Cadernos de Estudos Lingüísticos
Número 6, 1984
Páginas 183 - 186

2. UM MÉTODO QUANTITATIVO PARA A ANÁLISE LEXICAL*

Enzo Del Carratore
UNESP (Marília)

O trabalho em epígrafe veio preencher uma lacuna no campo ainda inexplorado entre nós dos estudos quantitativos de textos literários; muito embora essa afirmação constitua um dos muitos lugares-comuns de que se valem usualmente os críticos, é felizmente uma verdade alvissareira. Pela primeira vez estamos diante de um trabalho de elogiável honestidade, seriedade e rigor metodológico, que alia os preceitos teóricos extraídos das melhores fontes à prática da análise lexical, exercida sobre a obra de três poetas simbolistas brasileiros: Alphonsus de Guimaraens, Cruz e Sousa e Edgard Mata. A própria A. reconhece que mais poderia ter sido feito: "poderíamos proceder a diversos tipos de análise quantitativa do estilo, visando especialmente à determinação de desvios significativos, denotadores de traços pertinentes à expressão poética de Edgard Mata, que os identificasse com os traços paradigmáticos do Simbolismo brasileiro, inferidos da obra de Cruz e Sousa e Alphonsus de Guimaraens. Nossa intensão, porém, é mais modesta. Pretendemos (...) configurar (...) os principais universos de significação individual e/ou comum aos três autores" (p. 47). Intenção modesta mas nem por isso, acrescentamos, menos válida e meritória. De passagem, porém, notamos que a modéstia revelada nesse trecho contradiz o propósito contido à p.23: "... o que vamos fazer é, de um lado, determinar as características da população pelo julgamento das amostras de Alphonsus de Guimaraens e Cruz e Sousa e, de outro lado, verificar, pelo julgamento comparativo das amostras, se Edgard Mata possui essas mesmas características". Apesar desse cochilo,que esperamos ver corrigido numa próxima reedição do trabalho, a A. consegue mostrar, através de uma análise bem conduzida e convincente do universo lexical dos três autores, os traços estilística e tematicamente mais significativos de cada um, chegando à conclusão da "predominância do etéreo e da sensualidade em Cruz e Sousa, do funéreo e do sensorial em Alphonsus de Guimaraens, e do funéreo e do místico em Edgard Mata" (p. 72).

No mais, não tentaremos resumir o trabalho da profa. Cilene C. de Souza: achamo-lo por demais interessante para enquadrá-lo nos estreitos limites de uma recensão; preferimos recomendar sua leitura, sem dúvida instrutiva e proveitosa, ainda que pareça por vezes um pouco árida para quem não está suficientemente familiarizado com uma linguagem necessariamente técnica. Aliás, apontamos como um dos principais méritos do livro a sua simplicidade: a A. conseguiu, de maneira sóbria e muito didática, apresentar conceitos estatísticos numa linguagem clara e acessível, de modo a tornar a lei

tura de seu trabalho bastante fácil. Excelente a visão, forçosamente sucinta e simplificada, das principais noções básicas de estatística e dos métodos de análise que serão utilizados, agora acessíveis aos leigos, e que constituem o Capítulo I intitulado "Fundamentos teóricos do método", onde se reconhece a orientação segura de Charles Muller. Por falar nesse autor, observamos que não estão relacionadas na Bibliografia as suas obras mais recentes: Initiation aux méthodes de la statistique linguistique (1973) e especialmente Principes et méthodes de statistique lexicale (1977).

O trabalho traz evidentemente falhas e omissões. As principais são resultantes das limitações que os linguistas brasileiros enfrentam quando pretendem utilizar um instrumental de trabalho insólito - os métodos quantitativos; a própria A. se lamenta do fato de ter sido forçada a um levantamento manual das ocorrências vocabulares nos poetas estudados, tendo que reduzir a sua amostragem a uma quantidade razoável, mas não ideal, de poemas, por não poder dispor do instrumento mais poderoso que se conhece para trabalhos dessa natureza - o computador.

Entre as omissões que poderíamos apontar, a primeira decorre da limitação aqui exposta: a falta de um levantamento sistemático do "vocabulário característico" de cada autor, pela falta de um referencial indispensável: um léxico de frequências do idioma ou, no mínimo, de um corpus mais extenso que o das amostras (por exemplo, o léxico da poesia simbolista brasileira), que forneceria o modelo teórico em relação ao qual seria possível fixar o vocabulário característico de cada poeta, utilizando o teste do desvio reduzido.

Apontamos ainda duas opções de estudo que não foram aproveitadas - e o poderiam ter sido, contando apenas com os dados obtidos pela A. A primeira consiste no inaproveitamento dos vocábulos de frequência 1 na análise feita pela A.; com isso, embora explique o motivo, a profa. Cilene renuncia a um estudo muito interessante sobre a estrutura do vocabulário dos três autores, que poderia ser confrontada com os modelos teóricos obtidos pela lei binomial e pela distribuição de Waring, a fim de verificar o ajuste, satisfatório ou não, dos modelos aos dados reais. A segunda, que nos traria um elemento de apreciação muito importante, é o cálculo da riqueza do vocabulário de cada autor: vários processos poderiam ser utilizados e nos dariam informações muito úteis, e a A. deve dispor de todos os dados necessários para esse cálculo, mas que não são fornecidos no trabalho; fica aqui a indagação: qual dos três poetas simbolistas apresentam o vocabulário mais rico?

Outros reparos se fazem necessários: trata-se, com toda a evidência, de erros de impressão ou revisão, que não enganam o especialista, mas que podem induzir a dúvidas insolúveis ou a erros crassos um principiante curioso. Além da falta de um trecho entre "point de vue" e "thématique" na citação da p.15, assinalaremos os principais:

à p.27, o multiplicador do desvio padrão tal como está - $-\sqrt{n(n-1)}$ - levaria a absurdos; o correto será $\sqrt{n/(n-1)}$;

à p.29, na fórmula do desvio padrão está faltando o sinal de radical: $\sqrt{\frac{p.q}{n}}$;

à p.31, a fórmula correta do desvio reduzido é $z = \frac{x - np}{\sqrt{npq}}$, desde que o numerador representa o desvio absoluto $x-\bar{x}$, sendo aqui a freqüência média teórica identificada com np.

à p.38, o quadro "Verbos" apresenta valores incorretos:

σ	z	p
29	+2,24	0,028
28	-2,57	0,009
27	+0,26	0,764

devem ser:

σ	z	p
19	+3,42	0,00068
19	-3,79	0,00014
18	+0,39	0,689

;

à p.45, o valor do χ^2 para a última classe de efetivos é de 0,22 e não de 0,022, o que eleva o valor do χ^2 para 4,4336 - o que em nada altera a apreciação do desvio;

à p.56, há um erro na coluna das porcentagens do quadro e, consequentemente, nos valores calculados da linha correspondente: EM 10,08% (e não 16,08); efetivos teóricos 5 (e não 9); desvio zero (e não 4); χ^2 = zero (e não 1,77), o que reduz o valor do χ^2 para 26,80; aqui também a apreciação do desvio não se altera;

à p.57, os valores do χ^2 estão errados para os três autores: AG = 0,167 (e não 0,085); CS = 3,2 (e não 1,73); EM = 3,2 (e não 1,73); o χ^2 total é portanto 6,567, e não 3,545; aqui a conclusão deverá ser modificada: a probabilidade passa a ser igual a 3,75% em lugar de 17%, e o desvio deverá ser considerado significativo, rejeitando-se a hipótese de uma repartição regular do vocábulo entre os autores, na amostra considerada.

Uma observação final quanto aos Quadros que integram os Anexos ao trabalho. Teria sido conveniente indicar, ainda que fosse através de um único exemplo, o processo pelo qual o valor das estatísticas apresentadas foi obtido; assim, se nenhuma dúvida paira acerca da obtenção dos valores do χ^2 para os vocábulos do Quadro 1, o mesmo não podemos dizer dos valores do desvio reduzido z nos Quadros 2 a 5; vamos mostrar apenas um exemplo, o da palavra "triste" do Quadro 2.

Nós obteríamos o valor de z da seguinte maneira: desde que a probabilidade de ocorrência de uma palavra é igual à sua freqüência dividida pelo número de ocorrências do corpus, p = f/N, no caso, sendo f = 17 e N = 9311, p = 17/9311, isto é, p = 0,00183, e q será igual a 0,99817; se, na amostra considerada, f = 10, temos o seguinte quadro:

	f_t (=np)	f_o	d	σ (= npq	z	z (seg.a A.)
AG	1218 x 0,00183 = 2	4	2	1,49	1,34	1,42
CS	965 x 0,00183 = 2	1	-1	1,33	0,75	0,71
EM	1109 x 0,00183 = 2	5	3	1,42	2,11	2,13
	3292	10				

As diferenças aparecem em quase todos os itens lexicais analisados, mas são diferenças, como se nota, pouco significativas, a não ser em dois casos, sobre os quais chamamos a atenção (pode haver outros, pois não conferimos todos...): no Quadro 3, para a palavra "sonho" em EM, calculamos um z = 1,03, e não 0,10 como ali consta; no Quadro 5, para a palavra "saudade" em AG, calculamos um z = 1,19, em lugar do assinalado 2,89 (e isto faz diferença).

Estas observações, longe de desmerecer o trabalho, foram feitas com o exclusivo intuito de contribuir, ainda que modestamente, para melhorá-lo por ocasião de uma reedição, que esperamos se dê em breve.

Finalizando, gostaríamos de congratular-nos com a A. pelo seu livro, que, acreditamos, em muito contribuirá para divulgar uma metodologia de trabalho pouco praticada entre nós, mas cujos resultados, animadores e confiáveis, nos autorizam a alimentar fundadas esperanças em que cada vez mais os pesquisadores a ela recorram e dela se beneficiem.

NOTA:

* Resenha da obra homônima de Cilene Cunha de Souza, Rio de Janeiro, Editora Tempo Brasileiro/INL, 1979.