUNG, M. (org.) 1997. *Corpus-based studies in English - Papers from the Seventeenth International Conference on English Language Research on Computerized Corpora (ICAME 17)*. Rodopi.

MCENERY, T. & A. WILSON 1996. *Corpus linguistics*. Edinburgh University Press.

PERCY, C.E. et al (orgs.) 1996. *Synchronic corpus linguistics - Papers from the Sixteenth International Conference on English Language and Research on Computerized Corpora (ICAME 16)*. Rodipi.

RENOUF, A. (org.) 1998. *Explorations in corpus linguistics*. Rodopi.

SANCHEZ, A. 1995. Definicion e historia de los corpus. IN: A. SANCHEZ et al (orgs.) *CUMBRE – Corpus linguistico de espanol contemporaneo*. SGEL.

SANCHEZ, A. & P. CANTOS 1997a. El ritmo incremental de palabras nuevas en los repertorios de textos. Estudio experimental y comparativo basado en dos corpus linguisticos equivalentes de cuatro millones de palabras, de las lenguas inglesa y espanola y en cinco autores de ambas lenguas. *Atlantis*, **19**.2: 1-27.

____ 1997b. Predictability of word forms (types) and lemmas in linguistic corpora. A case study based on the analysis of the CUMBRE corpus: An 8-million word corpus of contemporary Spanish. *International Journal of Corpus Linguistics*, **2**.2: 258-280.

SIBSON, R. 1972. Order invariant methods for data analysis. *Journal of the Royal Statistical Society B (Methodological)*, **34**: 311-337.

SINCLAIR, J. 1991. *Corpus, concordance, collocation*. Oxford University Press.

____ 1996. EAGLES Preliminary recommendations on corpus typology. EAGLES Document EAG TCWG CTYP/P. Consiglio Nazionale delle Ricerche. Istituto di Linguistica Computazionale. Unpublished manuscript. Available at ftp://ftp.ilc.pi.cnr.it.

TEACHING AND LANGUAGE CORPORA 98. 1998. *Proceedings of TALC 98, Keble College, Oxford, 24-27 July 1998*. Humanities Computing Unit, Oxford University.

YANG, D.-H. & M. SONG 1998. How much training data is required to remove data sparseness in statistical language learning? NLP Lab., Department of Computer Science, Yonsei University, Seoul, Korea, http:// december.yonsei.ac.kr/ ~dhyang.

Tony Berber Sardinha holds a PhD in English (University of Liverpool, UK) as well as an MA in Applied Linguistics from the Catholic University of São Paulo (PUC/SP), where he has been an assistant professor of Applied Linguistics since 1998. His current interests revolve around Corpus Linguistics, including methodological issues, the compilation and exploration of a range of corpora (learner, Portuguese, business), Multidimensional analysis, and corpus-based research in translation.