QUE TIPO DE *CORPUS* É A WEB?

A. P. Berber Sardinha[*]

RESUMO: *Recentemente tem havido um aumento do interesse em ver a Web como um* corpus. *A Web é uma fonte gigantesca, renovável, gratuita e multilíngüe de dados lingüísticos. Este trabalho volta-se à questão da avaliação do tipo de evidência que a Web propicia por meio da realização de um contraste da evidência advinda da Web, em relação àquela fornecida por um* corpus *tradicional, sem textos da Web. O* corpus *tradicional é formado pelo componente falado do Banco de Português, um* corpus *eletrônico de grandes proporções de língua portuguesa brasileira contemporânea mantido na PUC-SP. O* corpus *da Web é formado por todas os resultados retornados pelo buscador Google. O trabalho detalha os vários problemas enfrentados durante a execução da pesquisa, interpreta os achados em vista dos desafios conceituais e metodológicos impostos pelo novo meio e discute questões como a possibilidade de a Web substituir* corpora *tradicionais como objeto de investigação lingüística, o tamanho da Web em relação a várias línguas, principalmente ao português, e as limitações das tecnologias atuais de busca de dados na Web disponíveis para o lingüista.*

PALAVRAS-CHAVE: Corpora *eletrônicos; a Web; Lingüística de* Corpus*; linguagem oral; língua portuguesa.*

Introdução

Um dos maiores desenvolvimentos humanos dos tempos recentes é, sem dúvida, a Internet. Esta rede de

[*] LAEL, Pontifícia Universidade Católica de São Paulo, PUC-SP.

computadores tem revolucionado os meios pelos quais os seres humanos se comunicam, permitindo acesso à informação numa escala nunca antes vista. Grande parte do sucesso da Internet é devido à World Wide Web, também conhecida como WWW, ou simplesmente Web, que permite a computadores e a seus usuários se comunicarem, usando interface gráfica por meio de elos (*links*) entre arquivos. A Web foi criada no início dos anos noventa e, devido à facilidade com que permite a conexão dinâmica entre pessoas, pode ser vista como uma criação mais social do que técnica (Berners-Lee, 1999: 169): a Web revoluciona a maneira como as pessoas vivem e se relacionam.

Tal revolução, nas palavras de Crystal (2001: viii), é de cunho essencialmente lingüístico:

> [...] what is immediately obvious when engaging in any of the Internet's functions is its linguistic character. If the Internet is a revolution, therefore, it is likely to be a linguistic revolution.

O caráter revolucionário da Web se dá tanto pelo que ela faz, mudando os recursos pelos quais as pessoas se socializam, à medida que se engajam em ambientes e veículos de comunicação novos (*chat, email, blog, Webpage, buddy list* etc.), quanto pelo que ela representa: o conjunto das interações e das informações disponibilizadas na rede, continuamente, dia a dia, em qualquer parte do mundo em que a infra-estrutura esteja presente, resulta num gigantesco *corpus* dessa interação. Recentemente, a comunidade de lingüistas, principalmente aquela que se volta à análise de *corpus* eletrônico (nas mais diversas áreas, como a Lingüística de *Corpus*, Processamento de Linguagem Natural e Lingüística Computacional) se deu conta da existência desse *corpus*. O número especial do periódico *Computational Linguistics*, um encontro especial da associação francesa de Lingüística de *Corpus* (TALN, *Traitement Automatique des Langues Naturelles*), além de comunicações em congressos (Kilgarriff, 2001; Renouf, 2001), todos destinados ao

assunto 'Web como *Corpus*',[1] são provas do aumento do interesse nessa área.

Um dos maiores defensores da Web como *corpus* é Adam Kilgarriff. Segundo ele (Kilgarriff, 2001: 344), 'the *corpus* of the new millennium is the Web'. Ele ressalta, porém, o problema que o aparente caos da Web traz para o lingüista:

> For the Web to be useful for language study, we must address its anarchy. If the Web is a torrent and nothing more, it is not useful; for it to be useful, we must channel off manageable quantities to irrigate the pastures of scientific and technological progress. (Kilgarriff, 2001: 342)

Ou seja, se é inquestionável que a Web é imbatível em relação à quantidade de dados, fornecendo um *corpus* de tamanho insuperável, ao mesmo tempo, dado o caos que parece reinar na sua organização, empregar a Web como *corpus* também deveria significar a busca da qualidade desses dados.

O trabalho relatado aqui pretende fornecer subsídios para uma pesquisa nessa direção. A pesquisa apresentada confronta dados obtidos na Web com aqueles retirados de um *corpus* tradicional (que não possui dados da Web). A pesquisa foi conduzida buscando-se aferir o grau de semelhança entre os dois *corpora* (Web e tradicional), de modo específico, discutido em mais detalhes nas seções que se seguem. Caso haja alto grau de similaridade entre os dois *corpora*, não seria ilegítimo supor que há redundância entre os dois *corpora*. Nesse caso, um dos *corpora* seria preterido, dependendo da situação em que o pesquisador específico estivesse envolvido. Àquele pesquisador que não possuísse um *corpus* tradicional seria mais prudente recomendar que usasse a Web, visto que ela possui, entre outras qualidades, variedade, renovação, abundância e baixo custo. Por outro lado, aquele pesquisador que dis-

[1] Respectivamente, www.itri.bton.ac.uk/~Adam.Kilgarriff/wac_cfp.html e http://conferences.atala.org /conferences /fiches /TALN,Corpus+Web.html

ponha de um *corpus* tradicional estaria mais propenso a preterir a Web em favor do *corpus* de que já dispõe. Aliás, para esse indivíduo, seria até mesmo legítimo questionar a validade da Web como *corpus*, já que, se ela nada acrescenta a um *corpus* tradicional, em termos de qualidade, a questão que surge é se realmente temos necessidade de maior quantidade. Entretanto, se a similaridade não for alta, nenhum desses cenários se delineará claramente, e o caminho mais sensato seria, então, não descartar nenhum dos *corpora*, mas integrar os dois. Como se vê, todas essas conjecturas são plausíveis, já que os dados de pesquisa que poderiam nos permitir propor respostas, mesmo que parciais, são escassos.

Um primeiro passo na pesquisa da Web como *corpus* é esclarecer o que se entende por Web. Segundo Crystal (2001: 13), 'the popular practice of using the terms *Internet* and *Web* interchangeably is very misleading'. A Web, nas palavras de seu criador, Tim Berners-Lee, é 'a global hypertext system [...] allowing anything to link to anything',[2] via Internet. A palavra 'Web', conforme dito na abertura deste trabalho, designa uma abreviação de World Wide Web, por isso é grafada com inicial em maiúscula.[3] Já Crystal (2001: 3) associa a grafia em letra maiúscula, como se fosse um nome próprio, à significância do meio, já que palavras relacionadas, como Internet e Net também são grafadas com a primeira letra em caixa alta:

> [...] there is no denying the unprecedented scale and significance of the Net, as a global medium. The extra significance is even reflected in the spelling, in languages which use capital letters: this is the first such technology to be conventionally identified with an initial capital. We do not give typographical enhancement to such developments as 'Printing', 'Publishing', 'Broadcasting', 'Radio', or 'Television', but we do write 'Internet' and 'Net'.

[2] http://www.w3.org/People/Berners-Lee/FAQ.html#Spelling

[3] http://www.w3.org/People/Berners-Lee/FAQ.html#Spelling

A Web, estritamente falando, é uma coleção de arquivos de computador em rede, que faz parte da Internet. A Internet, por sua vez, é uma rede de computadores. Parte dessa coleção de arquivos serve para fazer a própria rede funcionar; outra parte serve como 'conteúdo', isso é, como material que o usuário da rede pode acessar. Esses arquivos desempenham várias funções: alguns são programas de computador (executáveis para os mais variados propósitos, inclusive para causar dano à própria rede, como um vírus!), outros são páginas de texto para serem vistas ou lidas, outros são imagens (fixas ou em movimento, isto é, figuras, fotos, vídeo, cinema), outros, ainda, são arquivos de som, como música e voz. Na medida em que a rede guarda conteúdo, ela pode ser vista como um depósito de dados lingüísticos, isto é, de dados de linguagem humana (em oposição a dados de linguagem de computador, tais como *scripts* de programa).

Grande parte do material encontrado na Web destinado ao consumo humano[4] é de caráter lingüístico: *emails*, *chats*, *blogs*, *guestbooks*, fóruns, listas de discussão, *Webpages*, além dos formatos impressos disponibilizados *online*, como jornais, revistas e periódicos. Temos aí uma vasta coletânea de material, em grande quantidade, já em formato eletrônico. Em outras palavras, um *corpus*, com as vantagens adicionais de ser renovável, multilíngüe, gratuito, acessível remotamente e armazenado de tal modo que não ocupa espaço no computador do pesquisador.

Encarar a Web como *um corpus* é diferente de vê-la como *fonte de material para corpus*. No primeiro caso, a Web inteira é o *corpus*, ao passo que no segundo, a Web é fonte para coleta de material que irá compor um *corpus* nos moldes tradicionais. A situação na qual a pesquisa relatada aqui se insere é referente ao primeiro caso. O que distingue um *corpus* tradicional de um *corpus* de Web não é, portanto, o conteúdo, já que um *corpus* tradicional pode conter arquivos retirados da Web.

---

[4] Em oposição àquele voltado para os próprios computadores, como programas e código de programação.

Um dos grandes desafios da criação de *corpora* tradicionais é a incorporação de dados de fala. Dados de conversação, aulas, entrevistas, e demais manifestações da fala são notoriamente difíceis de obter e de transcrever. Dada a tecnologia de que dispomos, a transcrição da fala é um passo indispensável para tornar um evento falado disponível para inclusão num *corpus*, já que não é possível anexar apenas o conteúdo acústico das vozes (a gravação). Por isso, os *corpora de fala* são por definição *corpora de transcrição de fala*, ou seja, de registros feitos por meio de caracteres gráficos (letras do alfabeto). O componente acústico (as gravações das vozes dos falantes) é, por *default*, subtraído do *corpus*. *Corpora* que trazem consigo o componente acústico são raros e destinados a aplicações e a setores específicos, como aqueles voltados para o estudo da fonética, entoação ou para o treinamento de programas de reconhecimento de voz. À medida que a tecnologia avança, contudo, é de se esperar que esse cenário se modifique. É possível antever um crescimento na quantidade de *corpora* que tragam consigo tanto a transcrição quanto a gravação acústica da fala, além do aparecimento de programas de computador que façam transcrição de fala a partir de gravações de diálogo em ambiente natural (e não se restrinja a monólogos ditados, como é o caso hoje em dia) e, também, de programas de análise léxico-gramatical de *corpora* que atuem diretamente sobre os dados acústicos, dispensando a intermediação da transcrição gráfica. De todos esses possíveis avanços, claramente é o último que demandará mais tempo para se tornar disponível, mas se e quando se concretizar, tornará possível uma maior disponibilização de dados de fala na Web.

No que se refere ao conteúdo de textos escritos, a Web parece se assemelhar mais a um *corpus* tradicional, na sua composição. Muitos dos gêneros, registros e demais variedades encontradas em *corpora* tradicionais habitam naturalmente a Web, muitos até em profusão, como os vários tipos de textos encontrados em jornais e revistas, em periódicos especializados e no âmbito acadêmico (artigos e dissertações, por exemplo).

Por isso, diante do propósito desta pesquisa, pareceu mais natural perguntar qual seria o grau de semelhança entre a Web e um *corpus* tradicional no que se refere à fala, pois esse é o 'calcanhar de Aquiles' dos compiladores de *corpus*. Se a Web mostrar-se como sendo um depositório de registros de fala, então será possível pensar que, mesmo com os desafios existentes (delineados abaixo), é possível utilizá-la como substituta de *corpora* de fala.

Ao restringir o escopo da pesquisa dessa maneira, não será possível, nem é intenção do trabalho, caracterizar a linguagem da Web por completo. Mas, o argumento colocado aqui é que uma caracterização da linguagem da Web, a fim de entendermos melhor o tipo de *corpus* nela disponibilizado, passa, efetivamente, pela verificação da presença de linguagem falada nela.

1. *Corpora* utilizados no estudo

Esta seção apresenta detalhes acerca dos *corpora* usados na pesquisa.

Os *corpora* de estudo empregados na investigação são três: um relativo ao componente falado do *corpus* Banco de Português e um correspondente à Web. O primeiro faz parte do projeto DIRECT (da PUC-SP), que se destina à investigação da linguagem de negócios. O Banco de Português é um *corpus* variado. A extensão do *corpus* de fala aparece na tabela abaixo:

Palavras (*tokens*)	3.157.450
Formas (*types*)	47.852

O segundo *corpus* de estudo é a Web. Graças ao fato de Grefenstette e Nioche (2000), Berber Sardinha (2003) pode estimar o conteúdo da Web em o português como tendo a extensão abaixo:

Palavras (*tokens*)	5.972.909.999

Além desses dois *corpora* de estudo, um terceiro *corpus*, com a função de ser um *corpus* de referência, foi empregado para a extração de palavras-chave. O *corpus* de referência empregado foi o componente escrito do Banco de Português (o mesmo de onde foi retirado o *corpus* de fala). A sua extensão aparece em detalhes no quadro abaixo.

Palavras (*tokens*)	230.460.560
Itens (*types*)	607.392

O Banco do Português, assim como os demais *corpora* atuais (BNC, *Bank of English* etc.), não possui textos advindos do ambiente digital, como *e-mails*, *chats*, fóruns, *guest book* e *Webpages*.

2. Palavras-chave da fala

Conforme diz Hoey (1997), na Lingüística de *Corpus* começa-se a análise com a palavra. Por isso, o primeiro passo da análise mostrada neste trabalho foi a delimitação de quais palavras constituiriam o ponto de partida da investigação. Uma maneira de identificar quais palavras são mais típicas em um *corpus* é comparar a freqüência das palavras do *corpus* de estudo (no caso, o de linguagem falada) com um *corpus* de referência. O *corpus* de referência serve como termo de comparação para as freqüências. Aquelas freqüências que se apresentam estatisticamente mais altas no *corpus* de estudo do que no *corpus* de referência são consideradas palavras-chave. Esse procedimento, que está descrito em mais detalhes em outros trabalhos (Berber Sardinha, 1999b; Scott, 1997), é realizado por meio de computador, no programa WordSmith Tools (Scott, 1998), com a ferramenta KeyWords. O *corpus* de referência empregado foi o componente escrito do Banco de Português, conforme já mencionado.

O resultado da extração das palavras-chave aparece a seguir, de modo resumido:

	Palavra-chave	Fala		Escrita		p
		Freq.	%	Freq.	%	
1	NÉ	65.076	2,06	67.533	0,03	0,000000
2	EU	57.732	1,83	248.174	0,11	0,000000
3	PRA	29.878	0,95	48.410	0,02	0,000000
4	ENTÃO	21.486	0,68	89.223	0,04	0,000000
5	GENTE	19.613	0,62	70.789	0,03	0,000000
6	AÍ	17.435	0,55	51.773	0,02	0,000000
7	ASSIM	23.059	0,73	130.506	0,06	0,000000
8	AQUI	17.449	0,55	75.829	0,03	0,000000
9	É	76.958	2,44	1.971.886	0,86	0,000000
10	LÁ	16.313	0,52	70.894	0,03	0,000000

A palavra-chave mais característica do *corpus* de linguagem falada é 'né', que aparece na linha 1 do quadro. Sua freqüência no *corpus* de fala é de cerca de 65 mil ocorrências, enquanto no *corpus* de escrita ela é de aproximadamente 67 mil ocorrências. Como o *corpus* de fala é expressivamente menor, as ocorrências de 'né' respondem por pouco mais de 2% do total das palavras (*tokens*); por outro lado, no *corpus* de escrita, que é muitas vezes maior do que o de fala, as ocorrências de 'né' representam apenas 0,03% do total. Devido a essa diferença marcante, o teste estatístico (*log-likelihood*) atribui um valor de significância (a coluna 'p' do quadro) expressivo, o que confere a esse item o estatuto de palavra mais típica, em termos de freqüência, do *corpus* de fala.

Esse resultado corrobora tanto o senso comum quanto pesquisas anteriores. Intuitivamente, não resta dúvida de 'né' ser, de fato, um item extremamente comum e característico da fala, especialmente da fala não planejada. Na literatura referente à análise de textos orais de língua portuguesa (e.g. Preti, 1997), também, há estudos que notam a importância de 'né' como marcador da fala (Hilgert, 1997; Urbano, 1997). Urbano (1997: 100) classifica 'né' como um marcador de teste de participação ou de busca de apoio.

Assim, devido a essas características, 'né' parece ser um item adequado para iniciar a investigação. Infelizmente, devido às características da Internet, não é factível repetir o procedimento de

palavras-chave usando a própria Web como *corpus*: é impossível retirar todo o seu conteúdo e processá-lo por meio do programa KeyWords. Mesmo que fosse possível retirar o conteúdo da Web,[5] seria impossível conseguir um *corpus* de referência para viabilizar o procedimento, já que o *corpus* de referência precisa ser maior do que o de estudo, numa magnitude recomendável de cinco vezes (Berber Sardinha, 1999a). Uma outra alternativa seria obter uma porção de textos da rede e usá-los como amostra representativa do total. Contudo, como a composição da rede é desconhecida, esse procedimento se torna problemático.

3. Estimando o grau de similaridade entre os *corpora*

Os resultados da investigação aparecem apresentados e discutidos abaixo.

3.1. Similaridade de freqüência?

A primeira maneira pela qual é possível aferir o grau de similaridade entre dois *corpora* é por meio do exame das freqüências de palavras. Dado que 'né' é uma palavra-chave da fala, pareceu plausível examinar a freqüência dessa palavra como ponto de partida para aferir o grau de similaridade entre dois *corpora*. Assim, embora o estudo enfoque 'né', o objetivo não é o de estudar esse item em si a fundo, mas somente naqueles aspectos relevantes ao problema de pesquisa, que é o de aferir a similaridade entre a Web e a fala.

---

5 Isso seria possível se tivéssemos acesso, por exemplo, aos conteúdos dos servidores de um mecanismo de busca como o Google, que registra em seus computadores o conteúdo da rede, isto é, daquela porção aberta e atingível por meio de robôs (*spiders*) de pesquisa. Isso seria possível se tivéssemos acesso, por exemplo, aos conteúdos dos servidores de um mecanismo de busca como o Google, que registra em seus computadores o conteúdo da rede, isto é, daquela porção aberta e atingível por meio de robôs (*spiders*) de pesquisa.

As freqüências para o *corpus* de fala foram obtidas com a lista de palavras criada pelo programa WordSmith Tools WordList. Para a Web, foi usado o número de ocorrências exibido pelo Google na linha 'Resultados', no topo da página de resultados de busca.

O buscador Google foi escolhido por disponibilizar conteúdo abrangente, diversificado, atualizado e por ser rápido, o que também é uma qualidade indispensável, dada a vasta quantidade de material existente na Web. De modo prático, ele se mostrou mais satisfatório que os demais disponíveis, conforme é explicado a seguir.

Havia outras opções para acessar o conteúdo da Web além do Google. Uma delas é o *site* especializado 'WebCorp', que emula um concordanceador. Mas esse *site* não aceitava caracteres acentuados, inviabilizando a busca pela palavra-chave 'né'. Mesmo depois dessa limitação ter sido sanada, os resultados desapontavam, pois o *site* retornava apenas uma centena de ocorrências, enquanto Google reportava-se a dezenas de milhares.

Outra opção era o programa KWiCFinder, semelhante a WebCorp, mas funciona como um programa instalado na máquina do usuário, e não *online*. O funcionamento do programa mostrou-se desalentador, devido à demora na obtenção dos resultados e ao mau funcionamento (*crashes*).

Uma terceira opção, o utilitário 'WebGetter', disponível em versão beta (de teste) do WordSmith Tools versão 4, foi testado. Esse programa faz *download* de textos para a máquina do usuário, baseado na presença de um item determinado. No nosso caso, foi testada a busca por 'né'. Entretanto, o *software* ainda apresentava muitos problemas de execução na condução dessa pesquisa.

Para acessar o conteúdo exclusivo de português, foi usado o procedimento seguinte. Google oferece a opção de escolha de língua da página (no *site* brasileiro, diretamente na página de entrada com a opção 'Pesquisa páginas em português', e no americano, com a opção 'Advanced Search' / 'Return pages written in'). Contudo, os resultados não se mostraram confiáveis para 'né'. Muitas ocorrên-

cias vinham claramente de textos em francês (em que 'né' significa 'nascido'). Além disso, havia muitas ocorrências de NE, em maiúsculas e sem acento (significando muitas vezes Nordeste), o que se revelou um erro grosseiro do mecanismo de busca. Por isso, a opção mais segura foi restringir o domínio, para incluir somente endereços brasileiros, que possuem final .br. O termo de busca, portanto, foi né .br .[6]

Google mostra quantidade de ocorrências na linha 'Resultados', no topo da página com os resultados da busca. Esse número é um valor arredondado. Ocorrências repetidas são omitidas. Para essa pesquisa, foi desativada essa opção do programa, mas, mesmo assim, as quantidades encontradas eram bastante inferiores à freqüência relatada pelo próprio *site*. Apenas cerca de mil ocorrências foram mostradas.

A tabela abaixo traz os resultados obtidos, além do cálculo de freqüências por milhão de palavras:

Corpus	Palavras (*tokens*)	Freqüência de 'né'	Freq. de 'né' por milhão de palavras
Fala	3.157.450	65.076	20.610
BP	230.460.560	67.533	293
Web	5.972.909.999	79.100	13,2

[6] A decisão de realizar a busca apenas em *sites* brasileiros causa algum conflito com a estimativa do tamanho da Web apresentado acima, que levou em conta o conteúdo em português na Web toda, não somente em domínios .br . Entretanto, considerando que não há dados exatos sobre a parcela do conteúdo da Web em português corresponde aos domínios .br, seria um exercício de especulação tentar determinar esse universo. Assim, a melhor decisão foi manter a estimativa do tamanho da Web já calculado. Como se nota nos resultados do Anexo, contudo, os valores da estatística Escore T são bastante robustos, bem acima do mínimo recomendado (2), o que se deve ao tamanho extremamente grande da quantidade de palavras envolvida (na ordem dos bilhões). Isso nos leva a supor que a estatística de associação de palavra Escore T seria pouco afetada por mudanças menores que o tamanho da Web aceito para a pesquisa, o que, por sua vez, na prática, significa que os resultados, muito provavelmente, permaneceriam inalterados caso fosse feita uma estimativa do tamanho da Web em termos da quantidade de palavras em português existente somente em domínios .br.

Os resultados indicam que há uma semelhança entre os valores absolutos de 'né' nos dois *corpora*; na fala há cerca de 65 mil ocorrências, ao passo que na Web chega-se a quase 80 mil.

Há uma disparidade grande quando se calcula o valor 'por milhão'. Isso ocorreu devido ao fato de, no *corpus* tradicional, as freqüências se referirem apenas ao *corpus* limitado pelos arquivos referentes à fala. Na Web, entretanto, foi utilizado o valor referente à Web inteira, já que não há dados referentes ao que se constituiria na parte 'falada' da Web (até por que essa porção é desconhecida!). Caso seja calculada a freqüência por milhão de palavras no *corpus* tradicional, incluindo-se o componente escrito do Banco de Português, que chega a 230 milhões de palavras, a freqüência de 'né' por milhão baixa sensivelmente para 293.

Tomando como pressuposto o fato de 'né' ser um indicador de oralidade, então é legítimo sugerir que a freqüência desse item na Web indica que há oralidade nos textos disponíveis na rede. A próxima questão é saber se essa oralidade está presente em textos semelhantes aos do *corpus* tradicional,

3.2. Similaridade de fonte?

Para saber se a fonte de onde surgiram as ocorrências de 'né' na Web e de onde foram computadas as freqüências analisadas acima são semelhantes às amostras contidas no *corpus* tradicional, é preciso comparar as fontes dos dois *corpora*.

O desafio maior concentra-se, logicamente, na definição das fontes dos textos da Web. Para a pesquisa, antes de se conduzir a busca de 'né' no Google, foram feitos ajustes nas opções de exibição do buscador, de tal modo que cada tela de resultados da busca exibisse o maior número de ocorrências, para facilitar e apressar a gravação dos dados. O número máximo por *site* permitido era 100, que foi, portanto, o valor escolhido. Assim, cada tela do resultado da busca, que trazia 100 ocorrências, foi salva em um arquivo .txt

(texto simples). Após dez telas, ou mil ocorrências, chegou-se ao final do resultado da busca. Esse resultado de mil ocorrências estava muito abaixo das quase 80 mil ocorrências relatadas pelo próprio *site*. Infelizmente, essa limitação do *site* não é passível de mudança. Mesmo quando foi solicitado (dentro do menu 'preferências') que a busca não excluísse resultados repetidos ('repeat the search with the omitted results included'), poucas ocorrências foram adicionadas. Assim, a quantidade de ocorrências de 'né' na Web restringiu-se a mil.

Tendo em mãos, portanto, o resultado das telas de buscas do Google gravado em arquivo .txt, foi feita, então, uma listagem dos endereços visitados pelo buscador. Descobriu-se que as mil ocorrências vinham de 713 páginas da Web (conforme reveladas pelo endereço 'HTTP' indicado para resultado da busca). Para saber o conteúdo exato dessas 713 páginas, seria preciso visitar cada uma delas e fazer uma análise de seu conteúdo, sua classificação genérica ou de registro etc. Embora isso seja possível, mas cansativo, o procedimento deixa a desejar, visto que o que se deseja é um meio de conhecer a Web sem ter de visitá-la página por página. Por isso, um procedimento mais produtivo e adequado para os propósitos desta pesquisa foi seguido e não depende da consulta às páginas propriamente ditas. Esse procedimento baseia-se na análise dos endereços das páginas visitadas pelo buscador. Para tanto, somente os endereços constantes nas páginas de resultados do Google foram salvos num arquivo-texto, por meio de processamento com utilitários Unix (grep e sed). A lista de endereços resultante foi então processada pelo WordSmith Tools WordList, que, por sua vez, retornou uma lista de palavras com os itens que constavam nos endereços consultados. Foi feita, a seguir, uma leitura da lista até o ponto em que as palavras possuíam freqüência igual ou maior a três, a fim de se saber qual a fonte ou o conteúdo das páginas visitadas. As palavras que aparecem mais de três vezes, que podem trazer alguma indicação da fonte ou do conteúdo das páginas visitadas aparecem na relação a seguir:

BLOG	ENTREVISTAS	RECADOS
BLOGGER	FORUM	REPORTAGEM
BOTEQUIM	GUEST	REVISTA
CHAT	GUESTBOOK	SOM
COLETIVA	HIPHOP	
COLUNISTAS	HOBBIES	
CRONICAS	MURAL	
DEPOIMENTOS	MUSICA	
ENTREVISTA	NOTICIAS	
	PIADAS	

A relação revela que muitas páginas foram advindas de *sites* voltados para interações de caráter estritamente digital, tais como *blogs*, *chats*, *guestbooks* e fóruns. Há, ainda, páginas que vieram de *sites* com conteúdo de interação não digital, tais como entrevistas, depoimentos e coletivas. Há, ainda, outros que indicam que a fonte seja do modo escrito, como notícias, recados, reportagens.

O quadro abaixo traça uma comparação entre os dois *corpora*.

Corpus	Fontes
Fala	Aulas
	Conversação informal
	Entrevistas
	Palestras
	Reuniões
Web	*Blogs*
	Chats
	Crônicas
	Depoimentos
	Entrevistas
	Fóruns
	Guestbooks
	Murais

Como se percebe, há pouco em comum entre os dois *corpora*. Conforme o que foi dito anteriormente, há uma presença muito forte, na Web, de gêneros típicos do meio digital; do mesmo modo, no *corpus* de fala, há gêneros típicos desse meio. Assim, a similaridade de freqüência apurada tem de ser interpretada no contexto das diferenças descobertas aqui. Isso significa que a oralidade indicada pelo uso de 'nê', de modo freqüente na Web, manifesta-se

em gêneros diferentes daquele encontrado na fala. Ou seja, há oralidade na Web, mas não registro de fala.

3.3. Similaridade de uso?

O terceiro aspecto passível de revelar similaridade entre os *corpora* investigados aqui é o que concerne ao uso do marcador em questão ('né'). Colocado de outro modo, isso significa perguntar até que ponto o uso de 'né' nos dois ambientes é similar. Essa questão se traduz a partir do ponto de vista da Lingüística de *Corpus*, em uma investigação da padronização léxico-gramatical, mais especificamente, dos padrões de colocação de 'né'. Um padrão (*pattern*) é uma regularidade léxico-gramatical observável num *corpus* (Berber Sardinha, 2000; Hunston e Francis, 2000; Stubbs, 2001). Os padrões são unidades estruturais e de sentido, ao mesmo tempo advindas da co-ocorrência de unidades recorrentes do *corpus*. Há uma relação estreita entre a estrutura do padrão e o sentido que ele exibe, de tal modo que padrões distintos de um mesmo item indicam significados distintos (Sinclair, 1991).

Para descrever a padronização de 'né' foram seguidos alguns passos. Em primeiro lugar, foram extraídos os colocados de 'né' (a palavra nódulo, ou *node word*) do *corpus* de fala, com o programa WordSmith Tools Concord. Dado que 'né' é um marcador que ocorre fundamentalmente em posição final de oração (como um *tag*), pareceu sensato restringir a extração dos colocados àqueles que estivessem posicionados à esquerda do nódulo. Foram levados em conta os colocados que estivessem a até cinco palavras de distância do nódulo. Do total de colocados retornado pelo Concord, apenas os 50 mais freqüentes foram selecionados para análise. Esse número pareceu suficiente para o propósito dessa pesquisa, mais especificamente o de verificar a semelhança entre os dois *corpora*, já que a semelhança, caso exista, deva se manifestar entre os colocados mais freqüentes.

Em segundo lugar, foi feita uma procura por 'né' no *site* de busca Google. Google apresenta um trecho do texto onde ocorre a palavra de busca. Esse trecho é variável, mas preenche normalmente duas linhas de texto, na tela. Para 'né', foi indicado pelo *site* a existência de 79.100 ocorrências do nódulo, conforme já mencionado. A quantidade de ocorrências de 'né' utilizada foi mil, devido às razões já explicadas.

Em terceiro lugar, o arquivo-texto em que foi gravado esse conjunto de ocorrências de 'né', juntamente com as linhas de contexto retornadas pelo Google, foi submetido ao Concord do programa WordSmith Tools para extração dos colocados. Assim, tecnicamente, o arquivo com o resultado de busca do Google pode ser visto como um texto, a partir do qual foi realizada a análise dos colocados. O arquivo-texto foi processado para retirada de trechos indesejados, todos relacionados a textos de formatação da página do Google (tais como os referentes ao *link* 'Translate this page', que acompanham cada ocorrência. Essa 'limpeza' foi levada a cabo com utilitários de texto (sed e tr) do sistema operacional Unix (disponíveis no emulador Cygwin para a plataforma Windows). Os colocados foram, então, extraídos seguindo-se o mesmo critério usado para o *corpus* de fala, ou seja, foram considerados os colocados ocorridos à esquerda do nódulo, até a quinta palavra à esquerda.

Por fim, foi calculado o Escore T (*t-score*), que é estatística de associação (Stubbs, 1995), para cada um dos 50 colocados mais freqüentes de 'né' em cada um dos *corpora* (fala e Web). Tal estatística foi necessária para indicar quais das co-ocorrências entre nódulos de colocados eram mais significativas, ou seja, tinham ocorrido mais freqüentemente do que o esperado. O Escore T é calculado entre pares de palavras; no nosso caso, entre o nódulo 'né' e seus colocados.

O Escore T não é a única estatística desse tipo. Outra é a Informação Mútua (*Mutual Information*; Stubbs, 1995). A diferença básica entre elas é que o Escore T é mais indicado para revelar a

certeza da colocação (Hunston, 2002), já que tende a indicar como associados aqueles pares de palavras que sejam significativamente muito freqüentes. Já a Informação Mútua tende a mostrar como relevantes os pares que possuem baixa freqüência no *corpus*, embora possuam várias co-ocorrências em comum. Devido às características dos *corpora* em questão, notadamente a alta freqüência das palavras envolvidas, o emprego do Escore T foi mais indicado.

A fórmula para o cálculo do Escore T é a seguinte:

$$T = ((f(n,c)/N) - (f(n)/N * f(c)/N)) / \sqrt{(f(n,c) / N)}$$

Valores de T maiores ou iguais a dois são considerados indicativos de associação significativa entre o nódulo e o colocado (Stubbs, 1995). Por isso, o ponto de corte para seleção dos colocados significativos foi $T \geq 2$.

Para calcular o Escore T foi preciso obter os seguintes dados:

- Tamanho do *corpus* (N);
- Freqüência do nódulo (f (n));
- Freqüência do colocado (f(c));
- Freqüência de ocorrência do nódulo com o colocado (f(n,c)).

Para o *corpus* de fala, esses valores foram extraídos da tabela de colocados fornecida pelo Concord (o terceiro e quarto itens da relação acima) e pela lista de palavras do WordList (o primeiro e segundo itens), ambos parte do WordSmith Tools. Contudo, foi necessário um ajuste. O programa Concord, porém, tem como limite 16 mil ocorrências. Contudo, no caso de 'né' havia 65.076 ocorrências. Foi preciso, portanto, fazer uma seleção das linhas de concordância, na proporção de 1 para 4, usando a função 'at random', em 'settings'. Em virtude disso, o valor referente ao *corpus* precisou ser ajustado, já que, de fato, as freqüências dos colocados não se referem ao *corpus* todo, mas sim a ¼ dele. Assim, o valor referente ao tamanho do *corpus* (N), ajustado, foi de 789.362,5.

Para a Web, o único valor passível de obtenção com o WordSmith Tools Concord foi a freqüência de ocorrência do nódulo

com o colocado, a partir da tabela de colocados que teve como base o arquivo-texto de resultados do Google, conforme detalhado anteriormente. Detalhes do cálculo dos valores aparecem abaixo.

- Freqüência do nódulo: número de ocorrências exibidas pelo Google na linha 'Resultados [...] sobre [...]'. O valor desejado estava expresso logo após 'sobre'. A freqüência de 'né', apurada desse modo, foi de 79.100. Esse número não é fixo. A tendência é que esse valor cresça, à medida que o conteúdo da Web se expanda e que mais conteúdo seja indexado, tornando-se acessível pelo buscador.
- Freqüência conjunta do nódulo com o colocado. Esse valor foi expresso pela tabela de colocados do WordSmith Tools. Contudo, esses valores não puderam ser usados, pelo modo como estavam representados. A razão disso é que eles demonstravam valores referentes a apenas mil ocorrências de 'né', quando, na verdade, o valor total do nódulo é 79.100. Assim, foi necessário um ajuste nas freqüências. Para cada colocado, foi multiplicada sua freqüência por 79,1, que é o resultado da divisão de 79.100 por 1000. A freqüência conjunta ajustada reflete, em conseqüência, a freqüência que o nódulo possivelmente teria, caso tivesse sido possível obter as 79.100 ocorrências de 'né' relatadas pelo Google.
- Tamanho do *corpus*: 5.972.909.999 palavras, conforme explicado acima.
- Freqüência do colocado: foi feita uma busca para cada colocado no próprio Google. A freqüência de cada um foi obtida pela indicação do número de ocorrências encontradas pelo buscador, na linha 'Resultados', do mesmo modo realizado para a freqüência do nódulo.

Os valores do Escore T foram então calculados numa planilha Excel, utilizando a fórmula apresentada anteriormente. Foi, então, aplicado o ponto de corte de $T \geq 2$ para a seleção dos colocados. Dos 50 colocados iniciais para cada *corpus*, a quantidade dos que foram identificados como significantes pela estatística de Escore T aparecem na tabela a seguir.

Corpus	Colocados com T ≥ 2
Web	50
Fala	37
Total	87 (*tokens*)

Como se percebe, todos os 50 colocados iniciais da Web passaram pelo ponto de corte do Escore T. Isso se deve aos valores altos referentes aos colocados e ao tamanho do *corpus* (vide anexo). Já em relação ao *corpus* de fala, dos 50 colocados iniciais, apenas 37 (64%) mantiveram-se acima do ponto de corte. Assim, a aplicação do Escore T como medida de discriminação dos colocados foi válida, pois em um dos *corpora* ela permitiu selecionar mais criteriosamente quais colocados foram considerados para posterior análise.

Esses 87 colocados com Escore significativo correspondem, na verdade, a 58 unidades (*types*). O grau de compartilhamento e de exclusividade dos colocados foi, então, estimado por meio da contagem de quantas unidades eram específicas de cada *corpus* e quantas eram comuns aos dois *corpora*. Os resultados aparecem na tabela a seguir:

Colocados	Freq.	%
Em comum	29	50%
Somente Web	21	36%
Somente fala	8	14%
Total	58	

Os resultados indicam que há um grau semelhante de compartilhamento e de exclusividade entre os colocados. Exatamente a metade é compartilhada (29 dos 58). Entretanto, os restantes, que são exclusivos, não se dividem igualmente entre os dois *corpora*. A maioria dos colocados exclusivos está na Web (21) e apenas 8 no *corpus* de fala.

Mesmo assim, o resultado mais saliente é o referente à similaridade dos dois *corpora*, já que metade dos colocados ocorre tanto em um quanto em outro *corpus*. Isso seria diferente caso algum dos *corpora* fosse dominante no que diz respeito ao total de colocados;

porém, o que predomina é o compartilhamento. Uma decorrência disso é que é possível sugerir, com base no compartilhamento observado nos colocados, que os sentidos expressos pelo uso de 'né' nos dois *corpora* é em grande parte semelhante, já que, conforme exposto acima, o pressuposto da análise de colocação é o de que os padrões léxico-gramaticais revelados são unidades de sentido, além de unidades estruturais.

É preciso dar conta, ainda, do fato de a metade dos colocados ser exclusiva de um ou de outro *corpus*. A maioria dos colocados exclusivos provém da Web (21 dos 29, ou 72%). Antes de levar adiante essa parte da análise convém apresentar os colocados que ocorreram de modo compartilhado e exclusivo nos *corpora*:

	Colocados compartilhados	Colocados exclusivos à Web	Colocados exclusivos ao *corpus* de fala
1	A	BLOG	AS
2	AQUI	BOM	ELA
3	ASSIM	COMO	ELES
4	AÍ	E	ERA
5	BEM	FAZER	ESTÁ
6	COISA	FOI	OS
7	COM	LEGAL	PORQUE
8	DA	MAS	TINHA
9	DE	ME	
10	DO	MEU	
11	EM	MINHA	
12	EU	NEM	
13	GENTE	O	
14	ISSO	PARA	
15	JÁ	SABE	
16	LÁ	SÓ	
17	MAIS	TAMBÉM	
18	MESMO	TEM	
19	MUITO	TÁ	
20	NA	VOCÊ	
21	NO	WWW	
22	NÃO		
23	PRA		
24	QUE		
25	SE		

26	TUDO		
27	UM		
28	UMA		
29	É		

A listagem revela que mesmo os colocados exclusivos não são muito diferentes entre si. A maior similaridade entre os colocados exclusivos diz respeito à grande presença de pronomes pessoais (me, meu, minha, você, ela, elas) nos dois conjuntos de colocados exclusivos. A presença de pronomes pessoais da primeira pessoa, entre os colocados exclusivos da Web, requer uma interpretação. Uma possível leitura desse resultado é a de que esse uso dos pronomes parece refletir a necessidade de muitos usuários da Web se autodescreverem perante os seus interlocutores. Em *chats* e *blogs* o assunto é justamente o indivíduo, isto é, suas preferências, suas opiniões, seus passatempos etc. Desse modo, essa descrição ou apresentação individual envolve muitas vezes a representação, em palavras, da própria aparência do indivíduo, já que muitas vezes não há visualização do parceiro da interação. Para expressar tudo isso, os usuários precisam fazer uso dos pronomes pessoais. De modo geral, esse uso dos pronomes pessoais remete ao fato de a comunicação digital ser em grande parte individual: os usuários de computador estão quase sempre digitando sozinhos, ou um de cada vez, distantes uns dos outros. Na fala, por sua vez, não parece haver essa predominância do 'eu' porque, entre outras razões, a presença do falante já é suficiente para que seu interlocutor perceba sua aparência e outros aspectos que, na Web, precisam ser expressos em palavras.

Por sua vez, a maior diferença entre os dois *corpora* fica por conta da presença de itens específicos dos ambientes digitais (*blog* e WWW) entre os colocados da Web. Esses itens referem-se diretamente aos espaços de interação que são exclusivos da Web e que ficaram registrados nos trechos recolhidos pelo buscador Google durante a composição do *corpus*.

Assim, parece possível concluir que há um grau de similaridade marcante entre os dois *corpora* também no que se refere ao uso de 'né'.

Comentários finais

Os resultados obtidos, em resumo, indicam que (1) há oralidade nos textos disponíveis na Web, (2) a oralidade da Web não advém de transcrição de fala, (3) o emprego de um marcador típico da oralidade ('né') não é muito diferente na Web do que é na fala. A conclusão geral a que se pode chegar é a de que há similaridade entre a Web e um *corpus* tradicional de fala, constituído de transcrições de eventos presenciais, no que se refere à oralidade. Ou seja, a oralidade, que constitui a fala, está presente na Web; os eventos em que essa oralidade se manifesta, contudo, não são similares.

Assim, parece legítimo concluir que não há indicação da presença marcante de dados de fala na Web. A oralidade detectada não advém de dados de fala propriamente, como da transcrição de conversas, de entrevistas ou reuniões. A quase totalidade (senão toda) aparece em variedades lingüísticas nativas da própria Web, como *blogs*, *e-mails* e *chats*. Esses tipos de comunicação digital trazem consigo traços de oralidade que se manifestam, conforme visto, no emprego e nos padrões de uso de 'né'. Isso significa que os dados de oralidade da Web não devem ser vistos como dados de fala 'degenerados', mas devem, sim, ser apreciados pelo que são, ou seja, como formas legítimas de interação no meio digital.

Na medida em que a Web parece distinta de um *corpus* de fala tradicional, ela não parece (ainda) capaz de substituir os *corpora* tradicionais. Assim, voltando aos cenários de conduta colocados na introdução, a melhor opção, perante os achados dessa pesquisa, é tratar como objetos distintos a Web e o *corpus* tradicional (de fala, nos moldes do usado aqui). A Web não substitui o *corpus* tradicional de fala, ou vice-versa.

A Web não é representativa da fala humana. Para ser representativo de uma língua, um *corpus* precisa conter amostras da fala, já que ela é parte integrante e fundamental da linguagem. Por extensão, seguindo esse princípio, a Web não pode ser considerada representativa do português, embora seja bem maior do que qualquer *corpus* do português em existência.

Mas, como tudo na Web, essa situação pode ser reverter. À medida que mais transcrições de fala se tornem públicas na Web, mais ela poderá se aproximar de uma amostra representativa da linguagem presencial. De outro modo, a Web deve ser tornar cada vez mais representativa da linguagem humana, à medida que a própria comunicação humana ocorrer, cada vez mais, em ambientes digitais. A tecnologia tende a seguir nessa direção, substituindo o *input* de dados pelo teclado por outros meios, como a própria voz e imagem. Quando isso acontecer, a Web refletirá com maior propriedade a comunicação humana, mesmo que seja aquela realizada a distância.

Por sua vez, os *corpora* tradicionais atuais deveriam reconhecer a importância crescente da Web na comunicação humana e incorporar dados desse meio. Quando isso vier a ser feito, ocorrerá naturalmente uma maior aproximação entre o tipo de amostra da linguagem humana em que um *corpus* tradicional se constitui e a Web.

Por isso tudo, voltando à citação inicial, que afirma que a Web é o *corpus* do novo milênio (Kilgarriff, 2001), talvez seja mais apropriado dizer que ela *deverá ser* o *corpus* do novo milênio.

Os resultados apresentados aqui devem ser vistos, sempre, como referentes ao português brasileiro tal qual ele aparece na Web e no *corpus* de fala, ambos *corpora* limitados. Não é possível generalizar os tipos de achados reportados aqui para outras línguas, pois, entre outras coisas, os tamanhos das amostras de cada língua disponíveis na Web são diferentes. Além disso, talvez haja mais registro de fala para outras línguas do que há para o português.

Por exemplo, no âmbito das línguas européias (incluindo aí o português europeu), há uma quantidade considerável de transcrições de sessões de trabalho do parlamento europeu. Esses dados podem, em princípio, alterar o quadro apresentado aqui. Para o inglês há, ainda, outros tipos de dados que parecem relativamente comuns na Web, como palestras de políticos e executivos, depoimentos em tribunal etc. que também podem mudar o cenário dos dados demonstrados aqui.

Além disso, os achados advêm da exploração da Web segundo os instrumentos que estão disponíveis no momento, notoriamente aqueles permitidos pelo poder de busca de portais como o Google. Um problema com eles é que a quantidade de informação indexada em seus sistemas de busca e, portanto, disponibilizada para o usuário, é apenas uma pequena fração do total. Em primeiro lugar, conforme destacado, a quantidade de citações que é, de fato, mostrada ao usuário é muitas vezes menor do que aquela sinalizada pelo total de achados (*hits*). Em segundo lugar, há o chamado 'lado invisível da Web', que comporta, além do material que é de acesso restrito, aquele que, embora público, os indexadores não conhecem. Esse lado da rede é visto, cada vez mais, como gigantesco e, desse modo, como algo que afeta o dimensionamento que fazemos da Web. Segundo Corliss (2001), existe uma porção da Web, chamada de 'Deep Web', que engloba *sites* que possuem uma organização bastante profunda, com muitos níveis de divisão de diretório. Ela figura em contraste com a Web de superfície (*surface Web*), que é aquela que possui uma organização menos profunda, em que a informação não está 'enterrada' (*buried*) numa estrutura de diretórios tão profunda. Esses 'Deep Web sites' comportam dados em quantidades muitas vezes maiores do que aquelas que os indexadores (como Google e Yahoo) conseguem penetrar, e possuem uma quantidade de informação estimada em mais de 40 vezes a quantidade da Web de superfície. Além disso, os buscadores, mesmo detendo-se na rede de superfície, somente conseguem indexar parte dela

(por vários motivos), restringindo-se a algo em torno de 16% do total existente (Lawrence e Giles, 1999).

Para concluir, a pesquisa relatada aqui fez um trabalho que pode ser considerado inicial, no levantamento da similaridade entre um *corpus* tradicional e a Web. A Web é um patrimônio de valor inestimável, colocado à disposição de seus usuários e também, agora, a lingüistas. Jamais em nossa história tivemos tanto conhecimento disponível diante de nós. No âmbito da Lingüística, o mesmo pode ser dito: jamais tivemos um *corpus* tão extenso e renovável ao nosso dispor. Buscar entender como esse grande registro do conhecimento humano se configura e espelha a interação humana é uma tarefa que se coloca diante dos profissionais da linguagem.

ABSTRACT: *There has been an increased interest in seeing the Web as a corpus. The Web is a large, free, renewable multilingual source of linguistic data. This paper addresses the issue of evaluating the kinds of evidence provided by the web by means of contrasting this evidence with data from a traditional 'non-Web' corpus. The traditional corpus is formed by the spoken section of the 'Banco de Português', a large machine-readable corpus of contemporary Brazilian Portuguese held at PUC-SP. The Web corpus is made up of all the citations returned by Google for a query. The paper details a range of problems faced during the execution of the project, interprets the findings in view of the conceptual and methodological challenges imposed by this new medium, and discusses a number of issues such as the extent to which the Web may replace traditional corpora as an object for linguistic inquiry, the size of the Web corpus for several languages, mainly for Portuguese, and the limitations of current retrieval technologies for Web data available to linguists.*

KEYWORDS: *Electronic Corpora; the Web; Corpus Linguistics; Oral Language; Portuguese.*

Anexo

(1) Colocados do *corpus* de fala, ordenados decrescentemente pelo resultado do Escore T

	Colocado	Freqüência no *corpus* f (c)	Freqüência em conjunto com 'né' f(n,c)	Co-ocorrência observada (O)	Co-ocorrência esperada (E)	Escore T
1	MAIS	18470	767	0,00097	0,00047	14,17676
2	UM	19613	793	0,00100	0,00050	14,04298
3	EM	10100	526	0,00067	0,00026	14,00838
4	ERA	11919	566	0,00072	0,00031	13,63587
5	DE	68253	1940	0,00246	0,00175	12,63569
6	MUITO	16055	596	0,00076	0,00041	11,08310
7	É	76958	2043	0,00259	0,00198	10,68812
8	AÍ	9448	395	0,00050	0,00024	10,23887
9	UMA	21089	691	0,00088	0,00054	10,02538
10	EU	27093	824	0,00104	0,00070	9,57443
11	DO	19066	626	0,00079	0,00049	9,57400
12	ESTÁ	7773	328	0,00042	0,00020	9,41125
13	PRA	29878	885	0,00112	0,00077	9,39150
14	AS	6619	283	0,00036	0,00017	8,84738
15	TUDO	14277	483	0,00061	0,00037	8,80965
16	AQUI	13653	447	0,00057	0,00035	8,05303
17	COISA	10154	357	0,00045	0,00026	8,00147
18	ELES	13361	437	0,00055	0,00034	7,94942
19	GENTE	26395	741	0,00094	0,00068	7,56708
20	COM	20255	594	0,00075	0,00052	7,52667
21	ASSIM	30756	838	0,00106	0,00079	7,41291
22	MESMO	6956	253	0,00032	0,00018	7,04171
23	PORQUE	9564	314	0,00040	0,00025	6,78002
24	NA	19653	557	0,00071	0,00050	6,72192
25	DA	23291	640	0,00081	0,00060	6,63693
26	OS	13441	401	0,00051	0,00035	6,41985
27	TINHA	12455	372	0,00047	0,00032	6,19802
28	SE	19670	538	0,00068	0,00051	6,00559
29	BEM	9992	297	0,00038	0,00026	5,48153
30	LÁ	16313	440	0,00056	0,00042	5,21274
31	NO	21744	533	0,00068	0,00056	3,99620
32	NÃO	62952	1424	0,00180	0,00162	3,92181
33	ISSO	12502	324	0,00041	0,00032	3,92169
34	ELA	10475	271	0,00034	0,00027	3,56436
35	A	83716	1822	0,00231	0,00215	2,93117
36	QUE	113240	2433	0,00308	0,00291	2,79124
37	JÁ	17435	398	0,00050	0,00045	2,23564

(2) Colocados da Web, ordenados decrescentemente pelo resultado do Escore T

	Colocado	Freqüência no *corpus* f (c)	Freqüência estimada em conjunto com 'né' f(n,c)	Co-ocorrência observada ($O \cdot 10^5$)	Co-ocorrência Espe-rada ($E \cdot 10^5$)	Escore T
1	É	2510000	17560,2	0,293997398	0,00056	1176,332
2	QUE	2780000	15029	0,251619395	0,00062	1087,647
3	O	3580000	13526,1	0,226457455	0,00079	1030,741
4	DE	3760000	13367,9	0,22380883	0,00083	1024,469
5	A	3700000	11865	0,198646891	0,00082	964,7724
6	NÃO	2310000	10441,2	0,174809264	0,00051	906,1271
7	EU	808000	9650,2	0,161566138	0,00018	872,7192
8	E	3700000	8305,5	0,139052824	0,00082	805,7519
9	PRA	348000	6169,8	0,103296383	0,00008	698,0704
10	UM	2580000	6090,7	0,101972071	0,00057	690,2059
11	COM	3010000	6090,7	0,101972071	0,00067	689,5570
12	DO	3250000	6011,6	0,100647758	0,00072	684,6407
13	TEM	1700000	4904,2	0,082107382	0,00038	619,9748
14	JÁ	1690000	4429,6	0,074161506	0,00037	588,9394
15	MAS	1530000	4271,4	0,071512881	0,00034	578,5066
16	MAIS	2280000	4192,3	0,070188568	0,00051	571,7091
17	SE	2980000	4192,3	0,070188568	0,00066	570,4358
18	ISSO	1050000	3875,9	0,064891318	0,00023	551,7135
19	NO	3400000	3875,9	0,064891318	0,00075	547,2676
20	UMA	2330000	3796,8	0,063567005	0,00052	543,5671
21	TÁ	124000	3717,7	0,062242692	0,00003	542,0427
22	COMO	2220000	3717,7	0,062242692	0,00049	537,9938
23	WWW	2690000	3638,6	0,06091838	0,00060	531,2298
24	EM	2950000	3638,6	0,06091838	0,00065	530,7221
25	DA	3360000	3401,3	0,056945442	0,00074	511,9076
26	FAZER	1140000	3322,2	0,055621129	0,00025	510,2970
27	SABE	417000	3243,1	0,054296817	0,00009	505,6246
28	BOM	696000	3164	0,052972504	0,00015	498,8149
29	NEM	789000	3164	0,052972504	0,00017	498,6202
30	BEM	1370000	3084,9	0,051648192	0,00030	491,0741
31	FOI	1560000	3005,8	0,050323879	0,00035	484,2537
32	MUITO	1320000	2926,7	0,048999566	0,00029	478,2726
33	ME	1820000	2926,7	0,048999566	0,00040	477,1840
34	COISA	399000	2847,6	0,047675254	0,00009	473,7193
35	PARA	2920000	2847,6	0,047675254	0,00065	468,1550
36	NA	2970000	2847,6	0,047675254	0,00066	468,0446
37	SÓ	1160000	2768,5	0,046350941	0,00026	465,3652
38	MEU	687000	2689,4	0,045026629	0,00015	459,6680
39	TAMBÉM	1700000	2689,4	0,045026629	0,00038	457,3672

40	GENTE	522000	2531,2	0,042378003	0,00012	446,2351
41	ASSIM	1020000	2373	0,039729378	0,00023	430,7823
42	MESMO	1310000	2293,9	0,038405066	0,00029	422,7450
43	AQUI	1500000	2293,9	0,038405066	0,00033	422,2777
44	TUDO	993000	2135,7	0,03575644	0,00022	408,4848
45	BLOG	118000	2056,6	0,034432128	0,00003	403,0259
46	LEGAL	734000	2056,6	0,034432128	0,00016	401,4261
47	LÁ	402000	1977,5	0,033107815	0,00009	394,4352

BIBLIOGRAFIA

BERBER SARDINHA, A. P. (1999a). A influência do tamanho do *corpus* de referência na obtenção de palavras-chave. *DIRECT Papers, 38*. [*Online*] http://lael.pucsp.br/direct

____. (1999b) Word sets, keywords, and text contents: an investigation of text topic on the computer. *Delta, 15*, p. 141-9.

____. (2000) Lingüística de *Corpus*: Histórico e problemática. *Delta, 16*, p. 323-67.

____. (2003) Tamanho da web em português. Manuscrito inédito. LAEL, PUC-SP.

BERNERS-LEE, T. (1999) *Weaving the Web*. London: Orion.

CORLISS, P. (2001) Number of pages on the Internet. Discussão na lista *Corpora*, disponível em http://www.hit.uib.no/corpora/2001-4/0164.html.

CRYSTAL, D. (2001) *Language and the Internet*. Cambridge: Cambridge University Press.

GREFENSTETTE, G.; NIOCHE, J. (2000) Estimation of English and non-English Language Use on the WWW. Proceedings of RIAO'2000, "Content-Based Multimedia Information Access", Paris, April 12-14, 2000, p. 237-46.

HILGERT, J. G. (1997) Procedimentos de reformulação: a paráfrase. In: PRETI, D. (Ed.). *Análise de textos orais*. 3. ed. São Paulo: Humanitas, p. 103-28.

HOEY, M. (1997) From concordance to text structure: New uses for computer *corpora*. In: LEWANDOSWKA-TOMASZCZYK, B.; MELIA, P. J. (Eds.). *PALC'97 – Practical Applications in Language Corpora*. Lodz: Lodz University Press, p. 2-22.

HUNSTON, S. (2002) *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

HUNSTON, S.; Francis, G. (2000) *Pattern Grammar - A Corpus-Driven Approach to the Lexical Grammar of English*. Amsterdam/ Philadelphia: John Benjamins.

KILGARRIFF, A. (2001) Web as corpus. In P. Rayson, A. Wilson, T. McEnery, A. Hardie, & S. Khoja (Eds.). *UCREL Technical Papers: 13. Proceedings of Corpus Linguistics 2001 Conference*. Lancaster: UCREL, University of Lancaster, p. 342-4.

LAWRENCE, S.; GILES, C. L. (1999) Accessibility of information on the web. *Nature*, *400*, p. 107.

PRETI, D. (Ed.) (1997) *Análise de textos orais*. 3. ed. São Paulo: Humanitas.

RENOUF, A. (2001) The Web as a source of linguistic information. In: Rayson, P.; Wilson, A.; MCENERY, T.; HARDIE, A.; KHOJA, S. (Eds.). *UCREL Technical Papers: 13. Proceedings of Corpus Linguistics 2001 Conference*. Lancaster: UCREL, University of Lancaster, p. 492-3.

SCOTT, M. (1997) PC Analysis of key words: and key words. *System*, *25*, p. 233-45.

____. (1998) *WordSmith Tools Version 3*. Oxford: Oxford University Press.

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

STUBBS, M. (1995) Collocations and semantic profiles: On the cause of trouble with quantitative studies. *Functions of Language*, *2*(2), p. 23-56.

____. (2001) *Words and phrases:* corpus-based studies of lexical semantics. Oxford: Routledge.

URBANO, H. (1997) Marcadores conversacionais. In: PRETI, D. (Ed.), *Análise de textos orais*. 3. ed. São Paulo: Humanitas, p. 81-102.