Estudo Comparativo das Palavras-chave do Campo das Ações Afirmativas no Português Brasileiro e no Inglês Americano[1]

EDVAN PEREIRA DE BRITO

Bacharel em Letras (Português/Linguística) (USP/SP). Mestrando do Programa de Meios de Comunicação de Massa e Estudos de Mídia na John H. Johnson School of Communications, Howard University (Washington, D.C., EUA).
epbrito@yahoo.com.br

1 O presente texto representa o resultado parcial de uma pesquisa realizada durante um período de intercâmbio na Howard University (Washington, D.C., EUA), sob a coordenação da Profª. Dra Stella E. O. Tagnin (DLM/FFLCH/USP). Este trabalho fez parte do programa "Raça, Desenvolvimento e Desigualdade Social", financiado pela CAPES/FIPSE. Apresentado sob a forma de Comunicação no IV CONGRESSO BRASILEIRO DE PESQUISADORES NEGROS - COPENE, no Simpósio II – Ações Afirmativas, Estado e Movimentos Sociais, realizado no período de 13-16 set. 2006 em Salvador-BA.

Resumo
As discussões a respeito das ações afirmativas têm influenciado a produção de uma vasta quantidade de textos. O estudo desse material poderia revelar os aspectos socioculturais que podem estar intricados nos diversos posicionamentos com relação a este tema. A Linguística de Corpus, por se enquadrar numa perspectiva teórica na qual a linguagem é vista como um sistema probabilístico, se configura como uma metodologia capaz de demonstrar, através de análises de corpora linguísticos, dados importantíssimos acerca da cultura dos grupos que produziram os textos que compõem os corpora. Diante disso, o objetivo do presente trabalho é fazer um estudo comparativo das palavras-chave do campo das ações afirmativas no português brasileiro e no inglês americano e, a partir das análises quantitativa e qualitativa dos resultados dessas listas de palavras, identificar alguns dos aspectos socioculturais que favoreceram a ocorrência maior de determinadas palavras em uma língua ou outra, fornecendo assim dados objetivos para fomentar o debate sobre este tema. Por outro lado, a compilação do corpus nos forneceu material de pesquisa que poderá servir de base para uma série de outros estudos em linguagem.

Palavras-chave
Linguística de corpus. Ação afirmativa. Cultura. Análise constrativa inglês-português.

Introdução

Tendo em vista o fato de o Brasil apresentar um dos piores níveis de desigualdade social do planeta (ZIMMERMANN; SPITZ, 2005), organizações internacionais, órgãos do poder público e diversos setores da sociedade civil organizada têm discutido a implementação e a gestão de políticas públicas capazes de minimizar o problema da má distribuição de renda, que é uma das maiores causas da exclusão social no país. Dentre as ações pensadas neste sentido, as políticas de ação afirmativa são as que apresentam a maior polaridade de opiniões, o que repercute em posicionamentos públicos a respeito do assunto, por meio de pronunciamentos orais ou escritos.

A observação desse polêmico debate foi o que nos orientou para a elaboração deste trabalho. Além disso, apesar da existência de muitos pesquisadores que estudam as ações afirmativas sob a ótica das diversas áreas das ciências humanas, são poucos os estudiosos que, no Brasil, trabalham com essa temática pelo viés linguístico, a partir do estudo de corpora. Imaginamos, então, que uma análise dos textos produzidos nessa área poderia fornecer dados relevantes para fomentar as discussões em torno do assunto.

Diante disso, o presente trabalho foi organizado da seguinte forma: na seção 1, apresentamos a Linguística de Corpus e alguns dos estudos que serviram de base para a nossa pesquisa; a seção 2 foi dedicada à explicitação dos processos metodológicos relativos ao planejamento dos corpora utilizados neste estudo, assim como dos critérios de seleção, coleta e armazenamento dos textos que compuseram os corpora; na seção 3, apresentamos o software "Wordsmith Tools"

(SCOTT, 1999) e algumas de suas ferramentas, especialmente a "wordlist" (lista de palavras) e a "keyword" (palavra-chave), estudando analiticamente seus respectivos resultados; por fim, na seção 4, recapitulamos brevemente os processos executados, refletindo sobre as questões abordadas nesta pesquisa.

1 A Linguística de Corpus

Pesquisadores das mais diversas áreas das ciências humanas se interessam por estudar as políticas de ação afirmativa. No entanto, são poucos os estudiosos que, no Brasil, trabalham com essa temática pelo viés linguístico, especialmente a partir do estudo de corpora. Sendo assim, por considerar que os processos próprios dos estudos linguísticos realizados no âmbito da Linguística de Corpus seriam muito úteis para o propósito do presente trabalho, é que resolvemos tomá-la como metodologia para a sua execução.

> A Lingüística de Corpus ocupa-se da coleta e da exploração de corpora, ou conjunto de dados linguísticos textuais que foram coletados criteriosamente, com o propósito de servirem para a pesquisa de uma língua ou variedade lingüística. Como tal, dedica-se à exploração da linguagem através de evidências empíricas, extraídas por meio de computador (BERBER SARDINHA, 2000, p. 325).

O corpus é, portanto, o principal objeto de pesquisa desse campo de estudo, cujas bases se centram em exemplos concretos de uso linguístico (McENERY; WILSON, 2001). Mas, mesmo no âmbito da Linguística de Corpus, é possível encontrar várias definições para corpus. Entretanto, a que engloba todas as características essenciais para um trabalho de descrição linguística como o nosso parte da concepção de corpus como sendo:

> 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 (SANCHES; CANTOS, 1996, p. 8-9 apud BERBER SARDINHA, 2004).

Então, é com o propósito de descrever e/ou analisar os diversos aspectos concernentes ao uso linguístico e suas várias correlações, que, no âmbito da Linguística de Corpus, inúmeros

estudos vêm sendo feitos há muitos anos, todos baseados em corpora ou "coleções de textos, palavras, frases, trechos, diálogos, etc.," (TAGNIN; TEIXEIRA, 2004, p. 320) que, por sua vez, são exemplos concretos de uso da língua e podem conservar aspectos socioculturais relevantes a respeito dos grupos que a utilizam. Nesta perspectiva, encontramos pelo menos três trabalhos que serviram de base para esta pesquisa.

Um deles foi o estudo de Geoffrey Leech e Roger Fallon (1992), no qual, por meio do estudo comparativo das frequências das palavras do Brown Corpus e do LOB (Lancaster-Oslo/Bergen Corpus), descreveram dados interessantes a respeito dos aspectos sociais, políticos e culturais das línguas dos países que deram origem aos corpora, respectivamente, inglês americano e inglês britânico. Nesse mesmo trabalho, os autores apontam o estudo "Word frequencies in British and American English", de Hofland e Johansson, publicado em 1982, como a primeira tentativa desse tipo de abordagem.

Outro estudo bastante significativo usando palavras-chave foi apresentado por Michael Stubbs (1996), analisando o inglês britânico. O ponto de partida do estudo foi o fato de que é possível verificar padrões de uso das palavras, seja em suas frequências ou em suas colocações – "palavras que co-ocorrem em frequência maior do que se se tratasse de uma combinação aleatória" (TAGNIN, 2002, p. 194). Assim, existe uma probabilidade grande de tais padrões incorporarem e expressarem valores sociais particulares e visões de mundo. Stubbs objetivou, então, mostrar como a análise linguística por meio do estudo de corpora linguísticos pode auxiliar na descrição e análise dos elementos culturais presentes nos usos de uma dada língua.

Com uma proposta bem parecida, Mike Scott (1997), em "PC analysis of key words – and key key words", fez um estudo de palavras-chave em um corpus composto de 5000 textos publicados no jornal britânico "The Guardian" entre os anos de 1992 e 1994. Seu objetivo foi, a partir da observação e análise dos dados quantitativos e qualitativos apresentados pelo software "WordSmith Tools", desenvolver formas para identificar os aspectos culturais que estão por trás desses textos. Ilustrando com a descrição das palavras-chave (*keywords*), das palavras-chave chave (*key keywords*), dos associados (*associates*) e destes em *clumps*, o autor chegou ao que ele chamou de *schemata* ou *stereotype*, ou rede de ligações entre ideias determinadas socialmente.

Tais estudos representaram fortes evidências de que seria possível experimentar tais estratégias em outras línguas, tanto em um corpus geral como em um corpus mais específico, como, por exemplo, um corpus da área das ações afirmativas. Sendo assim, nossa proposta foi apresentar um estudo comparativo das palavras-chave do campo das ações afirmativas no português brasileiro e no inglês americano, descrevendo e analisando, o que os dados linguísticos podem indicar sobre o universo sociocultural das línguas que dão origem aos textos dos corpora.

2 Metodologia

2.1 Planejamento do corpus

Depois de considerar a possibilidade de perceber algumas das manifestações socioculturais na linguagem empregada pelos falantes de português brasileiro e inglês americano no que tange à questão das políticas de ação afirmativa, o passo seguinte foi pensar no planejamento de um corpus que pudesse dar conta de explicitar tais diferenças e/ou semelhanças, isto é, que fosse representativo da parte do sistema linguístico que compreende os discursos a respeito de tais políticas. Consideramos então que poderíamos coletar dois corpora de 100 mil palavras em cada língua, o que caracterizaria um corpus pequeno-médio, de acordo com Berber Sardinha (2004). Diante disso, a partir dos critérios para construção de corpus seguidos pelo Projeto COMET (Corpus Multilingüe para Ensino e Tradução), descritos por Pardo (2004) e Tagnin e Teixeira (2004), estabelecemos os pontos norteadores para essa nossa proposta, que foram os seguintes:

2.1.1 Quanto ao tipo de corpus

Como o objetivo deste trabalho era fazer um estudo contrastivo das palavras-chave da área das ações afirmativas no português brasileiro e no inglês americano, optamos por montar um corpus comparável bilíngue, isto é, que contivesse textos originais nessas duas línguas e que fossem apenas da área das ações afirmativas, como também do mesmo gênero. Neste último caso, só entrariam textos de duas categorias: 1) gênero científico, ou seja, textos escritos por especialista para especialista (textos acadêmicos); 2) gênero jornalístico - textos escritos por especialista para um público-alvo de não-especialistas (textos informativos).

2.1.2 Seleção dos textos

Os textos que comporiam os corpora seriam coletados em fontes idôneas, tais como, revistas especializadas, jornais, *sites* de universidades e de associações profissionais ou organizações que discutissem ações afirmativas. Procuramos, nesse caso, o maior grau de variabilidade possível com relação às fontes, mas alguns problemas dificultaram o alcance dessa meta, como, por exemplo, certas restrições ao acesso do arquivo de alguns jornais e revistas. Textos sem qualidade gramatical e ortográfica e que não apresentavam todos os dados necessários para a sua identificação (vide item 2.1.3.1), como a data de publicação, autoria, etc., não entraram no corpus.

Optamos também por delimitar o período de publicação dos textos. Assim, só foram incluídos textos publicados entre os anos de 2000 e 2005 porque, no Brasil, esse período representou um marco na história da luta dos movimentos sociais negros por melhores condições de acesso para a população negra e outros grupos discriminados historicamente.

Decidimos só coletar textos disponíveis eletronicamente, isto é, nenhum dos textos foi escaneado ou digitado. Com relação ao modo pelo qual esses textos foram pesquisados na Internet, as principais ferramentas de busca utilizadas foram o Google[2] no caso dos textos jornalísticos e o Google Scholar[3] para pesquisar especialmente os textos científicos em língua inglesa. Vale lembrar que recorremos muitas vezes diretamente aos arquivos digitais de jornais, revistas e universidades. Tal atitude partiu da percepção de que isso nos pouparia tempo, como também verificamos que nem sempre as ferramentas de busca citadas recuperavam textos mais antigos ou de *sites* pouco visitados. Contudo, em ambas as estratégias tentamos focalizar a busca utilizando sempre como palavras-chave juntas ou separadas as palavras da expressão "ação afirmativa" ou "ações afirmativas" no caso do português, e "affirmative action" no caso no inglês.

Dentre as principais fontes dos textos que deram origem aos nossos corpora, podemos destacar, no caso do português: *O Estado de São Paulo*, *O Globo*, *Folha Online*, *Estudos Afro-Asiáticos*, *Com Ciência*/SBPC-Labjor Brasil, *Revista Espaço Acadêmico*, *Universidade de São Paulo*, *Universidade de Brasília*, *Cadernos de Pesquisa*, etc. No caso do inglês, as principais fontes foram: *USA Today*, *The Wall Street Journal*, *Time Magazine*, *NAACP News*, *The Washington Post*, *Review of Higher Education*, *Newsweek Magazine*, *Civilrights.org*, *Negro Education Review*, *Detroit Free Press*, etc.

2.1.3 Coleta e armazenamento do corpus

Para fins de identificação, depois de serem obtidos através do recurso copiar/colar, cada texto foi gravado em extensão ".txt", nomeado de acordo com a sua origem (por exemplo, aaf01.ia ou aac01.pb) e recebeu um cabeçalho[4] com etiquetas (ver próximo item) que possibilitariam um maior controle e acesso aos dados do corpus.

2.1.3.1 Cabeçalho

O cabeçalho traz informações que categorizam cada texto, permitindo um melhor acesso aos dados até mesmo em futuras pesquisas (TAGNIN; TEIXEIRA, 2004, p. 327). No caso deste trabalho, utilizamos os seguintes campos:

2 <www.google.com>.

3 <www.google.scholar.com>. Na época da coleta do corpus, o Google não dispunha de uma ferramenta como essa para o português.

4 O cabeçalho utilizado no corpus foi o mesmo apresentado no curso sobre o *WordSmith Tools* ministrado pela Dra. Elisa Duarte Teixeira no segundo semestre de 2004 e oferecido pelo Serviço de Cultura e Extensão Universitária da FFLCH/USP.

- <tit> título do texto </tit>
- <filename> nome do arquivo, ex: aac01.pb </filename>
- <subcorpus> "divulgação" para os textos jornalísticos ou "científico" para os acadêmicos (artigos) </subcorpus>
- <language> "português do Brasil" ou "inglês americano" </language>
- <mode> "Internet" (não é o nosso caso, mas poderia ser "livro", "revista", etc) </mode>
- <publisher> nome do *site* ou empresa que o mantém </publisher>
- <editor> nome do editor do *site* (pessoa ou empresa), quando identificado </editor>
- <pubDate> data de publicação </pubDate>
- <pubPlace> endereço do texto na Internet </pubPlace>
- <accessDate> data de coleta do texto </accessDate>
- <comments> espaço reservado para comentários nossos </comments>
- <name> nome completo do(s) autor(es), quando mencionado </name>

O Quadro 1 mostra um exemplo de cabeçalho preenchido.

```
<Header>
        <title>
                <tit> Oportunidades para cotistas </tit>
                <filename> aaf43_pb </filename>
                <subcorpus> divulgação </subcorpus>
                                        </title>
        <sourceText>
                <language> português do Brasil </language>
                <mode> Internet </mode>
                <publisher> Unb Agência </publisher>
                <editor> Assessoria de Comunicação/UNB </editor>
                <pubDate> 17/05/2005 </pubDate>
                <pubPlace> http://www.unb.br/acs/unbagencia/ag0505-38.htm </pubPlace>
                <accessDate> 15/06/2005 </accessDate>
                <comments> inclusão racial </comments>
        </sourceText>
        <author>
                <name> DIEGO AMORIM </name>
        </author>
</Header>
```

Quadro 1 – Exemplo de cabeçalho utilizado para a identificação dos textos dos corpora

Com o objetivo de isolar determinadas partes de texto ou mesmo indicar os elementos não compatíveis com a linguagem do software "Bloco de Notas", que não aceita figuras, fotos, tabelas e outros elementos gráficos, foram inseridas também outras etiquetas no corpo dos textos nos lugares em que tais elementos apareciam no original. Assim, incluímos as seguintes etiquetas:

- <subtit> subtítulo dos textos, quando claramente expressos </subtit>
- <abstract> resumo em inglês, no caso dos textos científicos em português </abstract>
- <resumo> resumo em português </resumo>
- <keyword> palavras-chave em inglês e/ou em outra língua </keyword>
- <palchave> palavras-chave em português </palchave>
- <affiliation> informações gerais a respeito do autor do texto </affiliation>
- <bibl> bibliografia </bibl>
- <nota> notas explicativas colocadas geralmente no fim do texto </nota>
- <photo> fotos com numeração que corresponde à sequência em que elas aparecem nos textos, por exemplo, foto 1, foto 2, etc. </photo>
- <legphoto> legenda de fotos, tabelas e gráficos </legphoto>
- <table> tabelas numeradas de acordo com a sua sequência no texto </table>

Depois dessa etapa, os textos foram sendo agrupados em pastas correspondentes. Ao final do processo de coleta, tínhamos 54 textos originais em português e 75 textos originais em inglês americano. Em seguida, iniciamos o processo de análise dos corpora, utilizando o software "Wordsmith Tools" (SCOTT, 1999) que é "o mais completo e versátil conjunto de ferramentas para Linguística de Corpus" (BERBER SARDINHA, 2004, p. 16). Os procedimentos executados e seus respectivos resultados estão descritos na próxima seção.

3 Resultados e análise

A primeira ferramenta do "Wordsmith Tools" (SCOTT, 1999) utilizada foi a "wordlist", que fornece dados quantitativos e qualitativos a respeito do corpus por meio de listas de palavras, em 3 janelas diferentes: "uma contendo uma lista de palavras ordenadas por ordem alfabética, outra com uma lista classificada pela frequência das palavras, e uma terceira janela com estatísticas simples a respeito dos dados" (BERBER SARDINHA, 2004, p. 91). O processamento feito pela "Wordlist" nos mostrou que os 54 textos da pasta correspondente ao corpus de português e 75 textos da pasta de inglês contabilizavam aproximadamente 100 mil palavras em cada uma das línguas, o que era a nossa meta inicial. No entanto, observando a lista alfabética do corpus de inglês, percebemos que aparecia a palavra "divulgação", o que era um indício de que o software

não estava restringindo a leitura dos dados apenas ao conteúdo dos textos, isto é, fazendo a leitura dos dados sem as etiquetas. Depois de algumas tentativas mal-sucedidas de resolução deste problema, retiramos dos corpora as etiquetas correspondentes ao cabeçalho e, no caso específico dos textos científicos do corpus de português, também as etiquetas que contivessem texto em inglês como, por exemplo, a bibliografia, o *abstract* e as *keywords* (vide seção 2.1.3.1). Desse modo, ao refazer a "Wordlist", obtivemos resultados diferentes. Assim, a lista de estatísticas nos indicou que tínhamos 88.104 palavras no corpus de português e 92.866 no corpus de inglês. Por fim, resolvemos então trabalhar com esse último resultado por considerá-lo mais confiável.

Com relação à desproporção no número de textos (54 x 75), o fato deveu-se à dificuldade de encontrar textos em português que atendessem a todos os critérios adotados. Além disso, nessa língua, geralmente os textos encontrados eram muito mais longos se comparados aos textos do mesmo tipo escritos em inglês. De qualquer forma, isso não representou um problema para a leitura dos dados, como pode ser mostrado ao observamos o *type/token ratio* ou a razão vocábulo/ocorrência da "wordlist"[5], que indica a riqueza lexical dos textos do corpus. Cada palavra do corpus corresponde a um *token*, já *type* é cada palavra distinta que pode se repetir 2, 10 ou 100 vezes até perfazer o total de *tokens* (TAGNIN; TEIXEIRA, 2004, p. 342). Na frase "a menina pegou a maçã que estava sobre a mesa", temos então 10 *tokens* (palavras) e 8 *types* (palavras distintas). O cálculo feito pelo *Wordsmith Tools* se baseia na seguinte regra de 3:

$$x = \frac{(\text{no. de } \textit{types}) \times 100}{\text{no. de } \textit{tokens}}$$

No exemplo acima, a razão seria de 80% (8 x 100 ÷ 10), isto é, para cada 100 palavras, 80 são distintas e 20 se repetem. No caso da nossa pesquisa, o corpus de português apresentou 88.104 *tokens* e 10.838 *types*, já o de inglês tinha 92.866 *tokens* e 8.895 *types*. Portanto, nos corpora analisados, o português apresentou 1.943 palavras distintas a mais que o inglês.

Depois desse primeiro processamento, verificamos então que o corpus de inglês apresentava um índice de riqueza lexical de 9,58%, contra 12,30% do corpus de português. Sendo assim, como vimos, o corpus de inglês apresenta o maior número de palavras (*tokens*), mas o de português apresenta o número de palavras distintas (*types*). De qualquer modo, essa inversão pode ser explicada pelo fato de que "o aumento do *type/token ratio* é inversamente proporcional ao aumento de *types*; isto é, quanto mais palavras no corpus, maior a probabilidade de repetição" (BIBER, 1993 apud TAGNIN; TEIXEIRA, 2004). Depois de executar tais procedimentos utilizando a ferramenta "wordlist", passamos a fazer a análise das palavras-chave.

5 Janela de estatísticas.

3.1 As palavras-chave

Palavras-chave (*keywords*) podem ser definidas como palavras que têm presença ou ausência estatisticamente significativa em um determinado texto ou conjunto de textos. De acordo com Scott (1997), essa frequência não está necessariamente ligada à quantidade, já que são ignoradas as palavras que geralmente têm uma frequência absoluta incomum, ou seja, são listadas as palavras que, num conjunto de textos de uma determinada área, assumem uma posição de destaque neste corpus, se compararmos com um corpus mais geral.

O primeiro passo para se chegar às palavras-chave de um corpus utilizando o "WordSmith Tools" é a criação de duas listas de palavras (*wordlist*), como as que mencionamos na seção anterior: uma contendo a *wordlist* do corpus de estudo, o qual se pretende descrever, como é o caso do nosso corpus da área das ações afirmativas; e outra contendo a *wordlist* de um corpus de referência, que será interpretado pelo programa como parâmetro para a comparação das frequências do corpus de estudo[6]. Tal procedimento pode ser mais bem entendido de acordo com o que diz Berber Sardinha (2004, p. 97) a este respeito:

> A comparação é feita por meio de uma prova estatística selecionada pelo usuário (qui-quadrado ou *log-likehood*). *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.

Em linhas gerais, a frequência de cada palavra é contrastada no corpus de estudo e no de referência. Se a ocorrência de uma determinada palavra é proporcionalmente maior no primeiro, é bem provável que ela seja chave. Assim, hipoteticamente, se a palavra "maçã" têm uma ocorrência de 3,5 % no corpus de estudo, e 0,5% no corpus de referência, então "maçã" ocorre 7 vezes mais no corpus de estudo, sendo provavelmente chave neste corpus.

Para maior eficiência na leitura dos dados, ao usar a ferramenta "keyword", foi estabelecido o valor mínimo de significância, ou valor de p (p value), em 0,0001, o que equivale a dizer que existe 1 probabilidade em 10.000 da ocorrência de erro no nosso cálculo. Estabelecemos 50 como o número máximo de palavras-chave que o programa listaria, sendo que só poderiam constar nesta lista palavras cuja frequência mínina não fosse menor que 3 ocorrências. Feito isso, as 50 primeiras palavras-chave em ordem de chavicidade no corpus de português brasileiro foram listadas (Tabela 1).

6 No caso desta pesquisa, os corpora de referência utilizados foram os American National Corpus (ANC) para o inglês e o Lácio-ref (CR-LW – Corpus de Referência Lácio-Web) para o português.

Tabela 1 – Lista das palavras-chave do corpus em português

N	PALAVRA	FREQ.	CORPUS DE ESTUDO (%)	FREQ.	CORPUS DE REFERÊNCIA (%)	CHAVICIDADE	VALOR DE P
1	NEGROS	438	0,49	548		2.537,7	0,000000
2	RACIAL	340	0,38	154		2.401,5	0,000000
3	COTAS	297	0,33	44		2.368,7	0,000000
4	IGUALDADE	240	0,27	190		1.539,3	0,000000
5	AFIRMATIVA	193	0,22	32		1.525,5	0,000000
6	AFIRMATIVAS	171	0,19	7		1.455,1	0,000000
7	RACIAIS	181	0,20	44		1.381,2	0,000000
8	RACISMO	164	0,18	123		1.062,9	0,000000
9	DISCRIMINAÇÃO	169	0,19	165		1.037,2	0,000000
10	NEGRO	241	0,27	848	0,01	1.003,1	0,000000
11	RAÇA	164	0,18	309		848,9	0,000000
12	DESIGUALDADES	112	0,13	94		709,8	0,000000
13	AÇÕES	211	0,24	1.345	0,02	665,3	0,000000
14	NEGRA	128	0,14	306		614,1	0,000000
15	BRANCOS	113	0,13	210		587,2	0,000000
16	POLÍTICAS	186	0,21	1.337	0,02	548,4	0,000000
17	COTA	80	0,09	45		545,9	0,000000
18	AÇÃO	215	0,24	2.506	0,03	459,9	0,000000
19	SOCIAL	246	0,28	4.118	0,06	384,3	0,000000
20	UNIVERSIDADES	145	0,16	1.275	0,02	378,0	0,000000
21	VAGAS	122	0,14	925	0,01	348,6	0,000000
22	SOCIEDADE	188	0,21	3.077	0,04	300,2	0,000000
23	VESTIBULAR	71	0,08	245		297,7	0,000000
24	AFRO	52	0,06	73		292,3	0,000000
25	POPULAÇÃO	159	0,18	2.241	0,03	291,2	0,000000
26	COR	97	0,11	703		284,6	0,000000
27	DESIGUALDADE	59	0,07	156		273,4	0,000000
28	UNB	47	0,05	61		269,6	0,000000
29	PARDOS	38	0,04	20		262,1	0,000000
30	LEI	146	0,16	2.192	0,03	252,7	0,000000
31	PRINCíPIO	95	0,11	819	0,01	250,8	0,000000
32	RESERVA	62	0,07	246		245,4	0,000000
33	PÚBLICAS	99	0,11	1.010	0,01	233,4	0,000000
34	AFRODESCENDENTES	37	0,04	42		219,3	0,000000
35	DIREITOS	93	0,10	949	0,01	219,3	0,000000
36	SOCIAIS	125	0,14	1.865	0,03	217,6	0,000000
37	OPTANTES	25	0,03	1		212,8	0,000000
38	CRITÉRIO	73	0,08	547		210,0	0,000000
39	BRASIL	268	0,30	7.836	0,11	206,9	0,000000
40	UERJ	31	0,03	26		196,5	0,000000
41	UNEB	24	0,03	3		193,7	0,000000
42	MINORIAS	37	0,04	68		192,9	0,000000
43	QUOTAS	33	0,04	42		190,2	0,000000
44	POBREZA	51	0,06	246		185,0	0,000000
45	MITO	47	0,05	196		182,1	0,000000
46	COTISTAS	20	0,02	0		177,0	0,000000
47	AFFIRMATIVE	22	0,02	3		176,5	0,000000
48	DISCRIMINAÇÕES	24	0,03	10		171,5	0,000000
49	RACISTA	30	0,03	42		168,8	0,000000
50	CONSTITUCIONAL	40	0,04	137		168,2	0,000000

Como foi dito acima, as palavras estão enumeradas de acordo com a ordem de chavicidade. Então, da esquerda para a direita: na terceira coluna temos o número de vezes que a palavra ocorreu no corpus de estudo, seguido da sua porcentagem (coluna 4); nas quinta e sexta colunas temos, respectivamente, o número e a porcentagem de ocorrência da palavra no corpus de referência; a penúltima coluna traz o valor resultante do processo de comparação dos dois corpora, o que se denomina chavicidade; e, na última coluna, temos o valor de p. Vale lembrar que espaços em branco representam valores abaixo de 0,01. Os resultados do corpus de inglês com relação às palavras-chave seguem essas mesmas especificações. Assim, as 50 palavras-chave do inglês foram listadas (Tabela 2).

Os resultados apresentados sugerem que, apesar de pertencerem ao mesmo domínio, os contextos de produção dos textos dos corpora são bem diferentes. Por outro lado, ao contrastar as duas listas apresentadas, percebemos que algumas palavras são comuns (ou correspondentes) aos dois corpora. Entretanto, mesmo considerando que haja correspondência de uso dessas palavras, é importante refletir sobre as especificidades de emprego delas nos textos, assim como questões de ordem morfológica, sintática e semântica. As palavras que seriam correspondentes aos dois conjuntos de textos estão descritas na Tabela 3.

Observamos, então, que o discurso em torno das políticas de ação afirmativa nos dois contextos estudados tende a focalizar a questão racial, sobretudo a relação entre brancos, negros e os grupos considerados minoritários. A ocorrência de palavras como "lei/ law", "direitos/ rights", "discriminação(ões)/ discrimination" e "universidades/ university(ies)/ college(s)" nos indica também a relação que tais políticas têm com a questão legal e/ou constitucional do debate, especialmente voltada para o ensino superior. Além disso, em termos absolutos, a frequência da maior parte das palavras em inglês supera a de suas correspondentes em português.

No caso das palavras que só aparecem no português, a lista das palavras-chave aponta para um cenário cujos pontos centrais são a discussão a respeito de cotas para os negros nas universidades. As palavras seriam as seguintes: cotas (297) /cota (80) /quotas (33); igualdade (240); racismo (164); desigualdades (112) / desigualdade (59); social (246) / sociais (125); vagas (122); sociedade (188); vestibular (71); população (159); cor (97); UNB (47); pardos (38); princípio (95); reserva (62); públicas (99); optantes (25); critério (73); Brasil (268); UERJ (31); UNEB (24); pobreza (51); mito (47); cotistas (20); racista (30); constitucional (40).

A ocorrência de palavras como "cota(s)", "igualdade", "racismo", "vagas", "vestibular", "UNB", "reserva", "públicas", "cotistas" parece reforçar a ideia de que a adoção de cotas no ensino superior foi a grande questão que permeou o debate sobre ação afirmativa entre 2000 a 2005. Outra constatação interessante é o aparecimento de palavras que estão no cerne da questão das relações raciais no Brasil, tais como, "cor", "pardos", "mito", dentre outras.

Tabela 2 – Lista das palavras-chave do corpus em inglês

N	PALAVRA	FREQ.	CORPUS DE ESTUDO (%)	FREQ.	CORPUS DE REFERÊNCIA (%)	CHAVICIDADE	VALOR DE P
1	AFFIRMATIVE	746	0,80	342		6.559,2	0,000000
2	ACTION	758	0,82	3.136	0,02	4.230,2	0,000000
3	STUDENTS	396	0,43	20.458	O,01	1,923,6	0,000000
4	DIVERSITY	302	0,33	798		1,915,9	0,000000
5	RACE	368	0,40	2.208	0,01	1,809,8	0,000000
6	ADMISSIONS	239	0,26	312		1,782,0	0,000000
7	RACIAL	258	0,28	525		1,747,1	0,000000
8	MICHIGAN	226	0,24	395		1,584,8	0,000000
9	UNIVERSITY	340	0,37	3.445	0,02	1.351,0	0,000000
10	MINORITIES	166	0,18	190		1.269,2	0,000000
11	MINORITY	210	0,23	735		1.231,8	0,000000
12	COURT	321	0,35	3.684	0,02	1.203,3	0,000000
13	COLLEGES	138	0.15	199		1.008,3	0,000000
14	EDUCATION	231	0,25	2.261	0.O1	932,1	0,000000
15	PREFERENCES	122	0,13	188		879,1	0,000000
16	UNIVERSITIES	125	0,13	258		843,4	0,000000
17	SUPREME	162	0,17	955		801,5	0,000000
18	BLACK	249	0,27	4.548	0,02	726,4	0,000000
19	COLLEGE	182	0,20	2.207	0,01	663,9	0,000000
20	APPLICANTS	94	0,10	165		658,4	0,000000
21	MICHIGAN'S	64	0,07	14		604,9	0,000000
22	SCHOOL	240	0,26	5.724	0,03	588,5	0,000000
23	DISCRIMINATION	106	0,11	484		572,5	0,000000
24	AFRICAN	127	0,14	1.074		545,9	0,000000
25	O'CONNOR	71	0,08	81		543,1	0,000000
26	BAKKE	52	0,06	7		508,1	0,000000
27	JUSTICE	155	0,17	2.359	0,01	502,3	0,000000
28	LAW	211	0,23	5.556	0,03	481,6	0,000000
29	RIGHTS	153	0,16	2.463	0,01	480,5	0,000000
30	UNDERGRADUATE	63	0,07	84		467,6	0,000000
31	WHITE	226	0,24	6.804	0,04	464,9	0,000000
32	STUDENT	120	0.13	1.285		464,6	0,000000
33	POLICIES	113	0,12	1.084		459,9	0,000000
34	BLACKS	91	0,10	543		448,2	0,000000
35	EDUCATIONAL	88	0.09	524		433,7	0,000000
36	UNDERREPRESENTED	46	0,05	16		416,8	0,000000
37	SCHOOLS	119	0,13	1.579		414,8	0,000000
38	CIVIL	122	0,13	1.782		404,2	0,000000
39	INSTITUTIONS	82	0,09	550		386,7	0,000000
40	COURT'S	64	0,07	219		377,8	0,000000
41	PREFERENCE	66	0,07	299		357,3	0,000000
42	WHITES	67	0,07	336		350,7	0,000000
43	JUSTICES	50	0,05	89		349,1	0,000000
44	DECISION	126	0,14	2.622	0,01	338,6	0,000000
45	FACULTY	63	0,07	302		334,8	0,000000
46	AMERICANS	127	0,14	2.819	0,02	327,0	0,000000
47	PROGRAMS	137	0,15	3.434	0,02	324,1	0,000000
48	ADMISSION	70	0,08	530		314,8	0,000000
49	GRUTTER	29	0,03			307,3	0,000000
50	AMERICAN	211	0.23	9,507	0,05	297,7	0,000000

Tabela 3 – Lista de palavras correspondentes aos dois corpora

PALAVRA(S) EM PORTUGUÊS / FREQUÊNCIA(S) NO CORPUS	PALAVRA(S) EM INGLÊS / FREQUÊNCIA(S) NO CORPUS
NEGROS (438); NEGRO (241); NEGRA (128)	BLACK (249); BLACKS (91)
RACIAL (340); RACIAIS (181)	RACIAL (258)
AÇÕES (211); AÇÃO (215)	ACTION (758)
AFIRMATIVA (193); AFIRMATIVAS (171); AFFIRMATIVE (22)[1]	AFFIRMATIVE (746)
DISCRIMINAÇÃO (169); DISCRIMINAÇÕES (24)	DISCRIMINATION (106)
RAÇA (164)	RACE (368)
UNIVERSIDADES (145)	UNIVERSITY (340); UNIVERSITIES (125); COLLEGES (138); COLLEGE (182)[2]
POLÍTICAS (186)	POLICIES (113)
AFRO (52); AFRODESCENDENTES (37)	AFRICAN (127)
MINORIAS (37)	MINORITIES (166); MINORITY (210)
BRANCOS (113)	WHITE (226); WHITES (67)
LEI (146)	LAW (211)
DIREITOS (93)	RIGHTS (153)

(1) Em função do tempo, não nos foi possível verificar o contexto de ocorrência desta palavra no corpus, mas a nossa hipótese é de que se trata de referências feitas ao sistema norte-americano de políticas de ação afirmativa (Affirmative Action).

(2) Apenas por uma questão semântica "College" e Colleges" foram colocadas em comparação com "universidade(s)".

O corpus de inglês também apresenta dados interessantes. Os itens que aparecem como palavras-chave apenas nessa língua são: students (396)/ student (120); diversity (302); admissions (239)/ admission (70); Michigan (226)/ Michigan's (64); Court (321)/ Court's (64); education (231); preferences (122)/ preference (66); Supreme (162); applicants (94); school (240)/ schools (119); O'Connor (71); Bakke (52); justice (155)/ justices (50); undergraduate (63); educational (88); underrepresented (46); civil (122); institutions (82); decision (126); faculty (63); Americans (127)/ American (211); programs (137); Grutter (29).

Temos, então, a indicação de que, pelo menos no período escolhido para esta pesquisa, foi recorrente a discussão em torno das decisões da Suprema Corte norte-americana com relação às

ações contra a Universidade de Michigan. Poderíamos afirmar, portanto, que a discussão esteve voltada para o campo jurídico. Além disso, enquanto no português tiveram destaque algumas das universidades que foram pioneiras na discussão sobre a adoção de reserva de vagas nos seus cursos de graduação (UERJ, UNB, UNEB), no inglês, se destacaram alguns dos casos históricos de decisões da Suprema Corte sobre ações afirmativas (Bakke, Grutter), assim como um dos ícones do caso das ações contra a Universidade de Michigan (Sandra Day "O'Connor"). Por fim, as duas listas de palavras-chave podem indicar também um outro fato que é bastante peculiar em cada um dos contextos socioculturais estudados. Trata-se da ocorrência da palavra "igualdade" entre as palavras-chave do português e da palavra "diversity/diversidade" entre as palavras-chave do inglês. Tais palavras carregam e sustentam boa parte dos argumentos em favor ou contra as políticas de ação afirmativa nos dois contextos, cabendo a cada pessoa a escolha pela abordagem que se pode dar a elas.

4 Considerações finais

A língua, vista aqui como um dos aspectos da cultura, é capaz, através da observação dos usos, de nos dar uma demonstração das visões de mundo presentes nas comunidades que se utilizam dessas línguas. Essa foi uma das principais ideias que permearam as nossas reflexões enquanto realizamos o presente estudo. Dessa forma, com o auxílio dos processos teórico-metodológicos da Linguística de Corpus pudemos analisar a parte do sistema linguístico que corresponde ao campo das ações afirmativas, no português brasileiro e no inglês americano. Para tanto, tivemos a necessidade de compilar os dois corpora, já que não havia, pelo menos até então, corpora tão específicos e que fossem apropriados aos nossos propósitos. Posteriormente, utilizamos o software "Wordsmith Tools" (SCOTT, 1999) que, através das suas ferramentas de análise linguísticas, nos forneceu uma série de dados, não só sobre as duas línguas em questão, mas também sobre o contexto sociocultural dos falantes dessas duas línguas.

Entretanto, no decorrer do nosso estudo, percebemos que muitas poderiam ser as possibilidades de abordagem dos corpora, mas mantivemos o nosso foco no estudo comparativo das palavras-chave por não dispormos de tempo e recursos para levar adiante estudos muito mais aprofundados, inclusive verificando os contextos de uso/ocorrência de algumas das palavras mais frequentes, utilizando assim outras ferramentas do programa. De qualquer modo, como as observações aqui feitas não esgotam e nem têm a pretensão de esgotar as discussões sobre as políticas de ação afirmativa, o caminho está aberto para que outras leituras sejam feitas, assim como novas possibilidades de pesquisa sejam implementadas.

Comparative Study of Keywords in the Field of Affirmative Actions in the Brazilian Portuguese, and in the American English

Abstract
The current discussions on affirmative action have influenced the production of several texts on the subject. The study of these texts may reveal the socio-cultural aspects behind the different views on the theme. The Linguistics of Corpus, which fits a theoretical perspective in which language is seen as a probabilistic system, is a method capable of showing, through the analysis of linguistic corpora, relevant data about the culture of the groups that produced the texts of the corpora. Thus, this work intends to conduct a comparative study of affirmative action keywords in Brazilian Portuguese and American English. Through the quantitative and qualitative analysis of these lists of words, it also intends to identify the socio-cultural aspects that may explain a greater use of certain words in a language, which could provide data to encourage the debate on the issue. On the other hand, the compilation of the corpus will provide research material which could serve as a basis for a series of other language studies.

Keywords
Linguistics of Corpus. Affirmative action. Culture. Contrastive English-Portuguese analysis.

Referências

BERBER SARDINHA, Tony. Lingüística de corpus: histórico e problemática. **D.E.L.T.A.**, São Paulo, v. 16, n. 2, p. 323-367, 2000.

_____. **Lingüística de corpus**. Barueri: Manole, 2004.

LEECH, Geoffrey; FALLON, Roger. Computer corpora: what do they tell us about culture? **ICAME Journal**, n. 16, p. 29-50, 1992.

McENERY, Tony; WILSON, Andrew. **Corpus linguistics**. 2. ed. Edinburgh: Edinburgh University Press, 2001.

PARDO, Roberto Martinez. **Critérios de construção e organização de um corpus de especialidade**: o corpus técnico-científico de ortodontia. 2004. 135 f. Dissertação (Mestrado em Letras) – Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2004.

SCOTT, Mike. PC analysis of key words: and key key words. **System**, Liverpool, v. 25, n. 2, p. 233-245, 1997.

_____. **Wordsmith tools version 3**. Oxford: Oxford University Press, 1999.

STUBBS, Michael. **Text and corpus analysis**: computer-assisted studies of language and culture. Oxford: Blackwell, 1996.

TAGNIN, Stella Esther Ortweiler. Os corpora: instrumentos de auto-ajuda para o tradutor. **Cadernos de Tradução**, Florianópolis, n. 9, p. 192-219, 2002. Disponível em: <http://www.periodicos.ufsc.br/index.php/traducao/issue/view/432>.

_____; TEIXEIRA, Elisa Duarte. Lingüística de corpus e tradução técnica: relato da montagem de um corpus multivarietal de culinária. **TradTerm: Revista do Centro Interdepartamental de Tradução e Terminologia/FFLCH-USP**, São Paulo, n. 1, p. 313-358, 2004.

ZIMMERMANN, Patricia; SPITZ, Clarice. Brasil é oitavo país em desigualdade social, diz pesquisa. **Folha Online**, São Paulo, 7 set. 2005. Disponível em: <http://www1.folha.uol.com.br/folha/cotidiano/ult95u112798.shtml>. Acesso em: 8 ago. 2006.

Correspondência

EDVAN PEREIRA DE BRITO
Estrada Mato das Cobras, 79 - Ponte Alta
07179-701 - Guarulhos - SP
Fone: 1+202.904.1891
epbrito@yahoo.com.br

Recebido em 10.07.2009

Aprovado em 20.08.2009