FREQÜÊNCIA DE PALAVRAS: UM DIAGNÓSTICO DO VOCABULÁRIO DE REDAÇÕES DE VESTIBULAR

Roseli Imbernom do NASCIMENTO[1]
Aparecida Negri ISQUERDO[2]

■ RESUMO: Este artigo discute resultados de pesquisa léxico-estatística, realizada com base em um *corpus* de 450 redações de vestibular (1999 e 2000), de duas universidades do interior do Estado de São Paulo, uma pública e outra particular. Os resultados estatísticos dessa amostra do vocabulário de ingressantes universitários assemelham-se aos dados de outras pesquisas dessa natureza, sobretudo quanto à alta concentração da freqüência sobre um pequeno número de unidades lexicais: 11.151 unidades léxicas diferentes totalizaram 113.638 ocorrências do *corpus*. O confronto dos dados dessa pesquisa com os do *Dicionário de freqüências* do léxico do português brasileiro contemporâneo, de Biderman (1998), aponta um pequeno grupo de palavras (pouco mais de trezentos) comuns nas duas bases de dados e com índices de freqüência similares, o que indica a existência de um possível núcleo do vocabulário analisado que reúne palavras, provavelmente, usadas em qualquer tipo de texto. Esse fenômeno é marcado, sobretudo, nos vinte verbos mais freqüentes no *corpus*. Dados resultantes de recortes e de cruzamentos sociolingüísticos, a partir de variáveis como universidade de ingresso dos estudantes; conclusão do Ensino Médio em escola particular ou pública; sexo e renda familiar confirmam esse fenômeno. Dessa análise, dois fatores merecem particularmente destaque: a amplitude maior do vocabulário daqueles que ingressam na universidade pública e a "escolaridade" como variável que suplanta todas as demais.

■ PALAVRAS-CHAVE: Léxico-estatística; vocabulário; redações de vestibular; estudantes universitários; verbos.

Preliminares

Uma série de previsões e de constatações sobre o funcionamento da língua e sobre os elementos gramaticais presentes nos discursos orais ou escritos pode ser feita por meio da léxico-estatística, comprovando que "o quantitativo é uma das propriedades do vocabulário" e que "a freqüência é uma característica típica da palavra". Des-

1 Departamento de Letras das Faculdades Integradas Toledo – 16015-270 – Araçatuba – SP – Brasil. Endereço eletrônico: fersmile@terra.com.br.

2 Departamento de Comunicação e Expressão do Campus de Dourados – UFMS – 79825-070 – Dourados – MS – Brasil. Endereço eletrônico: aparecida.negri@uol.com.br.

venda-se, assim, a *norma lexical* vigente, descrita por Biderman (1998, p.162) como "a média dos usos freqüentes das palavras que são aceitas pela comunidade de falantes".

Entretanto, a língua não pode ser diretamente observável, por isso, de acordo com a equipe de pesquisadores portugueses que atuaram no projeto do *Português Fundamental*, quando se pretende verificar a freqüência de uso das palavras e a aferição da sua média, é necessário "tentar apreender e isolar, através de uma amostragem, o léxico comum de uma comunidade lingüística" (CRUZ, 1987, p.314).

Tomando por base parâmetros teórico-metodológicos fornecidos pela léxico-estatística, realizamos uma pesquisa lexicográfica (NASCIMENTO, 2001) que estudou uma amostragem do vocabulário de estudantes universitários, aqui referido por VEU, registrado em 450 redações de vestibular de duas instituições de ensino superior do interior do Estado de São Paulo, uma pública – Unicamp/Campinas, 170 redações do vestibular de 1999 – e uma particular – Faculdades Toledo/Araçatuba, 280 redações do vestibular de 2000 –, que tiveram como tema "A comemoração dos 500 anos do Brasil".

De um modo geral, os resultados apurados são bastante semelhantes aos dados obtidos por outros trabalhos voltados ao estudo da face quantitativa da linguagem e, de um modo muito especial, aos constantes do *Dicionário de freqüências* do léxico do português brasileiro contemporâneo, de Biderman (1998), doravante DIF, também elaborado mediante utilização de métodos estatísticos e computacionais. Nos dois trabalhos, observa-se que um determinado grupo de palavras – pouco mais de trezentos – registra um índice muito elevado de freqüência, ocupando, de forma quase idêntica, o topo das diversas listas de classificação. Esse fenômeno confirma as declarações de Biderman (1998, p.178-9) de que "por enorme que seja o léxico de uma língua, é reduzido o repertório efetivamente utilizado pelos falantes, até mesmo na língua escrita, que é a variante da língua que se serve de um vocabulário mais rico e mais variado".

Essas palavras com índice elevado de freqüência, possivelmente, constituem o núcleo do vocabulário das redações dos ingressantes universitários, fenômeno que corrobora a constatação de Biderman (1998, p.169) de que existem "palavras multiuso que aparecem em qualquer texto, independentemente de seu conteúdo temático".

Os procedimentos estatísticos não fornecem apenas dados objetivos, mas também indicam, por meio das escolhas de determinados itens lexicais feitas pelos falantes, a competência lingüística e a visão de mundo dos indivíduos. O fato de serem "escolhidas" determinadas palavras – e não outras – demonstra uma dada realidade vivida, uma vez que as palavras são capazes de testemunhar a história e de sintetizar o pensamento humano.

Neste artigo, pretende-se discutir parte dos resultados obtidos com a pesquisa em questão. Para tanto, o trabalho foi dividido em três partes: visão geral dos resultados da pesquisa e análise comparativa com os dados do DIF; análise do comportamento dos vinte verbos mais freqüentes no VEU; breve discussão de alguns resultados a partir de variáveis sociolingüísticas.

Visão geral do *corpus*

As pesquisas léxico-estatísticas não deixam dúvidas de que não existe uma distribuição homogênea da freqüência das palavras usadas nos textos, e de que o alto

índice da freqüência deve-se, fundamentalmente, a algumas poucas palavras, sobretudo às cinco primeiras. Esse fato ocorre em qualquer vocabulário que se analise e marca-se no *corpus* desta pesquisa, constituído de 11.151 unidades lexicais para um total de 113.638 ocorrências. A Tabela 1, na seqüência, comprova esse fato.

Tabela 1 – As cinco palavras mais freqüentes no VEU

Unidades/ocorrências	Freqüência acumulada
1ª - **de** = **4.693** ocorrências	
2ª - **que** = **3.959** ocorrências	**18.932**
3ª - **e** = **3.696** ocorrências	ocorrências
4ª - **a** = **3.358** ocorrências	**(17% do *corpus*)**
5ª - **o** = **3.226** ocorrências	

A alta freqüência registrada para um número tão ínfimo de palavras chama-nos a atenção: apenas cinco instrumentos gramaticais atingem, sozinhos, 17% da freqüência do *corpus*. Essa freqüência elevada torna-se ainda mais significativa quando se estabelece uma relação com as palavras de baixa freqüência e verifica-se que a freqüência assinalada por essas cinco primeiras palavras é três vezes maior que a de todas as unidades de freqüência 1 (os 5.813 *hapax legomena* – palavras que ocorrem uma só vez no *corpus*) (cf. BIDERMAN, 1998). Especialmente a soma das duas mais freqüentes é muito elevada (8.652 ocorrências). Confirma-se com isso o fenômeno do *dequeísmo*, ou tendência contemporânea de uso do "de" (preposição) e do "que" (pronome relativo), conforme argumentam Paiva & Scherre (1999, p.206).[3]

O levantamento e a contagem das unidades lexicais foram feitos com base nas "palavras gráficas", ou "formas flexionadas", efetivamente realizadas nos textos, tendo sido considerados dois paradigmas para a avaliação das palavras de maior ocorrência no *corpus*: o das cem palavras mais freqüentes e o das palavras de freqüência igual ou superior a quarenta ocorrências (F40) – limiar instituído pelos lingüistas que desenvolveram o projeto do *Português fundamental* (1987).

As cem palavras mais freqüentes no *corpus* somam 51,3% – mais da metade do total. Esse índice, que se concentra sobre menos de 1% do total de unidades léxicas, é bastante semelhante àquele obtido por Duncan (1972, apud BIDERMAN, 1978), quando elaborou o primeiro dicionário de freqüência do português[4] e constatou que as 100 primeiras palavras, referentes a 2% do total de unidades, registraram 61,98% da freqüência total.

3 Nesse trabalho, as autoras analisam o uso variável de determinadas preposições em diferentes processos de regência verbal. A respeito do "de" e do "que", Paiva & Scherre (1999, p.206) ponderam: "a instabilidade do sistema preposicional fica evidente ... na tendência à inserção da preposição 'de' em contextos em que não se prevê sua ocorrência – dequeísmo – ... ou de sua queda em contextos onde é esperada – queísmo".

4 *A Frequency Dictionary of Portuguese Words* (FDPW), de John C. Duncan Jr., PhD. Dissertation, Standford University, 1972, citado por Biderman (1978, p.265-72), selecionou as 5.000 palavras que mais freqüentemente ocorreram num *corpus* de 500.000 palavras.

O Quadro 1 apresenta as cem palavras mais freqüentes no *corpus*, divididas em duas categorias – palavras plenas e palavras instrumentais – e confirma a predominância do segundo grupo.

Quadro 1 – As cem palavras mais freqüentes no VEU

21 *palavras plenas* (substantivos e adjetivos)	**79 *palavras instrumentais*** (advérbios, artigos, conjunções, contrações, denotadores expressivos, numerais, pronomes, preposições e verbos auxiliares/instrumentais)
anos, Brasil, brasileiro, brasileiros, cara, cultura, descobrimento, grande, história, índios, melhor, mundo, nação, país, países, pessoas, população, portugueses, povo, problemas, terra.	a, à, ainda, ao, aos, aqui, as, assim, até, cada, com, como, da, das, de, desde, do, dos, e, é, em, essa, esse, está, este, foi, há, hoje, isso, já, mas, mais, mesmo, muito, na, não, no, nos, nós, nossa, nosso, nossos, o, onde, os, ou, outros, para, pela, pelo, pode, pois, por, quando, que, quem, quinhentos, são, se, sem, ser, será, seu, seus, só, sobre, somos, sua, suas, também, tão, tem, temos, ter, todos, tudo, um, uma, você.

O grupo de palavras de F40 soma 359 unidades lexicais, atingindo 66,7% da freqüência total registrada, e é constituído, fundamentalmente, por palavras instrumentais, como artigos, pronomes, preposições, conjunções, advérbios, e por alguns verbos bastante freqüentes.

Essas palavras podem ser consideradas possíveis representantes do núcleo do vocabulário das redações dos universitários, tomadas como objeto de análise, na fase de ingresso no Curso Superior.

Pelo Gráfico 1, constata-se a alta concentração da freqüência sobre esse grupo de palavras de F40, comparada à dos demais intervalos de freqüência registrados.

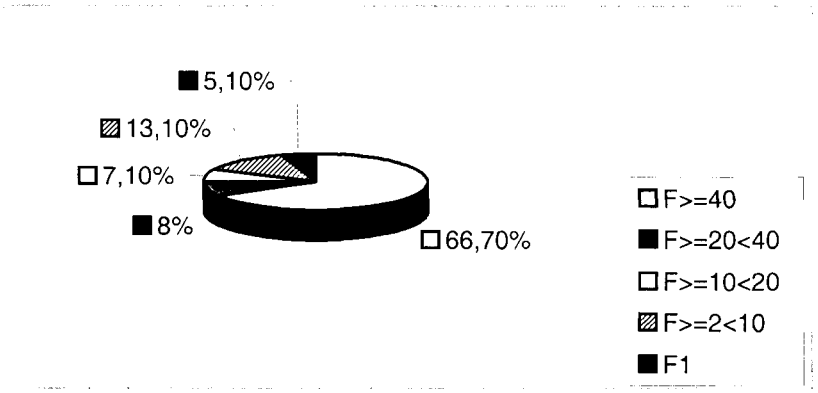

GRÁFICO 1 – Distribuição da freqüência das palavras por intervalos.

Na correlação desses dados com os do DIF, observa-se que, desse grupo das 359 palavras de freqüência igual ou superior a quarenta ocorrências, 308 palavras, ou seja, mais de 85% das mais freqüentes do *corpus* da pesquisa, coincidem com as mais fre-

qüentes do DIF.[5] Cremos, por isso, que essas 308 palavras legitimam o seu papel de representantes do vocabulário ora analisado. Talvez possam, ainda, por meio de resultados de pesquisas futuras, ser consideradas "multiuso", por aparecerem em qualquer texto.

Cerca de 25% dessas 308 palavras não estão "presas" à temática e ao gênero textuais, nem são de uso genérico. São, portanto, palavras plenas, ou de conteúdo nocional (substantivos, adjetivos e alguns verbos), as quais convencionamos chamar de *especiais*, que sintetizam o pensamento contemporâneo dos acadêmicos, pois correspondem ao que Matoré (1953) denominou de *mots témoins*, ou palavras-testemunhas da nossa história.

A freqüência assinalada por algumas dessas palavras atesta bem esse fato. Com o estudo das variáveis, observou-se, por exemplo, que, no vocabulário dos estudantes – frise-se, em uso nos anos de 1999 e de 2000 –, a palavra *desemprego* apresenta uma freqüência bastante alta, denotando a grande preocupação dos universitários com esse problema social existente no Brasil, atualmente. Na pesquisa de Biderman (1998), porém, que contemplou o período de 1950 a 1990, ela não figura entre as mais freqüentes, todavia, está presente o seu antônimo – a palavra *emprego* –, registrando um momento histórico oposto.

Sublinhamos, dessa forma, a importante "abonação" que esse *Dicionário* concede aos resultados da pesquisa.

Os verbos mais freqüentes

Nenhum vocabulário possui uma distribuição homogênea da freqüência das palavras, embora exista uma certa regularidade das estruturas léxicas. Esse fenômeno pode ser observado por meio da Tabela 2, a seguir, que visualiza a lista dos vinte verbos mais freqüentes, tanto no DIF quanto no vocabulário ora analisado. Os verbos extraídos do VEU foram lematizados e o número de ocorrências de cada uma das formas inclui todas as flexões existentes no *corpus*. O total de freqüência desses vinte verbos é de 9.549 ocorrências, que corresponde a 8,4% do *corpus*.

Como se observa, treze verbos – *ser, ter, estar, poder, fazer, haver, ver, saber, dever, ir, dizer, chegar* e *dar* – estão tanto na lista dos vinte primeiros do DIF quanto na lista dos vinte primeiros do VEU. Apenas sete verbos – *querer, ficar, achar, falar, precisar, começar* e *olhar* –, que figuram entre os vinte mais freqüentes do DIF, não estão entre os vinte do VEU. No lugar desses ausentes, constam, entre os vinte primeiros do vocabulário dos estudantes, os verbos *vir, viver, mostrar, possuir, existir, comemorar* e *descobrir*.

Cabem aqui dois tipos de considerações: uma sobre a semelhança dos treze verbos, outra sobre a dessemelhança dos sete. No primeiro caso, não nos surpreende a coincidência, uma vez que não só nos trabalhos desenvolvidos por Biderman (1998), mas também nas pesquisas realizadas em Portugal (*Português fundamental*) consta-

5 De acordo com Biderman (1998, p.161), foram consideradas mais freqüentes 1.078 palavras de F>=500. O *corpus* desse Dicionário foi constituído por cinco diferentes tipos de literatura: *romanesca* (LR), *dramática* (LD), *tecnocientífica* (LT), *jornalística* (LJ) e *oratória* (LO), compreendidas no período de 1950 a 1990.

tou-se que o comportamento desses verbos era parecido. Os dados do VEU retomam, pois, as seguintes declarações de Biderman (1998) acerca da conclusão a que chegou Müller, em 1974, ao pesquisar os vinte verbos mais freqüentes no francês – exatamente os mesmos da lista do Dicionário:

> esses vinte verbos mais freqüentes situam-se na escala decrescente de freqüência em posições quase idênticas; isso confirma também que distribucionalmente eles operam de maneira muito similar na língua, não importando o tipo de variáveis lingüísticas consideradas, a saber: língua falada ou escrita ... Os resultados demonstram, portanto, que o comportamento lingüístico desses verbos tem-se mantido quase imutável ao longo de duzentos anos. São, pois, verbos muito estáveis no idioma. (BIDERMAN, 1998, p.171)

Tabela 2 – Vinte verbos mais freqüentes nas bases de dados comparadas

Os vinte verbos mais freqüentes no DIF (*corpus*: 5 milhões)	**Os vinte verbos mais freqüentes no VEU (*corpus*: 113.638)[6]**
Classificação/verbo/ocorrências	*Classificação/verbo/ocorrências*
1º - ser [50.222]	1º - ser [3.770]
2º - ter [34.586]	2º - ter [1.006]
3º - ir [28.965]	3º - estar [702]
4º - estar [27.746]	4º - poder [567]
5º - poder [16.593]	5º - fazer [414]
6º - dizer [15.445]	6º - haver [330]
7º - haver [15.004]	7º - ver [293]
8º - fazer [14.279]	8º - vir [254]
9º - dar [10.792]	9º - saber [251]
10º - ver [10.391]	10º - viver [239]
11º - saber [10.247]	11º - dever [238]
12º - querer [9.986]	12º - ir [231]
13º - ficar [8.605]	13º - dizer [214]
14º - achar [7.980]	14º - mostrar [183]
15º - dever [7.758]	15º - possuir [159]
16º - falar [5.259]	16º - existir [148]
17º - chegar [4.628]	17º - chegar [147]
18º - precisar [4.039]	18º - comemorar [139]
19º - começar [3.596]	19º - descobrir [134]
20º - olhar [3.383]	20º - dar [130]

Esses resultados corroboram também a hipótese formulada pela mesma pesquisadora de que as conclusões de Müller sobre o francês talvez sejam válidas para o português. Veja-se, nesse caso, o fato de o primeiro e o segundo classificados – *ser* e *ter* – serem exatamente os mesmos, havendo apenas uma pequena variação na ordem de classificação dos demais.

6 Como se pode observar, o DIF baseou-se num *corpus* cinqüenta vezes maior que o do VEU. Assim, essa diferença proporcional deve permear quaisquer comparações estabelecidas.

A título de exemplificação do trabalho de lematização das formas verbais mais freqüentes no *corpus*, apresentamos na Tabela 3 todas as flexões dos cinco verbos mais empregados nas redações dos ingressantes universitários (ser, ter, estar, poder, fazer).

Tabela 3 – Lematização dos cinco verbos mais freqüentes no VEU

1º colocado: SER = 3.770 ocorrências	2º colocado: TER = 1.006 ocorrências	3º colocado: ESTAR= 702 ocorrências	4º colocado: PODER = 567 ocorrências	5º colocado: FAZER = 414 ocorrências
1. ser [340]	1. ter [178]	1. estar [42]	1. poder [82]	1. fazer [93]
2. é [1393]	2. tem [301]	2. está [229]	2. pode [141]	2. faça [03]
3. és [08]	3. têm [62]	3. estado [41]	3. pôde [02]	3. façam [03]
4. era [139]	4. temos [198]	4. estamos [102]	4. podem [35]	4. faço [01]
5. eram [61]	5. tende [07]	5. estando [04]	5. podemos [114]	5. fará [09]
6. éramos [11]	6. tendo [37]	6. estão [111]	6. podendo [04]	6. farão [02]
7. éramos [03]	7. tenha [23]	7. estará [11]	7. poderá [18]	7. faremos [06]
8. foi [292]	8. tenham [09]	8. estarei [01]	8. poderão [05]	8. faria [03]
9. fomos [46]	9. tenhamos [08]	9. estarem [02]	9. poderei [03]	9. fariam [01]
10. for [16]	10. tenho [11]	10. estaremos [10]	10. poderem [02]	10. faz [65]
11. fora [43]	11. terá [09]	11. estaria [05]	11. poderemos [18]	11. fazê [04]
12. foram [134]	12. terão [06]	12. estaríamos [01]	12. poderia [30]	12. fazem [45]
13. forem [06]	13. terem [11]	13. estarmos [05]	13. poderiam [10]	13. fazemos [04]
14. fosse [27]	14. teremos [16]	14. estava [55]	14. poderíamos [06]	14. fazendo [40]
15. fossem [08]	15. teria [17]	15. estavam [18]	15. podermos [06]	15. fazermos [02]
16. fôssemos [01]	16. teriam [9]	16. estávamos [04]	16. podia [04]	16. fazia [04]
17. fui [22]	17. tido [04]	17. estáveis [01]	17. podiam [03]	17. faziam [04]
18. são [359]	18. tinha [32]	18. esteja [08]	18. podíamos [01]	18. fazíamos [01]
19. sê [01]	19. tinham [14]	19. estejam [03]	19. possa [20]	19. feita [14]
20. sede [04]	20. tínhamos [08]	20. estejamos [01]	20. possam [12]	20. feitas [9]
21. seja [104]	21. tive [03]	21. esteve [08]	21. possamos [14]	21. feito [34]
22. sejam [23]	22. tivemos [09]	22. estive [02]	22. posso [13]	22. feitos [8]
23. sejamos [06]	23. tiver [01]	23. estiver [01]	23. pude [06]	23. fez [30]
24. sendo [123]	24. tivera [02]	24. estiveram [05]	24. pudemos [03]	24. fiz [03]
25. será [143]	25. tiveram [18]	25. estivermos [01]	25. puderam [01]	25. fizemos [02]
26. serão [09]	26. tiverem [01]	26. estivesse [05]	26. pudermos [02]	26. fizeram [15]
27. serem [28]	27. tivermos [03]	27. estivessem [01]	27. pudesse [06]	27. fizermos [02]
28. seremos [08]	28. tivesse [09]	28. estou [25]	28. pudessem [02]	28. fizesse [04]
29. seres [17]			29. pudéssemos [04]	29. fizessem [02]
30. seria [55]				30. fizéssemos [01]
31. seriam [09]				
32. seríamos [01]				
33. sermos [27]				
34. sido [44]				
35. sou [07]				
36. somos [252]				

Com relação aos sete verbos – *querer, ficar, achar, falar, precisar, começar* e *olhar* –, presentes entre os vinte primeiros do DIF, mas ausentes na lista dos vinte primeiros do VEU, cumpre-nos, inicialmente, registrar a seguinte ressalva: não obstante esses verbos não integrarem o conjunto dos vinte primeiros do VEU, todos eles possuem uma freqüência alta, qual seja: *querer* (total: 126), *ficar* (total: 127), *achar* (total: 47), *falar* (total: 71), *precisar* (total: 123), *começar* (total: 125) e *olhar* (total: 58). Esses números

demonstram que, com exceção dos verbos *achar, falar* e *olhar*, todos os demais – *querer, ficar, precisar* e *começar* – foram reiteradamente usados nos textos dos acadêmicos, registrando-se uma freqüência quase idêntica entre eles e também muito próxima da freqüência do vigésimo classificado – verbo *dar* –, com 130 ocorrências. Esse fato é sintomático de que, no *corpus*, esses verbos comportaram-se semelhantemente aos vinte primeiros.

Já os verbos *querer, ficar, precisar* e *começar* não integram o grupo dos vinte mais freqüentes porque, dentre os sete verbos que assumiram os seus lugares na lista – *vir, viver, mostrar, possuir, existir, comemorar* e *descobrir*, quatro deles – *vir, mostrar, comemorar* e *descobrir* – estão presos à temática da proposta da redação e, dessa forma, justificam a sua alta freqüência.

Uma última e, talvez, decisiva razão da não coincidência de sete verbos decorre do fato de que, também na lista dos vinte verbos do DIF, há alguns cujo uso é específico de uma determinada época ou gênero textual, não sendo, portanto, muito utilizados na produção de textos dissertativos. Esta incompatibilidade relativa ao gênero pode ser percebida no Quadro 2, que demonstra em qual tipo de literatura houve maior ocorrência de uso dos vinte primeiros verbos do DIF.

Quadro 2 –Tipo de literatura predominante em cada um dos vinte verbos do DIF[7]

1º - **ser**: LT (tecnocientífica)	11º - **saber**: LR (romanesca)
2º - **ter**: LJ (jornalística)	12º - **querer**: LR (romanesca)
3º - **ir**: LJ (jornalística)	13º - **ficar**: LR (romanesca)
4º - **estar**: não há dados	14º - **achar**: LD (dramática)
5º - **poder**: LT (tecnocientífica)	15º - **dever**: LT (tecnocientífica)
6º - **dizer**: LR (romanesca)	16º - **falar**: LR (romanesca)
7º - **haver**: LR (romanesca)	17º - **chegar**: LR (romanesca)
8º - **fazer**: LD (dramática)	18º - **precisar**: LR (romanesca)
9º - **dar**: LR (romanesca)	19º - **começar**: LD (dramática)
10º - **ver**: LR (romanesca)	20º - **olhar**: LD (dramática)

Esse quadro revela que, com exceção do verbo *estar*, do qual não possuímos dados, há um predomínio de formas verbais retiradas da literatura romanesca (dez verbos), quais sejam: *dizer, haver, dar, ver, saber, querer, ficar, falar, chegar, precisar*; existindo, depois, quatro verbos – *fazer, achar, começar, olhar* – pertencentes à literatura dramática; três verbos – *ser, poder* e *dever* – à literatura tecnocientífica; e dois verbos – *ter* e *ir* – à literatura jornalística.

O mesmo quadro também demonstra que, de um lado, os sete verbos ausentes na lista dos vinte primeiros do VEU – *querer, ficar, achar, falar, precisar, começar* e *olhar*

7 Os vinte verbos estão colocados na tabela de acordo com a ordem de classificação da sua freqüência, conforme se pode conferir em Biderman (1998, p.172). Quanto à predominância dos tipos literários, esclarecemos que esses dados ainda não foram publicados, mas a eles tivemos acesso em razão da gentil cessão das informações pela mesma pesquisadora.

– pertencem a um gênero determinado, isto é, quatro verbos – *querer, ficar, falar, precisar* – registram no DIF uma freqüência mais elevada na literatura romanesca e três verbos – *achar, começar, olhar* – apresentam maior incidência na literatura dramática. Por outro lado, constatamos que "todos" os verbos cuja ocorrência é maior na literatura tecnocientífica e na literatura jornalística – *ser, ter, ir, poder* e *dever* – também são os mais freqüentes no vocabulário dos estudantes. Isso faz supor que esses cinco últimos verbos provenientes das literaturas tecnocientífica e jornalística sejam indispensáveis a qualquer tipo de texto, justificando-se, pois, a sua presença significativa no vocabulário dos ingressantes universitários.

Algumas variáveis sociolingüísticas

Não obstante a análise do aspecto quantitativo do vocabulário apresentar marcas de regularidade das estruturas lexicais nele existentes, é necessário destacar as diferenças "qualitativas", ligadas às condições socioeconômico-culturais dos falantes, que apenas podem ser analisadas à luz das orientações da Sociolingüística.

Cumpre informar que foram vários os critérios utilizados, tanto isoladamente quanto em conjunto, para avaliar a "qualidade" do vocabulário. Uma vez que, aqui, não cabe descrevê-los com detalhes, destacamos apenas dois: o da freqüência maior de determinadas palavras plenas, denominadas *especiais*, cujo conteúdo referencial fosse bastante específico, e o da maior adequação do vocabulário aos padrões da norma culta.

Dos vários cruzamentos dos dados sociolingüísticos, constata-se a independência de algumas variáveis e a interdependência de outras. Independentemente de quaisquer outros fatores externos, a *escolaridade*[8] suplanta as demais variáveis, não havendo, entretanto, outras variações significativas, nem mesmo no que tange às diversas camadas socioeconômicas. No estudo da variável que separa os dois grupos de estudantes (universidade particular ou pública), constata-se que o vocabulário daqueles que se preparam para concorrer a uma vaga na universidade pública é mais específico e diversificado, conforme mostra o Quadro 3.

Um dado muito significativo, sobretudo em termos qualitativos, refere-se aos itens lexicais que representam legitimamente a preferência de uso dos dois grupos de estudantes. Referimo-nos às 137 palavras colocadas na primeira coluna do quadro e cuja freqüência atingiu ou superou o limiar de freqüência 40, tanto no conjunto das universidades quanto isoladamente em cada um dos *corpora*. Como se observa pelas formas destacadas, essa mesma listagem contém 32 palavras plenas, das quais 12 são consideradas *especiais*. São elas: *cultura, falta, governo, grande, história, melhor, mundo, problemas, social, sociedade, tempo* e *vida*. A maioria dessas palavras é de conteúdo abstrato e expressa tendências sócio-históricas, ou pensamentos da nossa época.

8 É importante assinalar que esse termo foi utilizado genericamente, aqui, tanto para se referir à Universidade de ingresso quanto ao tipo de escola (pública ou privada) na qual o estudante cursou o Ensino Médio.

Quadro 3 –Palavras de F>=40, segundo a Universidade *Particular* ou *Pública* (parâmetro considerado: as 359 mais freqüentes no *corpus* total)[9]

Palavras de F>=40 no *corpus* total e nos dois *corpora* (137)	Palavras de F>=40 no *corpus* total e no *corpus* da **PARTICULAR** (39)	Palavras de F>=40 no *corpus* total e no *corpus* da **PÚBLICA** (50)	Palavras de F>=40 no *corpus* total, mas com freqüência inferior a 40 nos dois *corpora* (133)
a, à, ainda, **anos**, ao, aos, apenas, apesar, aqui, as, assim, até, bem, **Brasil**, **brasileira**, **brasileiro**, **brasileiros**, cada, com, como, **cultura**, da, das, de, **descobrimento**, desde, **dia**, do, dos, e, é, eles, em, então, entre, era, essa, esse, esta, está, estão, este, **falta**, fazer, foi, foram, **governo**, **grande**, há, **história**, hoje, **índios**, isso, já, mais, mas, **melhor**, mesmo, muitas, muito, muitos, **mundo**, na, **nação**, não, nas, nem, no, nós, nos, nossa, nossas, nosso, nossos, o, onde, os, ou, outros, **país**, **países**, para, **parte**, pela, pelo, pelos, **pessoas**, pode, pois, **população**, por, porque, **portugueses**, pouco, **povo**, problemas, qual, quando, Quase, que, quem, **quinhentos**, **riquezas**, são, se, seja, sem, sempre, sendo, ser, será, seu, seus, sim, só, sobre, **social**, **sociedade**, somos, sua, suas, também, tanto, tão, tem, temos, **tempo**, ter, **Terra**, **terras**, todo, todos, tudo, um, uma, **vez**, **vida**.	agora, antes, brancos, cara, cientistas, comemorar, costumes, crânio, descoberta, descobertas, diferentes, dizer, ele, estamos, fome, futuro, gente, habitantes, Luzia, mistura, mostrar, mudar, nada, negros, podemos, políticos, povos, raça, raças, realmente, rosto, talvez, tecnologia, vai, vamos, várias, vários, violência, vivemos.	além, alguns, amigo, às, Cabral, carta, colônia, colonização, cultural, desenvolvimento, disso, economia, econômica, educação, enquanto, época, estava, eu, exemplo, exploração, fato, fomos, grandes, identidade, independência, índio, início, interesses, lhe, maior, maioria, manchetes, me, meu, milhões, minha, ouro, outras, passado, pau, poder, política, porém, Portugal, quanto, relação, séculos, situação, todas, você.	algumas, ano, atrás, através, atualmente, bom, busca, características, certeza, certo, chegada, coisa, coisas, continua, colonizadores, contra, comemoração, completar, condições, continente, corrupção, culturas, dar, depois, descoberto, desemprego, dessa, desse, desta, deste, deve, devemos, devido, dias, diferente, dinheiro, distribuição, durante, econômico, enfim, eram, escolas, esperança, essas, esses, estado, estar, estes, estrangeiros, europeus, existe, existem, existência, fatos, faz, fazem, fazendo, fora, forma, governantes, homem, homens, imagem, importante, isto, lado, lo, lugar, luta, maneira, mãos, meio, melhorar, menos, milênio, miscigenação, miséria, momento, mostra, muita, nacional, nações, nativos, naturais, negro, neste, notícias, nova, novas, novo, num, nunca, origem, outra, outro, parece, Paulo, Pedro, pelas, pobre, pobres, portanto, possível, possui, poucos, presente, primeiro, primeiros, própria, principalmente, qualquer, realidade, recursos, renda, respeito, sabe, sabemos, saber, saúde, século, seria, sido, sociais, tal, tantas, têm, toda, vem, trabalho, ver, verdade, verdadeira, viver.

9 Todas as palavras destacadas são plenas, ou de significação externa.

debate é inerente à ciência. Dessa forma, os resultados e as análises apresentados talvez se revistam de um caráter de incompletude muito maior do que propriamente conclusivo, podendo haver, ainda, muito mais perguntas que respostas envolvendo o objeto de estudo da pesquisa realizada.

À guisa de conclusão, cumpre-nos, assim, apenas retomar alguns dados mais expressivos e refletir sobre os aspectos mais relevantes do ponto de vista dos objetos pretendidos, lembrando que este trabalho apresenta resultados apenas parciais de uma pesquisa maior.

Preliminarmente, convém destacar que, ao se estudar a sistematicidade do vocabulário, deve-se atentar para a diversidade e a variabilidade também própria da sua natureza. Essa diferenciação não se refere apenas ao plano lingüístico, ou seja, às estruturas lexicais internas, mas expande-se para o mundo exterior. Sobretudo considerando os elementos da estruturação da realidade, é possível observar não existirem formas de se desenvolver uma análise – muito menos descrição – precisa e exaustiva dos fatos.

Há que se assinalar também a confirmação, na prática, de algumas regras de organização perante a diversidade e a variabilidade da língua, que se manifestam no vocabulário. Uma delas refere-se à existência de um comportamento *regular* quanto à distribuição da freqüência e ao número de unidades lexicais. Esse fenômeno da força da estrutura sobre a variação foi constatado não só com base no *corpus* total, mas também nos *corpora* resultantes dos recortes das variáveis estudadas. A *regularidade*, aqui observada, diz respeito à alta freqüência dos instrumentos gramaticais e de um determinado grupo de verbos, e à baixa freqüência das palavras plenas, ou de conteúdo externo – em especial, substantivos e adjetivos.

Outro fator a ser ressaltado refere-se à impossibilidade de se fazer qualquer tipo de avaliação do vocabulário sem operar recortes no *corpus*. Somente com base nos estudos das variáveis existentes nos grupos sociolingüísticos, foi possível perceber as semelhanças e/ou dessemelhanças e obter um resultado satisfatório, caracterizando a amplitude maior ou menor do vocabulário analisado, bem como a sua adequação ou inadequação de uso.

Os dados evidenciam, conclusivamente, maior competência vocabular por parte dos estudantes que ingressam na universidade pública, talvez em razão de estarem mais bem preparados para concorrer às vagas limitadas que ela oferece. Mostram, também, que aqueles que, até o Ensino Médio, estudaram em escola particular também possuem um conhecimento maior em termos vocabulares.

Se "o léxico é o indicador mais seguro de dificuldade do texto" e se as deficiências dos estudantes na compreensão da escrita, ou na percepção da função dos itens lexicais no texto, são decorrentes do ensino (KLEIMAN, 1989, p.132), deduz-se, então, que, sobretudo na escola da rede pública, os alunos não têm sido submetidos à prática de atividades adequadas ao desenvolvimento do seu vocabulário ativo. Diante desses resultados, não se pode deixar de ressaltar, mais uma vez, a necessidade de se refletir sobre o decisivo papel da Escola como "divisor de águas", também no que concerne ao ensino da língua materna.

Comparando as duas listas de palavras, que atingiram o limiar de freqüência 40 em cada *corpus*, observa-se que no rol da Instituição Pública há um grupo maior de palavras *especiais*. São dezoito palavras (*cultura, desenvolvimento, economia, econômica, educação, época, exploração, grandes, identidade, interesses, maior, maioria, milhões, passado, política, relação, séculos* e *situação*), que equivalem a 36% do total, contra onze palavras – 22% – registradas na lista da universidade particular (*brancos, diferentes, fome, futuro, gente, mudar, negros, políticos, violência, vivemos* e *tecnologia*).

Analisando-se aquele conjunto de dezoito palavras *especiais*, observa-se a presença de temas abrangentes, do ponto de vista do contexto vivido pela sociedade brasileira. Em particular, a palavra *política*, e não *políticos*, constante da lista da Instituição Particular, parece ser capaz de comprovar esse fenômeno e, ao mesmo tempo, de sintetizar o pensamento dos estudantes da Escola Pública. Na oposição entre *política/políticos* é possível supor que, enquanto este grupo trata da questão *política* do Brasil, aquele discute a situação dos *políticos* brasileiros.

Há também um dado curioso do ponto de vista gramatical: ocorre maior incidência de marca de plural nas palavras plenas do vocabulário dos estudantes da Instituição Pública; isso denota tendência de concordância em sintagmas nominais. Essa população apresenta, portanto, maior domínio dos elementos lingüísticos exigidos na variante de prestígio, ou modalidade escrita-culta.

Quando essas populações são subdivididas em quatro grupos sociolingüísticos, de acordo com a variável: escola (particular ou pública) em que concluíram o Ensino Médio, os dados são ainda mais expressivos. Tomando-se por parâmetro a extensão dos textos dos estudantes, ou seja, o número médio de palavras por redação, este trabalho apurou os seguintes resultados: a) textos de estudantes de universidade pública advindos da rede particular de ensino: média de *364 palavras*; b) textos de estudantes de universidade pública advindos da rede pública de ensino: média de *329 palavras*; c) textos de estudantes de universidade particular advindos da rede particular de ensino: média de *206 palavras*; d) textos de estudantes de universidade particular advindos da rede pública de ensino: média de *181 palavras*.

Todos esses elementos presentes e/ou ausentes no vocabulário de cada grupo de estudantes confirmam a existência de divergências qualitativas e contribuem para que se considere mais amplo o vocabulário dos ingressantes na universidade pública, sobretudo dos que concluíram o Ensino Médio em escola da rede particular.

Esses resultados demonstram a relação entre língua e sociedade, ou seja, o imbricamento entre fenômenos lingüísticos e aspectos sociopolíticos, e ainda denunciam problemas de desigualdades decorrentes da má qualidade do ensino de Língua Portuguesa, até o Ensino Médio, sobretudo na escola pública.

Verifica-se, portanto, que esse tipo de pesquisa no âmbito do vocabulário é de particular interesse para o ensino, pois somente com o conhecimento sobre quais vocábulos merecem maior atenção, ou sobre quais fatores desencadeiam semelhanças e/ou dessemelhanças com relação ao uso, é que se pode "evitar o empirismo na escolha do vocabulário para fins didáticos" (BIDERMAN, 1998, p.179).

Considerações finais

A árdua tarefa de se investigar os mistérios que envolvem a linguagem humana exige uma explicação científica, ao mesmo tempo sempre provisória, uma vez que o

Agradecimentos

Às Instituições de Ensino Superior – Universidade Estadual de Campinas (Unicamp-SP) e Faculdades Integradas Toledo de Araçatuba (SP) –, que forneceram o *corpus* para a pesquisa, e à Profª Drª. Maria Tereza Camargo Biderman (UNESP), por sua inestimável colaboração na cessão dos materiais de sua pesquisa.

NASCIMENTO, R. I. do; ISQUERDO, A. N. Frequency of words: a diagnostic of the vestibular compositions vocabulary. *Alfa*, São Paulo, v.47, n.1, p.71-84, 2003.

- *ABSTRACT: This paper presents the results of a lexical-statistical research on a* corpus *formed by 450 "vestibular" compositions (1999 and 2000) from a private and a public university in São Paulo state. The statistical results from this sample of the university applicants vocabulary resemble the data from another researches in the same area, specially when it comes to the high frequency level of the small amount of lexical units: 11,151 different types totaling 113,638 tokens in the* corpus. *The confrontation of this research data to the data from the* Frequency Dictionary of Contemporary Brazilian Portuguese *lexicon, by Biderman (1998), shows a small amount of words (a slight more than three hundred) which were common in the two databases and which shared similar frequency levels. This indicates the existence of a possible nucleus of the analyzed vocabulary which might gather words probably used in any kind of text. This phenomenon is, above all, marked by the twenty most frequent verbs in the corpus. Data resulting from different samples and from sociolinguistic crossings of a few variables such as the university the students applied for; public or private high school graduation, gender and family income, confirm this phenomenon. From this analysis, two findings deserve to be highlighted: the superiority of the vocabulary from those applying for a private university and the "school background" as a variable that overcomes all the others.*

- *KEYWORDS: Lexical-statistical; vocabulary; "vestibular" compositions; university applicants; verbs.*

Referências bibliográficas

BIDERMAN, M. T. C. *Teoria lingüística*: lingüística quantitativa e computacional. Rio de Janeiro: Livros Técnicos e Científicos, 1978. 277p.

_____. A face quantitativa da linguagem: um dicionário de freqüências do português. *Alfa*: Revista de Lingüística, São Paulo, v.42, n. esp., p.161-81, 1998.

CRUZ, M. L. S. da. A norma lexicográfica no tratamento do *Corpus* de freqüência. In: NASCIMENTO, M. F. B. do; RIVENC, P.; CRUZ, M. L. S. da. *Português fundamental*: métodos e documentos. Lisboa: Instituto Nacional de Investigação Científica, Centro de Lingüística da Universidade de Lisboa, 1987, p.311-421. v.2, tomo primeiro.

KLEIMAN, A. *Leitura*: ensino e pesquisa. Campinas: Pontes, 1989. 213p.

MATORÉ, G. *La méthode en lexicologie*: domaine français. Paris: Marcel Didier, 1953. 126p.

NASCIMENTO, R. I. do. *O vocabulário dos estudantes universitários*: um estudo com base em redações de vestibular. 2001. Dissertação (Mestrado em Letras) – Universidade Federal de Mato Grosso do Sul, Três Lagoas, 2001.

PAIVA, M. da C. de; SCHERRE, M. M. P. Retrospectiva sociolingüística: contribuições do PEUL. *Delta*, São Paulo, v.15, n. esp., p.201-32, 1999.

Bibliografia consultada

BIDERMAN, M. T. C. As ciências do léxico. In: OLIVEIRA, A. M. P. P.; ISQUERDO, A. N. (Orgs.) *As ciências do léxico*: lexicologia, lexicografia e terminologia. Campo Grande: Editora da UFMS, 1998. p. 11-20.

NASCIMENTO, M. F. B.; CRUZ, M. L. S. da. O inquérito de disponibilidade. In: NASCIMENTO, M. F. B.; RIVENC, P.; CRUZ, M. L. S. da. *Português fundamental*: métodos e documentos. Lisboa: Instituto Nacional de Investigação Científica, Centro de Lingüística da Universidade de Lisboa, 1987. p.27-40. v.2, tomo 2.

OLIVEIRA, A. M. P. P. de; ISQUERDO, A. N. (Org.) *As ciências do léxico*: lexicologia, lexicografia e terminologia. Campo Grande: Editora da UFMS, 1998. 262p.

REY, A. *La Lexicologie*. Paris: Klincksieck, 1970. 323p.

REY-DEBOVE, J. Léxico e dicionário. *Alfa*, Revista de Lingüística, São Paulo, v.28 supl., p.45-69, 1984.

TARALLO, F. *A pesquisa sociolingüística*. São Paulo: Ática, 1986. 96p.