Aplicação dos Métodos Estatísticos e da Teoria da informação e da Comunicação na Análise Lingüística: estudo da linguagem jornalística

Laís **A. Kibeiro**
Professora do Curso de Pós-Graduação (Mestrado) em Ciência da Informação IBBD/UFRJ

RESUMO

Possibilidades de utilização de métodos estatísticos e conceitos de Teoria da Informação e da Comunicação na análise lingüística foram demonstrados. A amostra foi extraída dos editoriais do Jornal do Brasil de 1959 a 1973. Os resultados confirmam a validade da Lei de Zipf com as necessárias restrições para a língua portuguesa. Palavras-chave foram determinadas indicando o ponto de vista dos editoriais e outras variáveis lingüísticas são discutidas.

A pesquisa examina os recursos, conceitos e possibilidades instrumentais dos métodos estatísticos e da Teoria da Informação e da Comunicação.
Apóia-se em hipótese de trabalho veiculada por autores como Charles Muller[1], CB Williams[2], George Kingsley Zipf[3] e Pierre Guiraud[4], entre outros, supondo uma aplicabilidade dos métodos estatísticos na análise lingüística, numa visão crítica do alcance desse tipo de enfoque.
Observa que essa direção tem sido tomada por vários pesquisadores, depois do desenvolvimento da Teoria da Informação e da Comunicação. Mencionamos alguns trabalhos, elaborados no Brasil: os de Affonso Romano de Sant'Anna[5], Arlindo Chave:.[6], Décio Pignatari[7,8] e Tulo Hostílio Montenegro[9,10].
Entre os métodos, escolhemos os propostos por Zipf[11] e Williams[12].
Definimos o corpus a partir de amostra extraída da linguagem jornalística. A escolha foi feita basicamente pelos seguintes motivos:

— caracterização de uma determinada área do universo lingüístico: a linguagem empregada nos jornais, onde se pode prever certa homogeneidade no léxico, empregado dentro das normas da sintaxe regular.

Ao nível formal, a linguagem jornalística tende a apresentar um menor índice de desvios em relação à chamada normalidade de cada sistema lingüístico, ao contrário de uma literatura proposta pela Estética informacional (onde o desvio caracteriza exatamente a informação, a novidade).

— o estudo desse gênero de discurso, sob o ângulo lingüístico, serve de base a análises científicas mais amplas, como o da ideologia de que esse discurso é exteriorização.

l - O CORPUS

O corpus: Propõe-se uma amostra de 10.093 unidades. Os textos escolhidos foram extraídos dos editoriais do Jornal do Brasil, entre os exemplares publicados durante o período 1959-1973.
Supõc-sc que esse discurso varie no decorrer desse período. Para essa variação julgou-se que seria uma divisão em períodos mais curtos.

1959-1963
1964-1968
1969-1973

As amostras foram extraídas dos editoriais. Considerou-se que o editorial de cada uma das seções do jornal (política, artes, esporte etc.. .) tem um determinado tipo de leitor. Entretanto, se essas diferentes seções podem estabelecer um apelo aos vários gêneros de leitor, cada periódico se autocaracteriza através do editorial, onde propõe uma imagem ao receptor.
Assegurou-se estatisticamente a validade da amostra (nível de significância do resultado do teste situado no intervalo (0,05; 0,2). *

* Os cálculos estatísticos foram realizados por Henrique Gurvitz do IBGE, Rio de Janeiro, GB.

2 - MÉTODO

Para o levantamento dos dados utilizou-se um programa que lê os cartões que contém o texto e pesquisa as unidades nele utilizadas.
As linguagens empregadas foram SNOBOL 4 — destinada especificamente a lidar com cadeias de caracteres, registrando-os. — FORTRAN — em seguida, foi utilizada essa linguagem, que permite ler as ocorrências de unidades idênticas, imprimindo essa contagem.
Os resultados foram aferidos a partir da leitura dos relatórios emitidos pelo computador. Os dados em questão foram apresentados a um sistema computacional, a fim de que este processasse as informações. Cumprindo as etapas lógicas do processamento de dados, foram apresentados ao referido sistema os objetivos do processamento e as informações a serem processadas.

3 - DESCRIÇÃO DOS RESULTADOS

— A *equação de Zipf* — A pesquisa realizada demonstra a comprovação, para a linguagem jornalística em português, da relação $ab^2=k$, a chamada lei do contexto de Zipf, com as restrições devidas.[13]

Na equação de Zipf:
b: nº de ocorrências de cada palavra,
a: nº de palavras que ocorrem *h* vezes no texto.
k: constante.
O conceito de palavra, linguisticamente polêmico, indica uma reunião de diversos componentes, dos quais se examinam, em cada tipo de análise, um ou mais níveis (sintático, semântico, fonológico; do ponto de vista paradigmático ou do ponto de vista sintagmático). No caso presente designa a unidade, separada por intervalos em branco, no texto preparado para ser submetido ao processamento dos computadores.
A aparente ordenação proposta por Zipf se exterioriza, portanto, quando é aplicada a textos escolhidos aleatoriamente em português. Isto é: funciona para o português, assim como, segundo constatações de Zipf, funciona para o latim de Plauto, para o chinês de Pequim (coloquial) e para o inglês dos jornais americanos (pesquisa de Eldridge).
Dentro do procedimento estatístico, estimou-se, a partir da distribuição de a e b no texto, qual seria o valor *h* na equação $k = ab^h$, através do método de estimação dos mínimos quadrados.
Tendo sido obtido o valor de *h*, verificou-se se ele diferia de 2 apenas por perturbações aleatórias, tcndo-se utilizado, para tanto, um teste de hipótese estatística onde a variável

$$t' = \frac{S_1 \sqrt{n-2}\ (h-2)}{S2,\sqrt{1-r^2}}$$

tem distribuição *t* de Student com n—2 graus de liberdade e

S1 : desvio-padrão de log b

S1 : desvio-padrão de log b

r : coeficiente de correlação entre $\log_e$ a e $\log_e$ b

n : Número de classes de freqüência

A partir do texto citado, obteve-se a seguinte distribuição, levando-se em consideração as palavras que nele aparecem até 26 vezes:

b	*a*
1	2161
2	441
3	187
4	101
5	69
6	34
7	34
8	18
9	9
10	17
11	4
12	6
13	9
14	4
15	3
16	5
17	2
18	4
19	4
20	4
21	4
23	1
24	o
25	1
26	1

Para a estimação de *h* efetuou-se a seguinte transformação na equação

$$K = ab^h$$

$$\log K_e = \log_e a + h \log_e b$$

$$\log a_e = \log_o K - h \log_e b$$

Procedeu-se à determinação de uma reta de regressão linear que nos forneceu os seguintes parâmetros:

r = 0,9780 (coeficiente de correlação entre log *a* e log *b)*

h = 2,2261

log K = 7,6040

K = 2006,75

Portanto, a equação obtida foi:

$$ab^{2,2261} = 2006,75$$

O passo seguinte foi a verificação da significância do desvio de *h* em relação a 2, tendo-se procedido à determinação de t', obtendo-se:

t' = 2,2837

Em vista desse valor foi rejeitada a hipótese nula a um nível de significância de 0,05; isto quer dizer que a probabilidade de o valor encontrado para *h* ser diferente de 2 em função somente de perturbações aleatórias é menor do que 0,05, ou seja, a probabilidade de o valor do expoente ser "realmente" diferente de 2 é maior do que 0,95. De onde se conclui que devemos rejeitar 2 para o valor de *h*, embora o valor de - 0,9780 encontrado para *r* nos conduza à aceitação dos termos gerais da lei de Zipf, rejeitando-se apenas o valor particular do expoente.

Além da aceitação dessa lei referente à organização das unidades no texto, verificamos, a partir dessa variação de expoente, que a repetição das palavras é menos freqüente no português do que no inglês. Podemos estabelecer essa dedução, observando-se

em $K = ab^h$ que $a = \frac{K}{b^h}$

e no inglês $a = \frac{K_1}{b^2}$

e no português $a = \frac{K_2}{b^{2,2261}}$

Por mais diferentes que possam ser KI K_2, à medida que *b* cresce, K_1 e K_2 permanecendo constantes, temos:

— a partir de um determinado valor *de b o* correspondente valor de *a* para o português será sempre menor do que para o inglês. Isto pode ser explicado, em relação ao texto: em português, a orientação é não repetir as palavras muito seguidamente, recorrendo-se à sinonímia.

Com relação à assertiva de Zipf de que a variação do valor de *h* em ab = k está relacionado ao tamanho da amostra pode-se concluir que, cm termos estatísticos, essa afirmativa é inconsistente. Chega-se ao resultado de que esse expoente está relacionado, antes, à natureza da língua examinada.

Como se vê, constata-se a existência de um equilíbrio formal no discurso, no que se refere à disposição das unidades (independentemente de sua função lingüística e de sua condição de lexemas ou relacionantes).

— A *verificação do princípio de economia lingüística proposto por Zipf* —

Constata-se que no sistema lingüístico existe urna lei do menor esforço, manifestando-se da seguinte forma: predominam as unidades constituídas de um menor número de letras, isto é, a freqüência está inversamente relacionada ao comprimento da unidade.

Os resultados mostram que, no português, entre as 14 palavras mais freqüentes da língua, as unidades de menor comprimento apresentam as freqüências mais altas; essas unidades são todas relacionantes, isto é, referem-se a relações e categorias gramaticais e não à representação dos objetos (por relacionantes designamos o que, p. ex., A. Martinet[14] chamaria de morfemas e B. Pettier[15] de gramemas).

— *O levantamento dos lexemas, para a possível análise ideológica do discurso* —

Com as restrições de uma análise atomística, da unidade lingüística isolada, procedeu-se ao exame dos lexemas existentes nos textos, caracterizando as palavras-chaves. Por palavra-chave se designam os lexemas de freqüência mais alta, na amostra examinada.

As palavras-chaves variam de acordo com o período examinado, permitindo uma caracterização desses a partir do critério em questão.

No lº período examinado (1959-1963) temos, portanto, como palavras-chaves do contexto dos editoriais — *política, presidente, Brasil.* No 2º período (1964-1968) a mais importante informação, em comparação com o período anterior, é o aparecimento da palavra-chave *nacional,* antes não incluída entre as unidades lingüísticas mais freqüentes. As palavras-chaves são: *governo, inflação, nacional e* esboça-se através dessa tríade a situação brasileira, suas preocupações c diretrizes políticas: a discussão do problema inflacionário, o culto ao nacional, a valorização do governo do país.

No 3º período (1969-1973) ocorre: - o desaparecimento da palavra *nacional;* — o aumento da freqüência da palavra *inflação;* — a redução da freqüência da palavra *continente.*

Nesse caso, aparece a principal restrição que esse sistema pode oferecer: na comparação dos resultados, certos dados ficam limitados a um determinado espectro de tempo (p. ex.: os termos de ordem administrativa). Os termos referentes a questões administrativas são restritos, não caracteri-

zando todo o período, embora, por vezes, a amostra escolhida apresente uma alta freqüência dessas unidades.

Mesmo em se tratando de uma pesquisa de signos isolados, com as limitações decorrentes, também para os estudiosos de sociologia e de política e da ciência interdisciplinar da sóciolingüística fica em aberto o exame da linha definida por essas palavras-chaves, como um meio para o exame ideológico do discurso e das variações desse discurso, durante o espaço de quatorze anos. É dentro de um quadro de apresentação ideológica por exemplo, que. dentro desse tempo, o jornalismo brasileiro discute a inflação e reitera o nacionalismo.

— *o funcionamento da análise matemática do estilo* — Supõe-se que há um "estilo" de linguagem escrita especificamente utilizado pelos jornais.

O 1° método diz respeito ao exame quantitativo do discurso, definindo a autoria o "estilo", por meio da contagem do nº de unidades que compõem os períodos. O 2º refere-se ao comprimento das unidades e à freqüência das palavras compostas por x, x +1, x+2,... x+n elementos ou sub-unidades.[16]

No l" caso, o resultado, após cálculos estatísticos, mostra ser o número médio de 27, 47 palavras p/período.

No 2º, como já observamos, constata-se que o comprimento das palavras está relacionada à sua freqüência, isto é, as palavras mais curtas são as mais freqüentes.

— *a redundância e a taxa de informação de vagais e consoantes* — Reportamo-nos a uma experiência nesse setor feita por Décio Pignatari e Luiz Ângelo Pinto[17]. Comprovamos que a conclusão sobre a redundância e taxa de informação de vogais e consoantes (vogais mais redundantes, consoantes mais informativas) é alterada quando se considera apenas o início da palavra. Em nossa pesquisa temos o seguinte resultado: cm posição inicial, a recíproca é verdadeira, isto é, vogais são merios redundantes e mais informativas do que consoantes. Os conceitos de redundância e taxa de informação são os utilizados pela Teoria da Informação c da Comunicação. Esses conceitos são explicitados, p. ex., na obra da A. Moles [18]. A informação é função da improbabilidade da mensagem recebida. A redundância é a repetição, que previne o erro dentro do sistema de comunicação.

4 - CONCLUSÃO PRINCIPAL

Julgamos poder oferecer, com esse trabalho, alguns dados que permitam ponderar sobre a contribuição das técnicas computacionais na lingüística.

Consideramos que, embora esse tipo de enfoque nos conserve ao nível do discurso, isto é, na superfície e não discuta a máquina de elaboração ao sistema lingüístico, contribui com esclarecimentos para o conhecimento desse discurso e de seu funcionamento, apresentando essa análise as necessárias características de informação (nova).

5 - REFERÊNCIAS BIBLIOGRÁFICAS

1 — MULLER, C. *Initiation à Ia statistique.* Paris, Larousse, 1968.

2 - WILLIAMS, C. B. *Style and vocabulary; numerical studies.* London, Charles Griffin & Co., 1970.

3 — ZIPF, G. K. *The Psycho-Biology of language.* Boston, Houghton Mifflin, 1935.

4 — GUIRAUD, P. *Langage et théorie de la communication: communication et information.* In: MARTINET, A., ed. *Le langage.* Bruxelles, Gallimard, 1968.

5 - SANT'ANNA, A. R. *de.Drumond, o gauche no tempo.* Rio de Janeiro, Lia Editor, INL, 1972.

6 — CHAVES, A. Identificação estatística das cartas chilenas. *Revista Brasileira de Estatística,* Rio de Janeiro :307, abr./jun. 1941.

7 - PIGNATARI, D. & PINTO, L. A. Crítica, criação e informação. *Invenção,* (4): 17-33, dez. 1964.

8 - PIGNATARI, D. *Semiótica e literatura.* São Paulo, Perspectiva, 1974.

9 - MONTENEGRO, T. H. *Análise matemática do estilo.* Rio de Janeiro, IBGE 1956.

10 - MONTENEGRO, T. H. O Comprimento do período como característica do estilo. *Revista Brasileira de Estatística,* Rio de Janeiro, 63:3-5, jul./set. 1955.

11 - ZIPF, G. K. op. cit.

12 - WILLIAMS, C. B. op. cit.

13 - ZIPF, G. K. op. cit.

14 - MARTINET, A. *Elementos de lingüística geral.* Trad de J. Morais Barbosa. Lisboa, Liv. Sá da Costa, 1968.

15 — POTTIER, B. *Presentation de Ia. lingüística.* Trad. de Antonio Quilis. Madrid, Alcalá, 1968.

16 - WILLIAMS, C. B. op. cit.

17 - PIGNATARI, D. & PINTO, L. A. op. cit.

18 — MOLES, A. *Teoria da informação e percepção estética.* Trad. de Helena Parente Cunha. Rio de Janeiro, Tempo Brasileiro, 1969.

ABSTRACT

Possibilities of using statistical methods and concepts of Communication and Information Theory in the linguistic analysis have been demonstrated. The sample was taken from newspaper editorial language of the Jornal do Brasil from 1959 to 1973. The results confirm the validity of the Zips's Law with the necessary restrictions for the Portugueses language. Key-words were determined indicating the ideological point of view of the analysed editorials and other linguistic variables were discussed.