# ESTATÍSTICA LINGÜÍSTICA

**Maria Tereza de Almeida Camargo**

## INTRODUÇÃO

Tanto quanto conheço os trabalhos lingüísticos realizados no Brasil (e para o português em geral), não sei de pesquisas sistemáticas realizadas nesse setor quer por lingüistas nossos, quer aplicações à nossa língua. Conheço apenas alguns estudos esparsos dedicados a problemas específicos em português. Por essas razões, pareceu-me lícito nesse encontro de lingüistas brasileiros, apresentar um relatório de tipo informativo, expondo e discutindo sucintamente o que se tem feito nesse domínio entre especialistas estrangeiros, para levantar a questão da aplicabilidade dos métodos da ciência estatística à nossa língua. Justifico-me assim de antemão por não apresentar um trabalho original, mas de fornecer simplesmente ocasião para discutirmos sôbre um setor importante da lingüística que tem sido descurado entre nós.

A minha comunicação dividir-se-á em três tópicos principais: 1) a Estatística e as Matemáticas como instrumento de pesquisa lingüística; 2) os problemas e métodos da Estatística Lingüística; 3) aplicações da Estatística Lingüística.

## I

Uma dificuldade básica com que todos os lingüistas deparamos ao afrontar a Estatística Lingüística é a da barreira instrumental que significa para nós o aparato técnico de que se serve essa ciência. A maioria dos lingüistas recuará com horror diante de tratados como o **The Calculus of Linguistic Observations** de Herdan, e torcerá o nariz a uma tábua de $X^2$ das Geórgicas de Virgílio. Por outro lado, os matemáticos e estatísticos que se têm dedicado à Lingüística Matemática não

têm suficiente formação lingüística para equacionarem devidamente os problemas lingüísticos dentro do universo estatístico. Muitas vêzes, como bem o denunciou Greimas num "Colóquio de Estatística e de Análise Lingüística" (Strasbourg, 1964), perdem-se em labirintos matemáticos tratando de problemas que não interessam à lingüística moderna, ou utilizam conceitos lingüísticos ultrapassados. Êsse desnível básico na nossa plataforma de sondagem dos problemas da Estatística Lingüística, levou os lingüistas e matemáticos presentes ao Colóquio acima referido a concluir que seria desejável que os estudantes inclinados aos estudos lingüísticos tivessem uma formação estatística elementar, durante os seus anos de licença universitária, como já acontece em outros domínios das Ciências Humanas — Sociologia, Psicologia. Êsse desnível explica também os desconcertos encontrados nos tratados e estudos que aplicam as técnicas da Estatística Lingüística — curiosa a dualidade antagônica do tratamento matemático do material lingüístico e do tratamento lingüístico do instrumental matemático. Lembramos pois, desde já, que é absolutamente indispensável uma formação estatística elementar, a fim de que o lingüista possa avaliar os resultados da Estatística, quando aplicada ao universo lingüístico.

Uma vez admitida a validade e o interêsse da Estatística no domínio lingüístico, um passo ulterior será o do estabelecimento de uma plataforma comum para essas pesquisas, conciliando dados lingüísticos e elementos estatísticos. Aqui são freqüentemente os estatísticos que se queixam do empirismo lingüístico e da anarquia de critérios reinantes entre os lingüistas. Na verdade, em todos os outros domínios da Língua, com exceção do fonético, os trabalhos dos lingüistas caracterizam-se por um largo contingente subjetivo. Por exemplo: dificilmente coincidem os lingüistas quanto à definição do vocabulário e mais ainda divergirão êles quando tiverem que decidir sôbre as unidades léxicas em uma compilação vocabular. Se passarmos ao nível morfêmico e sintático, as divergências serão ainda maiores. Ora, a estatística precisa partir de critérios seguros e bem estabelecidos para proceder à compilação de suas amostras. Sobretudo se pensarmos na automação das pesquisas dessa natureza, o problema torna-se ainda mais agudo, pois para que a máquina possa operar eficazmente é necessário apresentar-lhe um programa definido, definições objetivas, a partir do que ela fornecerá os resultados. Se êsses critérios não forem lingüisticamente válidos, ou pelo

menos se forem imprecisos, os resultados obtidos não terão significação lingüística.

No polo oposto, levanta-se a grita dos lingüistas contra os seus confrades matemáticos. Além do tecnicismo rebarbativo de suas fórmulas, queixam-se êles mui justamente de que alguns matemáticos utilizem a língua como instrumento de elucubrações abstratas, fazendo matemática por si mesma e esquecendo a língua como objetivo essencial de suas pesquisas. Para nós, dizem êles, interessam-nos métodos matemáticos que nos ajudem a penetrar e a descrever de modo mais exato o universo lingüístico e nada mais. Outras vêzes, são os escolhos do avanço dessa ciência que os deixam perplexos. Afinal, a tão propalada, aplicada e discutida lei de Zipf para a freqüência das palavras na língua, não tem nenhum valor matemático? Assim pontificam matemáticos ilustres como Herdan que têm demonstrado de modo meridianamente arrevesado a inverossimilhança dessa lei. Essas e outras contradições da Estatística Lingüística devem pôr de sobreaviso o lingüista não afeito aos meandros do mundo estatístico.

Feitas essas ressalvas, chegamos a uma afirmação axiomática: defender a aplicação dos métodos estatísticos no domínio da língua, significa formular a crença de que a língua é um código cujos símbolos obedecem a certas freqüências determinadas e previsíveis. Em outras palavras, empregando o jargão estatístico: a língua é uma população e as realizações do discurso podem ser consideradas como amostras dêsse universo. É só com base nesse postulado básico que podemos continuar aprimorando a nossa técnica no estudo estatístico da língua. Uma pergunta deve aflorar de imediato a um lingüista desacostumado a êsses horizontes: com que objetivo penetraremos em tal dédalo? que utilidade para nós poderão ter essas técnicas? Responderemos que tanto o lingüista preocupado essencialmente com a ciência da linguagem, como o historiador das línguas, o filólogo inclinado aos estudos literários e ao estabelecimento de textos, encontrará na prática da Estatística Lingüística um rico filão para explorar, revertendo-o em moeda sonante no comércio prático da sua ciência específica.

## II

A Estatística Lingüística encontrou ampla, exata e eficaz aplicação no nível fonêmico da língua. Os inúmeros trabalhos realizados sôbre as mais diversas línguas (das línguas do gru-

po indo-europeu às línguas semitas, chinês e línguas indígenas) revelaram que os fonemas obedecem perfeitamente às leis da probabilidade. O número de fonemas básicos de uma língua, oscilando levemente entre os aproximadamente 50 fonemas fundamentais da linguagem humana, combinam-se segundo leis aleatórias, sendo possível prever as diferentes probabilidades que afetam um determinado fonema em função de uma amostra qualquer do discurso. É claro que tal distribuição dos fonemas pode formular-se fàcilmente em têrmos estatísticos por duas razões essenciais: a primeira é que o número de unidades em ação é relativamente pequeno (o número de fonemas da língua considerada); por conseguinte, os graus de liberdade dêsse sistema são pequenos; além disso, como o fonema é quase independente do significado da mensagem, não intervêm elementos perturbadores das combinações aleatórias e o acaso atua pràticamente soberano. Isso explica a aplicabilidade prática e imediata da Estatística Lingüística na Teoria da Informação. As técnicas aplicáveis às telecomunicações procuram obter o máximo de mensagem através do mínimo de elementos, a fim de fornecer a informação ao menor custo possível.

Como êsse terreno só secundàriamente nos interessa, deixo-o de lado para tratar da aplicação da Estatística Lingüística aos outros níveis da língua: léxico, morfológico e sintático. Aqui começam as grandes dificuldades.

Os primeiros senões fàcilmente apreensíveis são constituídos pelos dois aspectos irredutíveis da realidade lingüística: o elemento qualitativo e o quantitativo. Ninguém negará o lado quantitativo da experiência lingüística. O próprio consenso lingüístico baseia-se em uma média de freqüência do uso geral, aceito pela comunidade falante. Mas... e é aqui que se pode tropeçar: tôda realização do discurso comporta em maior ou menor grau uma escolha por parte do falante, ou do escritor, dos elementos léxicos, morfológicos e sintáticos disponíveis da língua no nível em que êle a atualiza. Daí o título de uma das obras de Herdan — **Language as chance and choice.** E, na verdade, a margem de escolha é bem menor do que o estilista geralmente imagina. Contudo, os graus de liberdade em um sistema lingüístico são muito numerosos; daí a grande dificuldade de aplicação dos parâmetros estatísticos a êsse universo. Trabalhando-se exclusivamente com uma população vocabular em uma língua como a inglêsa onde, parece, os lexemas sobem a 50.000, será extremamente com-

plexo formular a aplicação dos parâmetros estatísticos em um universo com tal variedade. E por essa razão impõe-se a aplicação de métodos quantitativos, uma vez que os dados são muito numerosos. Sendo a lingüística uma ciência de observação como a Psicologia, a Sociologia, a Meteorologia, a análise estatística aí se impõe indubitàvelmente.

Para a Estatística Léxica, em especial, muitos trabalhos já foram realizados e alguns especialistas chegaram a estabelecer certos parâmetros específicos para êsse domínio. Guiraud propôs fórmulas para o cálculo do léxico potencial de um autor e para o cálculo de concentração de um vocabulário. Zipf estabeleceu a lei da

> "constância do produto da freqüência pela ordem (**rank**) ocupada por uma palavra em uma lista de distribuição de freqüências".

Mandelbrot introduziu ligeiras correções na lei de Zipf. Apesar de muito aplicada na literatura do gênero, Herdan vem contestando sistemàticamente a sua validade matemática com argumentos de pêso. Yule propôs outra fórmula, considerando o primeiro e o segundo momento de uma distribuição vocabular ao longo de um texto, de onde deduz uma característica (K), típica das distribuições estatísticas de palavras. Herdan, trabalhando como matemático, parece provar satisfatòriamente que as distribuições das freqüências vocabulares obedecem à lei complexa de Poisson. Vai mais além e formula uma teoria "quantum" da língua aproximando o universo lingüístico do universo físico, adaptando assim a estatística de Bose-Einstein para a lingüística. Procura demonstrar a semelhança do equilíbrio do sistema atômico e do sistema lingüístico. Valha o que valer lingüìsticamente a sua fórmula para o cálculo da entropia com relação aos dados lingüísticos, e outras fórmulas arrevezadas, fica de pé a sua proposta de aplicar esta Estatística Física ao universo da língua.

Alguns lingüistas estatísticos verificaram que os parâmetros da Estatística (a média, o desvio-padrão) não conservavam suas características próprias quando aplicados ao domínio léxico ou morfêmico da língua; isto é, o seu valor não era independente do trabalho da amostra. Por essa razão Yule propõe a sua característica K e Herdan a sua versão do teo-

rema multinominal. Verificou-se também freqüentemente que outros dados perturbavam enormemente os dados quantitativos de uma amostra: a influência exercida pelo tema sôbre o vocabulário utilizado pelo autor, as linguagens especiais (científicas), etc. Assim Muller no seu **Essai de Statistique Lexicale** hesita várias vêzes nas suas conclusões, sem saber a que atribuir uma determinada distribuição de freqüências, se ao tema, se ao gênero literário empregado... É por essa razão também que são criticáveis estudos estilísticos ou lingüísticos que utilizam como ponto de referência uma compilação genérica. Para exemplificar, comparações de um texto francês do século XVII ou de um autor contemporâneo com o "Índice de Freqüências" da lista de Vender Beke terá pouca validade, uma vez que essa lista foi estabelecida para o francês literário do século XIX. Crítica dêsse tipo pode ser feita ao estudo de Ellegard, "Estimating Vocabulary Size" (cfr. bibliografia). Êle opõe aqui os vocabulários de Chaucer, Shakespeare, Bíblia (Authorized Version of the Bible) e de J. Joyce (Ulysses) à lista de freqüências para o inglês estabelecida por Thorndike e Lorge, a partir de um material contemporâneo e extremamente heterogêneo. Como a estilística fundamenta-se na comparação, pouco relevantes serão os seus resultados se utiliza paradoxalmente os seus próprios princípios.

De tudo o que foi dito e do muito que se tem discutido nesse campo, lembremos um dado incontestável nesse oceano de discrepâncias: poderemos apontar, com certeza, tendências no universo lingüístico a que se pode aplicar o cálculo das probabilidades; mas dificilmente estabeleceremos leis que governem êsse universo.

Um último lembrete ainda neste capítulo: dois trabalhos paralelos e complementares são aqui necessários — o manual e o das máquinas. Não só a formulação dos programas a serem executados pelas calculadoras mecanográficas ou eletrônicas exige a presença do homem. Nem mesmo apenas faz-se necessário o seu concurso na utilização do material fornecido pelas calculadoras. É bom não esquecer que o antiqüíssimo trabalho das compilações manuais é de enorme utilidade lingüística e filológica, hoje como outrora. Realmente, apesar do aprimoramento das técnicas, as calculadoras, por mais especializadas que sejam, não substituirão nunca o cérebro do homem e a sua experiência e sensibilidade lingüística na organização dêsses levantamentos estatísticos.

## III

Passemos à última parte — as aplicações da Estatística Lingüística. Tratemos primeiro da Lingüística aplicada. Um domínio que tem utilizado amplamente a lingüística matemática tem sido o ramo da tradução automática. Para citar um um exemplo nesse setor: na Universidade de Nancy (França) trabalha um "Grupo de Tradução Automática" que pesquisa a tradução do inglês para o francês. Ali se utiliza uma calculadora eletrônica. Nessa mesma Universidade ainda um "Centro de Pesquisa para um Tesouro da Língua Francesa" trabalha com essas calculadoras a fim de fazer um levantamento total do vocabulário francês da Idade Média a 1950, num inventário de 250 milhões de palavras, a fim de traçar a história do léxico francês.

Outras pesquisas mecânicas são realizadas em vários centros similares na Europa (Liège, Bruxelas, Gallarate, Estocolmo, Besançon, Paris, Estrasburgo, Sarrebruck), trabalhando com máquinas eletrônicas ou mecanográficas e atendendo a programas diversos. O Centro de Paris (CREDIF) coletou um material imenso da língua francesa falada contemporânea, trabalhou-o mecânica e estatìsticamente e elaborou o "francês fundamental" em vários graus, donde resultou o método de ensino da língua francesa a estrangeiros — "Voix et Images de France". Essa uma das aplicações mais imediatas da Lingüística Matemática: o ensino de línguas estrangeiras através de métodos rápidos e eficientes. O Centro de Bruxelas, financiado pela Euratom, realiza pesquisas nessa linha, tendo em vista o aprendizado das principais línguas da comunidade européia. O Centro de Besançon já estabeleceu o "vocabulário básico" para o alemão e o espanhol, trabalha no "vocabulário científico" francês, e tem no seu programa o estabelecimento dos "vocabulários básicos" das principais línguas européias. O Centro de Besançon vem publicando também sistemàticamente "índices de palavras" de autores franceses da Idade Média aos nossos dias, a fim de fornecer instrumento de trabalho preciso para os estudiosos literários e lingüistas da língua francesa.

A Lingüística Geral já se beneficiou amplamente dessas compilações exaustivas e exatas. No capítulo da genealogia das línguas muita coisa já se fêz com relação ao indo-europeu, procurando-se estabelecer através dêle uma metodologia que possa servir no estudo da correlação de outras línguas despro-

vidas de documentação histórica (línguas indígenas das Américas, línguas africanas, polinésicas). Além dos trabalhos de Kroeber, Chrétien, Czekanowski, Collinder, seria bom lembrar a glotocronologia de Swadesh que utiliza cálculos matemáticos na tentativa de estabelecer a época de separação de duas línguas ou de um grupo de línguas entre si.

A Filologia desde sempre realizou compilações de tipo estatístico nos seus esforços para estabelecer textos, quer quando se tratava de casos de autoria discutida, quer de textos de datação insegura. Hoje ela pode contar com dados mais precisos e com um método de trabalho bem mais seguro para a solução dêsses problemas.

Finalmente, a Estilística e a Literatura utilizarão o material fornecido pela mecanização ou automação do "dépouillement" dos textos de um autor, de uma escola literária ou de um período da língua, a fim de estudar o estilo de um autor ou de uma escola, as características de um período da língua, de um gênero literário, etc.

Para concluir, relacionemos o que foi dito com a realidade brasileira. Parece-me que não padece dúvida o interêsse de promover e fomentar êsse tipo de estudos entre nós. Na Universidade de Toulouse o "Instituto de Português" dirigido pelo Prof. Roche já está procedendo ao "dépouillement" de textos portuguêses e brasileiros. A autora dêste trabalho também está fazendo o mesmo para a obra do poeta português Fernando Pessoa. Contudo, é evidente que se faz necessário tôda uma equipe bem formada para realizar um trabalho de fôlego tendo em vista um estudo mais aprofundado da língua portuguêsa. A Estatística Lingüística aplicada ao português viria colaborar eficientemente no aprimoramento das técnicas da lingüística brasileira. Talvez pudéssemos possuir, dentro de certo tempo, calculadoras que nos ajudassem no levantamento dos dados com menor perda de tempo e menor probabilidade de erros. O "dépouillement" de textos portuguêses e brasileiros forneceria "corpus" ideais para os trabalhos dos nossos lingüístas, filólogos, estilistas e literatos.

## BIBLIOGRAFIA

G. HERDAN — **Language as chance and choice.** Groningen, 1956.

G. HERDAN — **The Calculus of Linguistic Observations.** The Hague,

P. GUIRAUD — **Les caractères statistiques du vocabulaire.** Paris, PUF, 1954.

P. GUIRAUD — **Problèmes et méthodes de la statistique linguistique.** Paris, PUF, 1960.

C. MULLER — **Essai de statistique lexicale.** Paris, Klincksieck, 1964.

**Lexicologie et lexicographie française et romane.** Centre National de Recherche Scientifique, 1961 (Colloque International, Strasbourg, 1957).

**Statistique et analyse linguistique** (Colloque de Strasburg, 1964). Paris, PUF, 1966.

**Artigos**

A. ELLEGARD — "Estimating Vocabulary Size" — **Word**, XVI, 1960, 219-250.

A. ELLEGARD — "Statistical Measurement of Linguistic Relationship" — Language, v. 35, 1959, 131-156.

D. L. BOLINGER — "The Uniqueness of the Word" — **Língua**, v. 12, nº 2, 1963, 113-136.

H. K. COWAN — "A note on statistical methods in comparative linguistics" — **Lingua**, v. 8, nº 3, 1959, 233-246.

S. R. LEVIN — "Deviation — Statistical and Determinate — in Poetic Language" — **Lingua**, v. 12, 1963, 276-290.

H. MITTERAND et J. PETIT — "Index et Concordances dans l'étude des textes litéraires" — **Cahiers de Léxicologie** 3, Didier Larousse, 1962, 160-175.

R. MOREAU — "Au sujet de l'utilisation de la notion de fréquence en linguistique" — **Cahiers de Léxicologie** 3, Didier Larousse, 1962.

C. MULLER — "Le mot, unité de texte et unité de léxique en statistique léxicologique" — **Travaux de Linguistique et de Littérature** I, Strasbourg, Klincksieck, 1963, 155-173.

C. MULLER — "Les index de vocabulaire" — **Bulletin des Jeunes Romanistes** 4, Strasbourg, 1961, 9-14.

C. MULLER — "Les index de vocabulaire, II" — **Bulletin des Jeunes Romanistes** 8, Strasburgo, 1963, 44-45.

## INTERVENÇÕES:

ATICO VILAS BOAS

1) — A título de colaboração, informo a V. Sa. que é Joselice Macedo a pessoa que está pesquisando o português fundamental.

Prof. JOÃO PENHA

2) — Conhece, de Sampaio Dória, o índice de freqüência da colocação dos pronomes? Pôde observar alguns trabalhos estatísticos sôbre OS LUSÍADAS? E sôbre a acentuação de certos ditongos? Essa estatística se pode considerar lingüística? Que se entende aqui por estatística lingüística?

R.) — Não tenho conhecimento dêsse trabalho de Sampaio Dória. Quanto aos trabalhos de tipo estatístico a que fiz referência, realizados desde há muito sôbre obras literárias, só podemos classificá-los de estatísticos de maneira imperfeita. Não sei se é o caso dêsse trabalho. Desde o século passado os filólogos têm se preocupado com a datação de textos, mas freqüentemente não se procedia a levantamentos rigorosos e exaustivos, nem se utilizavam os métodos específicos da estatística que podem ajudar a precisar a procedência de uma obra. É claro que, em última análise, nem mesmo as leis da estatística poderão nos fornecer uma certeza total. Por exemplo, Guiraud, num dos seus trabalhos, procura datar a **Iphigénie**, procedendo de modo rigoroso, para indicar a posição exata da peça no conjunto das obras de Racine. E comenta que os resultados a que chegou eram os mesmos a que haviam chegado críticos literários que trabalharam antes dêle sem utilizar os métodos rigorosos da estatística. Mas, finalmente, ninguém poderia dizer que **Iphigénie** de Racine foi composta no ano X, tendo-se apenas um dado provável da sua composição.

ATICO VILAS BOAS

3) — Ainda como informação relembro os trabalhos de estatística que têm sido publicados na **Revista do Livro.**

Prof. MATTOSO CAMARA

4) — Atendendo a uma indagação que lhe foi formulada por um dos presentes, fêz uma intervenção declarando que na aplicação da Matemática à Lingüística, o que mais o atraj é a formulação algébrica. Pois é a álgebra a verdadeira matemática, uma vez que o número ainda tem qualquer coisa de concreto. A álgebra permite uma abstração muito fecunda no estudo estrutural da língua. E um exemplo muito preciso disso é a Glotocronologia de Swadesh — citado aliás pela relatora. Mas tudo isso não anula a grande utilidade da estatística lingüística desde que executada com os critérios apontados no trabalho da Professôra Maria Teresa Camargo.

STALEY CERQUEIRA

5) — Indago sôbre a possibilidade de delimitação do vocábulo numa obra como **Ulysses**, de James Joyce, em que o problema do léxico é complexíssimo.

R.) — Parece que um professor da Universidade de Cambridge tentou um estudo dêsses, mas não pude consultá-lo. O conhecimento que dêle tive foi obtido por meio do citado artigo de Ellegard. A única referência que tenho sôbre o critério ali adotado é que se distingue "lexical unity" de "word lexical unity".