Alfa, São Paulo
25:31-47, 1981

O CORPUS TIBULLIANUM: UMA ANÁLISE QUANTITATIVA

Enzo Del CARRATORE*
Cicília Yuko WADA**

RESUMO: A partir dos dados extraídos do Corpus Tibullianum, e submetidos a tratamento estatístico, os autores pretendem reforçar a opinião geralmente aceita de que a obra não foi composta por um único poeta, admitindo a pluralidade de autores. Para tanto, utilizam, a distribuição de Waring e os índices mais recentes de riqueza lexical, apontando, entre as duas partes em que dividiram o Corpus, *diferenças que corroboram a hipótese inicial.*

UNITERMOS: Estatística lexical; distribuição de Waring; estimadores; índices de riqueza lexical; extensão teórica do vocabulário.

INTRODUÇÃO

O presente trabalho irá apresentar uma abordagem quantitativa dos dados fornecidos pelo *Corpus Tibullianum,* com a intenção precípua de salientar a utilidade e a conveniência do emprego de técnicas probabilísticas de análise como recurso eficaz de convalidação de teorias lingüísticas ou literárias. Não propomos a substituição dos métodos qualitativos tradicionais de interpretação de fatos lingüísticos ou de apreciação de obras literárias; desejamos somente mostrar a críticos literários e a lingüistas alguns instrumentos de trabalho que lhes poderão ser úteis e valiosos.

O aparato matemáticos sobre o qual se alicerça o trabalho foi reduzido ao mínimo indispensável para a compreensão de qualquer leitor, por leigo que seja; fizemos isso para não espantar e afugentar a nossa clientela potencial de lingüistas e de literatos, tradicionalmente avessa neste país à ingerência de métodos de uma ciência "exata" num domínio presumivelmente insubmisso a leis outras que não sejam as da livre e espontânea criação. No entanto, a ciência lingüística muito lucraria com a aliança de que nos propomos apresentar uma breve e singela amostra. Mais e melhor fizéramos, se tivéssemos à disposição dados que, corriqueiros em outras terras, inexistem na nossa; esta limitação servirá ao mesmo tempo de apelo aos estudiosos patrícios, e de pedido de desculpas antecipado pelas eventuais falhas ou omissoẽs de nosso artigo.

1 DISTRIBUIÇÃO DE WARING

IRWIN[5] ao investigar o ajuste de modelos teóricos para descrever, na área das ciências biológicas, distribuições de freqüências com caudas longas, propôs a utilização de uma distribuição baseada na expansão de uma série de termos descrescentes, conhecida por série de Waring:

$$\frac{1}{x-a} = \frac{1}{x} + \frac{a}{x(x+1)} + \frac{a(a+1)}{x(x+1)\ (x+2)} + \ldots + \frac{a(a+1)\ldots(a+i-2)}{x(x+1)\ldots(x+i-1)} \quad (I)$$

* Professor Titular do Departamento de Lingüística e Língua Portuguesa do Instituto de Letras, História e Psicologia — Campus de Assis — UNESP, SP.

** Professora Assistente do Departamento de Análise Numérica e Estatística do Instituto de Biociências, Letras e Ciências Exatas — Campus de São José do Rio Preto — UNESP, SP.

Essa série, por ser convergente para $x > a > 0$ pôde ser transformada em distribuição discreta de probabilidades,

$$\frac{x-a}{x} + \frac{a(x-a)}{x(x+1)} + \frac{(x-a)a(a+1)}{x(x+1)\ (x+2)} + \ldots + \frac{(x-a)a(a+1)\ldots(a+i-2)}{x(x+1)\ldots(x+i-1)} = 1 \quad \text{(II)}$$

Irwin denominou essa distribuição de probabilidades de "Distribuição de Waring" *.

1.1. HERDAN[4] em 1964, investigou por suas vez a aplicação da distribuição de Waring a dados obtidos de textos literários, quando se realiza a contagem dos diferentes vocábulos com freqüência i = 1,2,3,... Em um texto de tamanho N, constituído de V vocábulos, observa-se que a distribuição dos V_i vocábulos, isto é, dos vocábulos que aparecem no texto dado com freqüência 1,2,3..., é decrescente; em virtude dessa propriedade da distribuição, Herdan comprovou o ajuste da distribuição de Waring a dados lexicais.

Assim:

$p_1 = \frac{x-a}{x}$ é a probabilidade de um vocábulo ter freqüência 1;

$p_2 = \frac{(x-a)a}{x(x+1)}$ é a probabilidade de um vocábulo ter freqüência 2;

$p_i = \frac{(x-a)a(a+1)\ldots(a+i-2)}{x(x+1)\ldots(x+i-1)}$ é a probabilidade de um vocábulo ter freqüência i.

Para se obter o valor de cada p_1 (isto é, a probabilidade de um vocábulo ter no texto a freqüência i), é necessário conhecer o valor dos parâmetros a e x - o que nos leva ao problema dos estimadores desses parâmetros, que será tratado no item seguinte.

Uma vez de posse dos valores de â e $\hat{x}$, pode-se construir o modelo teórico de distribuição das freqüências dos vocábulos num texto qualquer, aplicando a distribuição de Waring, da seguinte maneira:

$$E(V_1) = \frac{\hat{x}-\hat{a}}{\hat{x}} V$$

$$E(V_2) = \frac{(\hat{x}-\hat{a})\hat{a}}{\hat{x}(\hat{x}+1)} V \quad \text{ou:} \quad E(v_2) = E(V_1)\frac{\hat{a}}{\hat{x}+1}$$

$$E(V_3) = \frac{(\hat{x}-a)\hat{a}(\hat{a}+1)}{\hat{x}(\hat{x}+1)\ (\hat{x}+2)} V \quad \text{ou:} \quad E(V_3) = E(V_2)\frac{(\hat{a}+1)}{(\hat{x}+2)}$$

$$E(V_i) = \frac{(\hat{x}-\hat{a})\hat{a}(\hat{a}+1)\ldots(\hat{a}+i-2)}{\hat{x}(\hat{x}+1)\ldots(\hat{x}+i-1)} V \quad \text{ou:} \quad E(V_i) = E(V_{i-1})\frac{\hat{a}+i-2}{\hat{x}+i-1}$$

Desse modo, obtemos o número de vocábulos que se espera encontrar repartidos pelas classes de freqüência i, admitindo que eles se repartam de fato segundo a distribuição de Waring.

* Um estudo sucinto da distribuição (II), a partir da expansão da função $\frac{1}{x-a}$ será apresentado no Anexo que acompanha este artigo.

1.2 Estamos agora diante de duas distribuições de freqüência: uma observada, e a outra esperada — a construída sobre o modelo ora descrito. Formulamos então a "hipótese nula" H_0, isto é, a hipótese da "não diferença" entre as duas distribuições. O problema que se apresenta é o de verificar se os desvios existentes entre os valores efetivos observados V_i e os esperados $E(V_i)$ são significativos ou não: se os desvios se revelarem significativos, isto é, superiores a um valor crítico a que está associada uma probabilidade de ocorrência aleatória, rejeitamos a hipótese nula e concluímos que o modelo teórico de distribuição de freqüências não se ajusta aos dados observados; ao contrário, se os desvios não forem significativos, não podemos rejeitar a hipótese nula, o que nos permite aceitar, dentro de uma margem previsível de erro, o fato de que o modelo teórico se ajusta aos dados reais.

Para testar a hipótese nula, isto é, para tomar a decisão de rejeitá-la ou não, utilizamos o teste de aderência (x^2 de Pearson)*.

Rejeita-se H_0 quando o valor do x^2 obtido for superior ou igual ao valor do x^2 crítico, com um número de graus de liberdade dado por $n = i-j-1$, onde i é o número de classes de freqüência, e j o número de parâmetros estimados.

2. OS ESTIMADORES

Um modelo probabilístico é caracterizado por seus parâmetros. Como esses parâmetros normalmente são desconhecidos, torna-se necessário "estimá-los", isto é, encontrar, a partir de dados amostrais, valores que mais se aproximem dos verdadeiros valores dos parâmetros populacionais. Esses valores são resultados daquilo que na Estatística costuma chamar-se de "estimadores", que são então funções de variáveis aleatórias.

Para determinar com o máximo de precisão o valor de um parâmetro, podemos construir um número praticamente ilimitado de estimadores *ad hoc*, além daqueles baseados em técnicas convencionais como, por exemplo, os estimadores obtidos pelo método da máxima verossimilhança. Estes estimadores, embora possuam ótimas propriedads e por isso forneçam os melhores resultados, são de difícil obtenção na prática, porque as equações de máxima verossimilhança somente podem ser solucionadas por processos iterativos, exigindo o emprego de algoritmos e o uso de computador.

No caso dos parâmetros α e χ da distribuição de Waring, examinaremos a seguir três tipos de estimadores dentre os comumente usados.

2.1 Irwin propôs estimadores baseados na média observada $\hat{\mu}_1$ e na probabilidade $\hat{p}_1$ associada aos vocábulos de freqüência 1. Daí as fórmulas **

$$\hat{a} = \frac{1}{\frac{1}{\hat{q}_1} - \frac{1}{\hat{\mu}_1} - 1} \qquad \text{(III)}$$

e

* Por esse teste, o valor do x^2 obtido segundo a fórmula $x^2 = \Sigma \frac{(f_i - f_t)^2}{f_t}$ serve para avaliar, em termos de probabilidade, o desvio constatado entre os valores reais e os teóricos. Uma das limitações do teste é a de oferecer resultados pouco confiáveis para valores calculados muito pequenos; convenciona-se fixar o limite de 5 como o valor mínimo abaixo do qual o teste se torna ineficiente, e por essa razão costuma-se acumular as classes de freqüência mais elevada na cauda da distribuição, a partir do momento em que a primeira delas apresentar um valor calculado inferior a 5. É o que se verificará logo adiante nas tabelas que integram este texto.

** A obtenção dessas fórmulas será detalhada no Anexo.

$$\hat{x} = \frac{\frac{1}{\hat{q}_1}}{\frac{1}{\hat{q}_1} - \frac{1}{\hat{\mu}_1} - 1} \qquad \text{(IV)}$$

Lembramos que a média observada $\hat{\mu}_1$ é por esse autor obtida considerando i = 0,1,2,. . .:

$$\hat{\mu}_1 = 0p_1 + 1p_2 + 2p_3 + \ldots + (i-1)p_i \qquad \text{(V)}$$

Observe-se, a esta altura, que Irwin, ao lidar com problemas biológicos, considerou uma variável aleatória que assumia o valor inicial zero, o que não ocorre evidentemente com dados lingüísticos, onde a variável "classe de freqüência" assume usualmente os valores i = 1,2,3,. . .

Deste modo, a média observada f = é obtida por:

$$\bar{f} = 1p_1 + 2p_2 + 3p_3 + \ldots + ip_i \qquad \text{(VI)}$$

A diferença entre $\bar{f}$ e μ_1 é igual a 1, obtida pela subtração, membro a membro, de (VI) menos (V), que resulta em:

$$\bar{f} - \hat{\mu}_1 = p_1 + p_2 + p_3 + \ldots + p_i$$

Como, por definição, a soma das probabilidades é sempre igual a 1, logo

$$\bar{f} - \hat{\mu}_1 = 1 \quad \text{e} \quad \hat{\mu}_1 = \bar{f} - 1$$

Utilizaremos uma notação mais familiar aos que trabalham com dados lingüísticos, operando as seguintes substituições:

$$p_1 = \frac{V_1}{V} \; ; \; q_1 = 1 - \frac{V_1}{V} \; ; \; \bar{f} = \frac{N}{V}$$

Desta forma, os estimadores (III) e (IV) de Irwin podem ser notados por:

$$\hat{a} = \frac{1}{\frac{1}{(1 \frac{V_1}{V})} - \frac{1}{(\frac{N}{V} 1)} - 1} \qquad \text{(VII)}$$

e

$$\hat{x} = \frac{\hat{a}}{1 - \frac{V_1}{V}} \qquad \text{(VIII)}$$

2.2 Ao aplicar os estimadores de Irwin a dados de natureza lingüística, na tentativa de ajuste do modelo de Waring à distribuição de freqüências dos vocábulos de um texto literário, Herdan cometeu o equívoco de não perceber a diferença de 1 unidade entre a média $\bar{f}$, que considera i = 1,2,. . .,e a média $\hat{\mu}_1$, que considera i = 0,1,2,. . . Assim os estimadores de Herdan serão construídos segundo as fórmulas

$$\hat{a} = \frac{1}{\frac{1}{(1-\frac{V1}{V})} - \frac{1}{(\frac{N}{V})} - 1} \qquad (IX)$$

$$\hat{x} = \frac{\hat{a}}{1 - \frac{V_1}{V}} \qquad (X)$$

2.3 Um recente trabalho de RATKOWSKY[10] chamou-nos a atenção para novos estimadores, descobertos por aquele pesquisador, e que, aplicados à distribuição de Waring, fornecem em muitos casos um melhor ajuste de um modelo capaz de descrever satisfatoriamente a distribuição dos vocábulos de um texto segundo a ordem de freqüência de suas ocorrências. Isto porque, ao contrário de Irwin e de Herdan, que levaram em conta apenas os parâmetros f (freqüência média) e p_1 (probabilidade de ocorrência de um vocábulo de freqüência 1), Ratkowsky propõe um método de estimação que utiliza as classes de freqüência 1, 2 e 3 — o que proporciona maior flexibilidade ao processo de cálculo dos valores teóricos, resultando daí, geralmente, uma melhor concordância entre os dados observados e os calculados para as várias classes de freqüência.

Os novos estimadores para os parâmetros α e χ são obtidos pelas fórmulas

$$\hat{x} = d + (d^2 + e)^{1/2} \qquad (XI)$$

$$\hat{a} = \hat{x}(1 - \hat{p}_1) \qquad (XII)$$

que requerem o cálculo prévio dos seguintes termos:

$$b = \frac{\hat{p}_2 + \hat{p}_3}{\hat{p}_1}$$

$$c = 2 + \hat{p}_1^2 - 3\hat{p}_1 - b$$

$$d = \frac{1{,}5(+ \hat{p}_1 - 1)}{c}$$

$$e = \frac{2b}{c}$$

O autor desse novo método de estimação aplicou-o a 33 obras literárias, tendo obtido resultados sensivelmente melhores em 26 casos em relação aos obtidos através dos estimadores de Irwin, e em 21 casos em relação aos obtidos pelos estimadores de Herdan. Além disso, Ratkowsky mostrou que, para 28 das 33 obras, a distribuição de Waring, calculada a partir dos novos estimadores, oferece uma concordância aceitável com os dados observados, em termos de probabilidade, verificada através do teste x^2.

3. ANÁLISE QUANTITATIVA DO *Corpus Tibullianum*.

Estamos agora em condições de aplicar as noções teóricas expostas até aqui à obra que constituirá o principal objeto do nosso estudo.

3.1 À guisa de pré-teste e de exemplificação do método empregado, aplicamos o procedimento descrito por Ratkowsky, comparando os resultados com os valores obtidos com a utilização dos estimadores de Irwin e de Herdan. A obra escolhida para este primeiro ensaio foi o poema latino *Aetna*, de autoria desconhecida *. Como não dispunhamos de índice de palavras, procedemos ao levantamento manual de todas as ocorrências do texto e à contagem das diversas classes de freqüência, o que deu origem à tabela seguinte **:

TABELA 1

Valores observados e calculados dos vocábulos no poema Aetna

Dados: N = 4324, V = 1446, V_1 = 768, V_2 = 283, V_3 = 135

Freqüências i	Valores observados V_i	Valores calculados Estimadores de Irwin $E(V_i)$	Valores calculados Estimadores de Herdan $E(V_i)$	Valores calculados Estimadores de Ratkowsky $E(V_i)$
1	768	768,00	768,00	768,00
2	283	277,95	262,02	282,45
3	135	133,54	126,35	135,55
4	56	75,03	72,46	75,67
5	55	46,61	46,19	46,62
6	30	31,06	31,63	30,80
7	24	21,80	22,81	21,43
8	14	15,93	17,11	15,52
9	14	12,02	13,23	11,61
10	13	9,31	10,49	8,92
11	6	7,36	8,49	7,00
12	3	5,93	6,99	5,60
⩾13	45	41,46	60,23	36,83
	â	1,58652	1,25261	1,70548
	x̂	3,38364	2,67150	3,63735
	$\chi^2$10 g.l.	10,73	15,91	12,62
	distribuição	0,62106	0,89776	0,75430
	probabilidade	37,89	10,22	24,57
	resultado	não rej. H_0	não rej. H_0	não rej. H_0

* Utilizamos a edição a cargo de J. Vessereau, publicada por "Les Belles Lettres"[3].

** Para assegurar a indispensável coerência e comparabilidade dos dados do *Aetna* com os do *Corpus Tibullianum*, seguimos a norma lexicográfica adotada por Della Casa, implícita na lematização efetuada[1].

3.1.1 Comentários.

H_0:V_i = $E(V_i)$significa que os desvios entre os valores reais e os do modelo teórico não são significativos; isto é, os valores reais seguem a distribuição de Waring. A não rejeição da hipótese nula parece evidenciar que o modelo teórico (distribuição de Waring) se ajusta aos dados da realidade fornecidos pelo texto.

Note-se que o melhor ajuste do modelo é dado, no nosso exemplo, pelos estimadores de Irwin, embora os estimadores de Ratkowsky e mesmo os de Herdan também tenham apresentado resultados que podem ser considerados satisfatórios.

3.2. Aplicaremos o mesmo procedimento aos dados extraídos do *Corpus Tibullianum,* obtidos por contagem manual a partir do levantamento já citado de Della Casa, das "concordâncias" do *Corpus.* Alguns esclarecimentos preliminares se fazem necessários.

Inicialmente, acatando uma criteriosa observação de C. Muller [8] eliminamos da contagem todos os nomes próprios, os de acidentes geográficos e os gentílicos: com efeito, além do alto grau de arbitrariedade de sua escolha em textos literários, dá-se o caso, num *corpus* fragmentário constituído de poemas mais ou menos independentes um do outro, de esses nomes estarem sujeitos a constante renovação, sem a regularidade e a constância das repetições que se verificam no resto do vocabulário. O mesmo não ocorre, em tese, em textos unitários como o *Aetna,* anteriormente analisado, razão pela qual os nomes próprios foram, nesse poema, conservados. Entretanto, qualquer que seja a alternativa escolhida (manutenção ou eliminação), os resultados observados divergem de maneira pouco significativa, como tivemos oportunidade de verificar em nossos cálculos, em relação ao *Aetna;* a divergência, porém, torna-se muito mais significativa em relação ao *Corpus Tibullianum* — o que aliás, era esperado, pelo que acabamos de expor.

Com relação a esta última obra, uma dificuldade de outra ordem se apresenta: o problema da autoria. Sem entrar em pormenores, parece que esse problema pode ser equacionado, em linhas gerais, da seguinte maneira *: a autoria dos dois primeiros livros de Elegias, dos três em que modernamente costuma dividir-se o *Corpus,* é seguramente atribuída a Álbio Tibulo, poeta do I século a.C.; o terceiro livro, em sua maior parte, se não em sua totalidade, é indevidamente atribuído a Tibulo pela tradição manuscrita, quando na realidade fatos lingüísticos e estilísticos parecem invalidar essa atribuição, sugerindo antes a existência de outro autor (ou autores) para nós desconhecido, que teria composto, os poemas do terceiro livro procurando imitar a língua e o estilo de Tibulo. Esta é a teoria mais ou menos universalmente aceita e que parece destinada a ser definitiva, a não ser que algum fato novo, pouco provável aliás, como seria a descoberta de um desconhecido manuscrito, venha lançar reveladora luz sobre as trevas em que tateia a filologia clássica à procura de respostas a certas indagações. É na esperança de indicar um caminho — já que não podemos oferecer soluções — que o presente trabalho é proposto.

Mantivemos, para a nossa análise, a divisão já consagrada, e calculamos separadamente os subconjuntos constituídos pelos livros I e II de um lado, e pelo livro III do *Corpus* de outro lado, a fim de verificar se a hipótese nula é aceitável, ou, em outras palavras, se os efetivos reais das classes de frequência dos vocábulos dos dois subconjuntos obedecem à distribuição de Waring.

Obtivemos, assim, as tabelas que seguem:

* Os interessados poderão encontrar informações mais detalhadas sobre a controvertida questão da autoria do *Corpus Tibullianum* nos comentários à edição de "Les Belles Lettres", feitos por MAX PONCHONT[12].

TABELA 2

Valores observados e calculados para os nomes comuns do *Corpus Tibullianum* (Livros I e II)

Dados: N = 7844, V = 1834, V_1 = 842, V_2 = 303, V_3 = 171

Freqüências i	Valores observados V_i	Valores calculados Estimadores de Irwin $E(V_i)$	Valores calculados Estimadores de Herdan $E(V_i)$	Valores calculados Estimadores de Ratkowsky $E(V_i)$
1	842	842,00	842,00	842,00
2	303	351,94	341,75	311,58
3	171	185,03	179,27	152,41
4	128	110,99	108,23	99,71
5	74	72,58	71,46	67,46
6	66	50,45	50,22	48,68
7	38	36,70	36,95	36,79
8	41	27,66	28,16	28,78
9	22	21,45	22,07	23,13
10	18	17,02	17,69	19,00
11	17	13,77	14,45	15,89
12	13	11,32	11,99	13,48
13	12	9,44	10,09	11,58
14	6	7,97	8,59	10,06
15	9	6,80	7,39	8,82
16	7	5,85	6,41	7,80
17	6	5,08	5,61	6,94
≥18	61	57,95	71,67	119,89
â		1,83948	1,62607	1,17160
x̂		3,40081	3,00626	2,16604
$\chi^2$15 g.l.		25,30	23,02	51,74
distribuição		0,95394	0,91628	0,99999
probabilidade		4,60	8,37	0,00
resultado		rej. H_0	não rej. H_0	rej. H_0

3.2.1. Comentários.

Dois dos três modelos teóricos propostos recomendam a rejeição da hipótese nula, sendo que um deles — o construido pelos estimadores de Ratkowsky, que geralmente proporcionam resultados satisfatórios — não deixa qualquer margem de dúvida: o modelo não se ajusta de modo algum aos dados observados. O único resultado positivo é o apresentado pelo modelo construido sobre os estimadores de Herdan; no entanto, o erro sobr o qual se fundamenta o próprio modelo, como já foi examinado, nos autoriza a desconfiar dos resultados de sua aplicação, que devem por isso ser considerados suspeitos e admitidos com reservas. No nosso caso, a suspeição ao modelo de Herdan se fortalece na medida em que os dois outros modelos divergem dele, ao rejeitarem a hipótese nula. Isto significa que a distribuição de Waring não dá conta satisfatoriamente das distribuições de freqüencias dos vocábulos no subconjunto do *Corpus* que acabamos de analisar.

TABELA 3

Valores observados e calculados para os nomes comuns do *Corpus Tibullianum* (Livro III)

Dados: N = 4282, V = 1360, V_1 = 709, V_2 = 249, V_3 = 137

Freqüências i	Valores observados V_i	Valores calculados Estimadores de Irwin $E(V_i)$	Valores calculados Estimadores de Herdan $E(V_i)$	Valores calculados Estimadores de Ratkowsky $E(V_i)$
1	709,00	709	709,00	709,00
2	247,85	249	261,36	259,62
3	120,88	137	127,19	126,38
4	69,81	76	72,18	71,89
5	44,71	45	45,21	45,17
6	30,72	24	30,34	30,41
7	22,21	23	21,43	21,55
8	16,69	19	15,74	15,88
9	12,93	19	11,93	12,07
10	10,27	8	9,28	9,41
11	8,32	3	7,37	7,50
12	6,86	6	5,96	6,08
≥13	59,75	42	43,01	45,04
â		1,60344	1,29620	1,55805
x̂		3,34974	2,70789	3,25491
$\chi^2$10 g.l.		10,64	16,66	12,28
distribuição		0,61376	0,91777	0,73325
probabilidade		38,62	8,22	26,68
resultado		não rej. H_0	não rej. H_0	não rej. H_0

3.2.2 Comentários.

Os três modelos concordam na não rejeição da hipótese nula, o que parece indicar de forma irrecusável que a distribuição de Waring é adequada para descrever a distribuição real dos vocábulos no subconjunto ora analisado do *Corpus*, qualquer que seja o modelo de estimação utilizado. Aqui também, como no caso do *Aetna*, o melhor resultado é obtido através dos estimadores de Irwin, e o pior pelos de Herdan.

4. Conclusões.

Os dados apresentados até aqui se prestam, a nosso ver, não para uma solução definitiva do problema da autoria do *Corpus Tibullianum*, mas e tão somente para corroborar as conclusões de grande parte da crítica literária que não reconhece em Tibulo o autor único do *Corpus* que lhe foi atribuído.

A quantificação e a análise de dados liguísticos, na realidade, raramente oferecem respostas definitivas e cabais a dúvidas que porventura surjam; mas elas fornecem ao crítico literário e ao filólogo excelentes instrumentos que poderão ser utilizados em combinação com outros, dotando afirmações vagas e conclusões imprecisas de um rigor científico e de uma significância probabilística nada desprezíveis.

É nessa linha de raciocínio que se situa a conclusão central do nosso trabalho. Não se pretende "provar" que o *Corpus Tibullianum* foi escrito por dois ou mais autores; ou que, em outras palavras, deve-se atribuir a Tibulo apenas a autoria

dos livros I e II (e possivelmente de uns poucos poemas curtos do livro III, segundo alguns críticos), sendo o livro III de autoria de um poeta desconhecido, ou mesmo de vários (já se propuseram os nomes de Lígdamo, Válgio Rufo, Ovídio, Propércio, entre muitos outros): aqui como sempre cabe aos críticos a última, embora improvável, palavra. Nós apenas desejaríamos acrescentar mais um dado aos muitos que compõem a já *communis opinio* de que Tibulo não foi o único autor do *Corpus Tibullianum*. Nossa análise estatística aponta diferenças entre os dois subconjuntos do *Corpus* e diferenças significativas; a primeira delas — e talvez a mais relevante — se verifica na própria construção de um modelo teórico de distribuição de freqüencia que, se para um subconjunto (o livro III) se ajusta muito bem aos dados observados, para o outro subconjunto (os livros I e II) propõe uma concordância inaceitável; ora, parece pouco provável que isto tivesse ocorrido se autor do *Corpus* fosse um só.

4.1 Reconhecemos a fragilidade desta hipótese — uma presunção de probabilidade apenas. Mas a análise mais pormenorizada das unidades lexicais do Corpus revela fatos que, segundo cremos, favorecem a aceitação da hipótese: há vocábulos que aparecem em proporção muito maior no livro III do que no subconjunto formado pelos livros I e II (alguns exemplos: *aduersus*, 8 e 1 respectivamente; *aer*, 5 e 0; *alter*, *9 e 2; carus*, 10 e 3; *densus*, 6 e 1; *ergo*, 5 e 0; *maior*, 7 e 0; *minor*, 5 e 1; *pars*, 7 e 1; *pontus*, 5 e 1; *seu*, 27 e 13; *uel*, 14 e 4). Este é um fato de natureza estilística também capaz de sugerir, pela falta de homogeneidade na distribuição dos vocábulos pelas partes do texto, a duplicidade — ou mesmo a pluralidade — de autores. Há, evidentemente, casos notáveis do fenômeno inverso: vocábulos empregados um número de vezes muito maior nos livros I e II do que no livro III (exemplos: *ad*, 31 e 5 respectivamente; *ager*, 17 e 1; *agua*, 23 e 4; *bos*, 9 e 1; *caput*, 18 e 3; *plenus*, 11 e 0; *pes*, 32 e 4; *peto*, 10 e 0; *puer*, 21 e 4; *quotiens*, 6 e 0; *sequor*, 8 e 1; *sto*, 16 e 1; *tener*, 28 e 4; *tunc*, 25 e 7; *uua*, 10 e 1, entre muitos outros), quando a proporção esperada — e realmente observada em grande número de casos — seria a de 2:1, desde que o primeiro subconjunto contém pouco menos do dobro das unidades de texto contidas no segundo subconjunto. Observe-se que, de todos os exemplos apresentados, nenhum vocábulo é suscetível de revelar especialização lexical exigida por situações especiais dentro da temática desenvolvida no *Corpus;* trata-se, a nosso ver, de vocábulos (alguns deles meros instrumentos gramaticais desprovidos de conteúdo semântico) que pertencem ao léxico comum da poesia latina da época clássica, e o seu emprego diferenciado reflete apenas preferências individuais de cada autor — o seu estilo. Parece haver, portanto, uma sensível diferença de estilo entre as composições que constituem as duas partes em que convencionamos dividir o *Corpus* — fato, aliás, já abundantemente ilustrado pela crítica e hoje universalmente aceito.

4.2. Apesar da evidência dos fatos mostrados nas considerações expostas até aqui, procuramos buscar mais elementos que testassem e convalidassem a nossa hipótese, e os encontramos sob a forma de um índice que revelasse a riqueza lexical das duas partes do *Corpus Tibullianum* postas em confronto. A noção de riqueza lexical, como bem adverte MULLER[9] (especialmente Cap. 18 e segs.) a quem remetemos diretamente o leitor interessado, deve ser logo despojada de qualquer conotação elogiosa, subjetiva e impressionista, para ser considerada tão somente uma noção puramente técnica, suscetível de quantificação e de análise objetiva.

"Aplicado a um texto, o termo riqueza lexical é, pois, definido pelo número dos vocábulos, e nada mais" (p.116)*

O método dos índices, descrito por MULLER[9] e fundamentado na comparação dos parâmetros V (número de vocábulos), $V._1$ (número de vocábulos de freqüencia 1), f (freqüencia média) e q_1 (índice de repetição) de dois textos distintos, considerando N_a a extensão do texto constituído pelos livros I e II do *Corpus*, e N_b a extensão do livro III, não oferece resultado apreciável: não é possível rejeitar a hipótese de estrita igualdade de riqueza lexical entre os dois textos. Entretanto uma explicação, bastante óbvia, nos aponta o vício do resultado e nos obriga a percorrer outros caminhos: é que a diferença de extensão entre os dois textos é muito grande (N_a = 7844; N_b = 4282), e é bastante evidente que um texto de qualquer extensão terá provavelmente um vocabulário mais rico do que um texto de extensão igual a cerca da metade do primeiro. Isto significa que a riqueza do vocabulário de um texto é função da extensão desse mesmo texto. Uma comparação válida seria a efetuada entre o poema *Aetna* e o livro III do *Corpus Tibullianum*, aproximadamente de igual extensão (N_a = 4324; N_b = 4282); a tabela abaixo indica os resultados.

TABELA 4

Comparação dos índices

	C. Tib. I, II N_a = 7844		C. Tib. III N_b = 4282
V	1834	>	1360
V_1	842	>	709
$\tilde{f}$	4,28	>	3,15
q_1	0,54	>	0,48

Conclusão: $R_a \cong R_b$ (os dois textos não diferem em riqueza lexical).

Comparação dos índices

	Aetna N_a = 4324		C. Tib. III N_b = 4282
V	1446	>	1360
V_1	768	>	709
$\tilde{f}$	2,99	<	3,15
q_1	0,47	<	0,48

Conclusão: $R_a > R_b$ (o texto do Aetna é mais rico do que o livro III do C. Tib.)

No entanto, o índice $\tilde{f}$ (freqüencia média), menor em C. Tib. III do que em *C. Tib. I, II (3,15 < 4,28), sugere maior* riqueza de vocabulário naquele do que neste, muito embora isoladamente o fato não seja probante, porquanto $\tilde{f}$ varia com a extensão do texto. Verificaremos a seguir, por outros métodos, se realmente o vocabulário do livro III do *Corpus Tibullianum* é mais rico do que o dos livros I e II de Tibulo.

4.3 O método ideal seria aquele em que se obtivesse um índice de riqueza lexical independente do tamanho dos textos a ser comparados. Examinaremos aqui alguns entre os vários que têm sido propostos, embora todos eles apresentem alguma sensibilidade à extensão dos textos, variando segundo a extensão de cada texto.

4.3.1 O índice de riqueza lexical proposto por E. Brunet (*apud* 2), como base de cálculo para o vocabulário teórico de qualquer texto, obedece à fórmula.

$$W = N^{1/(V-b)^a}$$

onde b = 20 e a = 0,172.

* Não se deve confundir riqueza lexical com originalidade, excentricidade, ou emprego de termos raros, exóticos ou criativos. Aos interessados, recomendamos a leitura do artigo de N. MÉNARD[7] que procura correlacionar riqueza lexical e emprego de palavras raras a partir do exame de um corpus constituído de seis amostras extraídos de romances de seis autores franceses contemporâneos.

Aplicando essa fórmula aos nossos dados, obtivemos os seguintes valores:

no *Corpus Tib.*	I,II :	W =	11,79
no Corpus *Tib.*	III :	W =	11,29
no *Aetna*	:	W =	11,03

Como o valor de W é inversamente proporcional à riqueza do vocabulário de um texto, pode-se concluir que o vocabulário do livro III do *Corpus Tibullianum* é mais rico do que o vocabulário dos livros I e II; o do *Aetna*, por outro lado, é o mais rico dos três *.

4.3.2 Após criticar o índice W de Brunet, por não ser totalmente independente do tamanho do texto, DUGAST (2) propõe para o cálculo da extensão teórica de um vocabulário a adoção da relação logarítima log V/log N, simples e pouco sensível às variações de extensão dos textos; a partir dessa relação, o autor formula um novo índice de riqueza lexical de um texto:

$$U = \frac{n_2}{n\text{-}v} \quad **$$

onde n = log N e
v = log V.
A aplicação da fórmula de Dugast forneceu os seguintes valores:

no *Corpus Tib.* I II :	U	=	24,02
no *Corpus Tib.* III :	U	=	26,35
no *Aetna*	U	=	27,54

variando U em razão diretamente proporcional à riqueza do vocabulário de um texto, os resultados confirmam plenamente as conclusões expostas no item anterior.

4.3.3. Um outro método para comparar a riqueza do vocabulário de textos de extensão diferente baseia-se no princípio da redutibilidade do mais extenso para o tamanho do mais curto; isto é, calcula-se a estrutura lexical que um texto A de tamanho N passaria a ter, se a sua extensão fosse reduzida ao tamanho N' de um texto B. No nosso caso, o objetivo é saber qual seria o total V' de vocábulos do subconjunto *C. Tib.* I e II (e eventualmente a distribuição de freqüencias dos V_i), se reduzíssemos esse subconjunto de N = 7844 para N' = 4282.

Dois são os processos que podem ser utilizados para o cálculo do vocabulário teórico V' : o primeiro é o cálculo pelo modelo binominal da distribuição teórica, segundo o qual

$$E(V') = V - \Sigma q^i V_i ;$$

o segundo é o cálculo pelo conhecido modelo de WARING[8,9]. A lei binomial opera especialmente sobre as freqüências mais baixas, mas perde a eficácia com muita rapidez; a distribuição de Waring, por sua vez, reduz proporcionalmente os efetivos de todas as classes de freqüência indistintamente, o que pode ser questionado.

A combinação dos dois modelos (o binomial e o de Waring) deve-se a M. Dubrocard, e as tabelas numéricas que permitem a aplicação desse método a qualquer texto foram elaboradas por RATKOWSKY e HANTRAIS[11].

---

* C. MULLER[9] (Anexo 7), reproduz a fórmula de Brunet sem a correção representada pelo valor b:

$$W = N^{V^{-a}},$$

propondo a transformação para um índice R = (25-W)/1,5, tal que O<R>10; neste caso, obtivemos os seguintes valores:

no *Corpus Tib.* I, II: W = 11,733; R = 8,84
no *Corpus Tib.* III: W = 11,217; R = 9,19
no *Aetna*: W = 10,966; R = 9,36.

Como, ao contrário de W, R é maior quanto mais rico for o vocabulário de um texto, aceitar-se-á a mesma conclusão.

** E não como constou, por um lamentável lapso de revisão: $U = n^2 (n\text{-}v)$.

Depois de procedermos às interpolações necessárias, pois que a tabela não fornece os valores intermediários obtidos por nós (N'/N = 4282/7844 = 0,546, com uma taxa de redução de N igual a 0,454; V_1/V = 842/1834 que dá P_1 = 0,46; N/V = 7844/1834, que é a $\hat{I}$ = 4,28), o cálculo do vocabulário teórico do texto N'indica:

V' = 1834 X 0,652 = 1196.

Isto significa que, se o texto do *Corpus Tibullianum,* livros I e II, fosse reduzido ao mesmo tamanho do texto do *Corpus Tibullianum,* livro III, o novo texto assim obtido contaria com um vocabulário de aproximadamente 1196 vocábulos diferentes, com uma distribuição de freqüências que não nos interessou calcular, mas que muito provavelmente seguiria a distribuição de Waring. Ora, como na realidade o texto do *Corpus Tibullianum,* livro III, possui 1360 vocábulos, contra apenas 1196 do texto reduzido do *Corpus Tibullianum,* livros I e II, conclui-se imediatamente que o vocabulário do *Corpus Tibullianum,* livro III, é sensivelmente mais rico do que o vocabulário do outro subconjunto.

Em outras palavras, se Tibulo tivesse escrito suas Elegias utilizando um total de 4282 palavras em lugar das 7844 que realmente usou, teria utilizado provavelmente pouco menos de 1200 vocábulos diferentes, ao passo que o autor (os autores) do livro III do *Corpus* se valeu de um vocabulário bem mais rico, de 1360 unidades.

Parece difícil acreditar que essa diferença seja meramente aleatória. Tudo aponta para a existência de mais de um autor para o *Corpus Tibullianum,* analisado neste artigo.

Esta é a nossa conclusão final.

4.4. A quantificação dos dados do *Corpus Tibullianum* e a utilização de alguns procedimentos estatísticos para a sua análise levaram-nos às reflexões aqui apresentadas. Os autores destas linhas não reivindicam nenhuma originalidade; apena desejam chamar a atenção dos estudiosos, lingüistas ou estatísticos, para os inegáveis benefícios que a interdisciplinariedade pode trazer às pesquisas em vários setores. As ciências humanas, e a lingüística em particular, podem enriquecer-se sobremaneira aliando os tradicionais métodos qualitativos de análise a rigorosos métodos quantitativos; a estatística, por sua vez, poderá encontrar nos inúmeros dados fornecidos pela lingüística um fecundo campo de aplicação de suas teorias e de suas técnicas de análise. Em nosso país, quase tudo está por ser feito; poucos são os trabalhos de que temos conhecimento no domínio da estatística lingüística. Não demos certamente o primeiro passo, mas esperamos não ter dado o último.

CARRATORE, E. Del & WADA, C.Y. Le **Corpus Tibullianum**: une analyse quantitative. **Alfa**, São Paulo, 25:31-47, 1981.

ABSTRACT: D'après les données observées dans le Corpus Tibullianum *et soumises às des procédés statistiques, les auteurs prétendent renforcer l'opinion, généralement admise, que cette oeuvre n'a pas été composée par un poète unique, et ils admettent donc la pluralité d'auteurs. Pour cela, ils utilisent la distribution de Waring et les indices les plus récents de richesse lexicale, et ils signalent, entre les deux parties dont ils ont divisé le* Corpus, *les différences qui renforcent l'hypothèse initiale.*

KEY-WORDS: Statistique lexicale; distribution de Waring; estimateurs; indices de richesse lexicale; extension théorique du vocabulaire.

CARRATORE, E. Del & WADA, C. Y. O **Corpus Tibullianum**: uma análise quantitativa. **Alfa**, São Paulo, **25**:31-47, 1981.

REFERÊNCIAS BIBLIOGRÁFICAS

1. DELLA CASA, Adriana. *Le concordanze del Corpus Tibullianum*. Genova, Istituto di Filologia Classica e Medioevale, 1964.
2. DUGAST, Daniel Sur quoi se fonde la notion d'étendue théorique du vocabulaire. *Le Français moderne*, Paris, *46* (1): 25-32, 1978.
3. L'ETNA (texte établi et traduit par J. Vessereau), 2.éd. Paris, Les Belles Lettres, 1961.
4. HERDAN, Gustav. *Quantitative linguistics*. London, 1964; *apud* MULLER, C. *Principes et méthodes de statistique lexicale*. Paris, Hachette, 1977.
5. IRWIN, J. O. The place of mathematics in medical and biological statistics. *Journal of the Royal Statistical Society*, London, A-*126* (1): 1-41, 1963.
6. JOHNSON, N. L. & KOTZ, Samuel. *Discrete distribuition*. Boston, Houghton Mifflin, 1970.
7. MÉNARD, Nathan Richesse lexicale et mots rares. *Le Français moderne*, Paris, *46* (1): 33-43, 1978.
8. MULLER, Charles. Peut-on estimer l'étendue d'un lexique? *Cahiers de Lexicologie*, Besançon, *27* (2): 3-29, 1975.
9. MULLER, Charles. *Principes et méthodes de statistique lexicale*. Paris, Hachette, 1977.
10. RATKOWSKY, D.A. Une nouvelle approche concernant l'application de la distribution de Waring aux fréquences des vocables dans les textes littéraires. *Cahiers de Lexicologie*, Besançon, *34*(1): 3-18, 1979.
11. ATKOWSKY, D.A. & HANTRAIS, Linda. Tables for comparing the richness and structure of vocabulary in texts of different lengths. *Computers and the Humanities*, New York, *9*(2):69-75, 1975.
12. TIBULLE. *Elégies*. (texte établi et traduit par Max Ponchont), 4.éd. Paris, Les Belles Lettres, 1955.

ANEXO

A série de Waring foi obtida por esse matemático inglês no séc. XVIII, pela expansão da função $\frac{1}{x-a}$, através da fórmula de interpolação polinomial de diferenças finitas descendentes de Newton-Gregory:

$$\frac{1}{x-a} = \sum_{i=0}^{\infty} \binom{-a}{i} \Delta^i \left(\frac{1}{x}\right)$$

com

$$\Delta^i \left(\frac{1}{x}\right) = \frac{(-1)^i i!}{x^{[i+1]}}$$

onde

$$x^{[1]} = x, \quad x^{[2]} = x(x+1), \ldots, x^{[i+1]} = x(x+1) \ldots (x+i)$$

e portanto:

$$\frac{1}{x-a} = \binom{-a}{0} \Delta^0 \left(\frac{1}{x}\right) + \binom{-a}{1} \Delta^1 \left(\frac{1}{x}\right) + \binom{-a}{2} \Delta^2 \left(\frac{1}{x}\right) + \ldots$$

que é igual a:

$$\frac{1}{x} + \frac{a}{x(x+1)} + \frac{a(a+1)}{x(x+1)\ (x+2)} + \ldots + \frac{a^{[i]}}{x^{[i+1]}} + \ldots$$

Como a serie é convergente, pois x>a>o, a multiplicação da série por uma constante (x-a) produz uma distribuição discreta de probabilidades (6), a que Irwin denominou distribuição de Waring, e cuja função densidade é dada por:

$$P(I=i) = (x-a) \frac{a^{[i]}}{x^{[i+1]}} \quad i = 0,1,2\ldots$$

Essa distribuição discreta de probabilidades (de Waring) é um caso particular de uma distribuição mais geral formulada por Irwin, denominada distribuição fatorial inversa (5).

Entre os parâmetros que uma distribuição de probabilidades possui, os que melhor a caracterizam e a definem são a esperança matemática (que se identifica com a média) e a variância. Dada a função densidade de probabilidades, calculam-se a média e a variância que, no caso da distribuição de Waring, são dadas respectivamente por:

$$\mu_1 = \frac{a}{x-a-1}$$

e

$$\sigma^2 = \frac{a(x-1)\ (x-a)}{(x-a-1)^2(x-a-2)}$$

pode-se verificar facilmente que, quando $1 < (x-a) < 2$, a variância é infinita.

Os parâmetros μ_1 e σ_2 (populacionais) podem ser estimados por μ_1 e s^2 (amostrais).

Para a estimação dos parâmetros a e x, necessários para o cálculo das probabilidades esperadas da distribuição de Waring, Irwin utilizou a média observada $\hat{\mu}_1$ e a probabilidade associada à classe

de freqüência 1, $\hat{p}_1$, em lugar de utilizar s^2, visto que não é adequada como estimador de σ^2, infinita dentro do intervalo de 1 a 2.

Como, por definição,

$$P_1 = \frac{x-a}{x} \quad e\, q_1 = (1-p_1) = 1 - \frac{(x-a)}{x} = \frac{a}{x},$$

então

$$x = \frac{a}{q_1}.$$

Como

$$\mu_1 = \frac{a}{x-a-1},$$

obtém-se o valor de

$$x = a(1 + \frac{1}{\mu_1}) + 1$$

de onde

$$a(1 + \frac{1}{\mu_1}) = x-1$$

Substituindo nesta última relação o valor de x por $\frac{a}{q_1}$, temos:

$$a(1 + \frac{1}{\mu_1}) = \frac{a}{q_1} - 1,$$

que resulta em

$$a + \frac{a}{\mu_1} - \frac{a}{q_1} = -1,$$

que é igual a

$$a(1 + \frac{1}{\mu_1} - \frac{1}{q_1}) = -1,$$

e finalmente

$$a = \frac{1}{\frac{1}{q_1} - \frac{1}{\mu_1} - 1}$$

e portanto o estimador â em função de $\hat{\mu}_1$ e $\hat{q}_1$ é dado por:

$$\hat{a} = \frac{1}{\frac{1}{\hat{q}_1} - \frac{1}{\hat{\mu}_1} - 1}$$

Retomemos o valor de

$$x = a(1 + \frac{1}{\mu_1}) + 1;$$

substituindo nesta relação o valor de a por

$$\frac{1}{\frac{1}{q_1} - \frac{1}{\mu_1} - 1}$$

obteremos:

$$x = \left[\frac{1}{\frac{1}{q_1} - \frac{1}{\mu_1} - 1} \left(1 + \frac{1}{\mu_1}\right) \right] + 1 ,$$

que dá como resultado

$$x = \frac{1}{\frac{1}{q_1} - \frac{1}{\mu_1} - 1} + \frac{1}{\mu_1 \left(\frac{1}{q_1} - \frac{1}{\mu_1} - 1\right)} + 1 ;$$

então

$$x = \frac{\frac{\mu_1}{q_1}}{\mu_1 \left(\frac{1}{q_1} - \frac{1}{\mu_1} - 1\right)}$$

que é igual a

$$x = \frac{\frac{1}{q_1}}{1 \frac{1}{q_1} - \frac{1}{\mu_1} - 1}$$

e portanto o estimador $\hat{x}$ em função de $\hat{\mu}_1$ e $\hat{q}_1$ é dado por:

$$\hat{x} = \frac{\frac{1}{\hat{q}_1}}{\frac{1}{\hat{q}_1} - \frac{1}{\hat{\mu}_1} - 1}$$