ALGUNS ASPECTOS DO TRATAMENTO ESTATÍSTICO E COMPUTACIONAL EM LINGÜÍSTICA

Cidmar Teodoro Pais

Apesar da terminologia um tanto rebarbativa para os literatos que tentam abordá-la sem uma preparação matemática, a estatística lingüística — lingüística estatística, como preferem alguns —, desenvolvida em algumas Universidades européias e americanas, como Estrasburgo e Stanford, é uma técnica, ou antes, um conjunto de técnicas bastante simples, que busca aplicar os métodos estatísticos à análise das estruturas e funções lingüísticas.

Não se confunde, pois, com a assim chamada lingüística matemática, que consiste essencialmente numa tentativa de formalização da língua e que se vale dos quadros lógico-matemáticos.

Técnica ou conjunto de técnicas, a estatística lingüística não constitui pròpriamente uma escola lingüística independente mas põe à disposição das diferentes correntes atuais uma série de instrumentos de pesquisa eficientes e versáteis. Empregada em estudos lingüísticos, literários ou filológicos, não suprime o lingüista ou o filólogo e em nada diminui o seu papel.

Com efeito, a estatística lingüística fornece ao lingüista processos de quantificação de dados. É preciso, porém, que o lingüista diga quais são êsses dados, ou seja, quais são as estruturas lingüísticas qualitativas que devem ser quantificadas. A estatística lingüística ensina como contar, mas é preciso saber antes o que contar

Como técnica, a estatística lingüística tem suas limitações e suas vantagens. Fundamentando-se em noções como a freqüência, a distribuição, tomando por referência certa constância dos fenômenos — que permitem definir uma *esperança matemática* —, compete-lhe estudar as relações quantitativas que se estabelecem entre as formas, entre as funções, do código lingüístico. A análise das estruturas qualitativas lingüísticas escapa-lhe totalmente. Seu mérito maior é a objetividade no tratamento dos dados. O grande número dêsses dados, quando analisados, leva a uma compreensão dos fenômenos lingüísticos de alta precisão — até a margem de êrro pode ser calculada — e corrige, inclusive, as conclusões apressadas de uma lingüística por vêzes, impressionista ou intuitiva.

Se fôr empregado um computador, àquela objetividade acrescenta-se a "fiabilidade" e a velocidade da máquina. A capacidade de trabalhar absolutamente sem êrro e a velocidade que permite tratar um número muitíssimo grande de dados em pouco tempo e sem fadiga para o lingüista. O aspecto mais importante, como se vê, é a possibilidade de proceder ao tratamento de fatos lingüísticos em tal quantidade que o seu processamento por fichário manual levaria longos anos e exigiria equipes enormes, o que significa que seria irrealizável na prática. Essa mesma quantidade, ampliando a "amostragem", torna-se mais válida, como manifestação de um código que não pode ser completamente conhecido, e ainda torna o modêlo lingüístico dela inferido mais próximo do mesmo código. Finalmente, se o lingüista fornece ao computador um conjunto de textos e alguns modelos de análise, extraídos do estudo lingüísco qualitativo, a própria máquina fará o levantamento de certos dados e organizará sòzinha, por assim dizer, o seu "fichário"

A pesquisa estatística, em lingüística como nos demais campos a que se aplica, fundamenta-se em dois princípios básicos, a *normalização* e a *amostragem*. Tomemos, por exemplo, o poema épico *Os Timbiras,* de Gonçalves Dias. Se contarmos o número de palavras de um verso determinado, poderemos encontrar um resultado igual a 2, 6 ou 9. Mas, se contarmos o número de palavras de dez versos, dificilmente encontraremos oito versos com nove palavras ou com duas. Se repetirmos a operação em cem versos, veremos que a imensa maioria tem entre 5, 6 e 7 palavras, estabelecendo-se uma média de 6,5 palavras por verso. Continuando a operação por mais cem versos, verificamos que essa média se confirma (1) Um verso isolado pode, pois, apresentar um número excepcional de palavras, mas se tomarmos um número suficientemente grande de versos, o resultado se inscreverá dentro da média.

Assim, de um modo geral, um fato lingüístico isolado pode apresentar características excepcionais e o lingüista não tem condições para saber " a priori" se êle é realmente um caso particular ou representa a norma. Tomando-se um número suficientemente grande de fatos lingüísticos, as diferenças individuais tendem a compensar-se, anulam-se e chega-se a uma distribuição dos fatos em tôrno da média, em forma de uma curva da lei *normal* ou curva de Gauss. Em seguida, toma-se uma outra série de fatos do mesmo tipo e verifica-se se a curva se mantém. Atingido êsse ponto, em que o fenômeno estudado apresenta uma curva de Gauss constante, é inú-

---

(1) — Pesquisa realizada por um grupo de Professôres brasileiros, entre os quais estava o autor dêste artigo, sob a direção do Professor Ch. Muller, da Universidade de Estrasburgo, em 1969.

til aumentar o número de fatos, prolongar a pesquisa, o resultado será também constante.

Obviamente, não há uma curva única para o código lingüístico. Ao contrário, cada aspecto, cada elemento do código, cada estrutura e cada função têm a sua curva própria, têm um comportamento próprio, e o número de fatos lingüísticos concretos que se deve considerar para alcançar o nível de *normalização* varia de acôrdo com a natureza do fenômeno lingüístico estudado. Assim, no estudo do sistema fonológico, que apresenta um inventário fechado e extremamente restrito de elementos (24 no espanhol, 46 no sânscrito, por exemplo), a curva que representa a distribuição *normal* pode ser obtida com o exame de uma seqüência textual que contenha apenas dois mil fonemas. Se se tratar do léxico, cujo inventário é aberto e muitíssimo maior, seremos obrigados a lançar mão de textos bem mais extensos.

Graças ao princípio da normalização, torna-se possível aplicar o segundo, o da amostragem, torna-se possível, enfim, a própria pesquisa lingüística em têrmos estatísticos. Com efeito, ainda que pudéssemos gravar todos os atos de fala de todos os brasileiros durante um ano — hipótese puramente teórica, como se vê —, terminada a pesquisa, os brasileiros continuariam falando; os lingüístas nunca teriam em suas mãos todos os atos de fala que são as manifestações do código. E o conjunto de atos de fala, se pode dar-nos a *norma,* jamais esgota as possibildades do *sistema* (2) Impõe-se, por conseguinte, que se proceda por amostragem, ou seja, que o lingüista realize suas pesquisas sôbre uma *amostra* considerada válida.

Considera-se uma amostra válida, para um fenômeno determinado, aquela suficientemente ampla para que os fatos lingüísticos apresentem uma curva de Gauss que seja representativa da *norma,* que atinja um nível de constância verificável noutras amostras. Vê-se, pois, que uma amostra nunca é válida em têrmos absolutos mas apenas em relação a um aspecto determinado das estruturas lingüísticas, que se vai estudar. Como observamos, um texto que tenha uma extensão de dois mil fonemas dá conta da distribuição e da combinatória dêsses fonemas, com uma margem de êrro negligenciável. A mesma amostra, entretanto, não permitirá um estudo seguro dos processos sintáticos, por exemplo.

Assim, pois, a primeira preocupação de um lingüista que se dedica a uma pesquisa dêsse tipo, é o estabelecimento de um *corpus*

---

(2) — V "Sistema, norma y habla" nos estudos de E. Coseriu, *Teoria del Lenguaje y Lingüistica General,* 2. ed., Madrid, Gredos, 1969, p. 11-113.

de trabalho. O *corpus* é um conjunto de textos, ou seja, de *atos de fala* que constituirão o objeto da pesquisa, que serão submetidos ao tratamento estatístico, computacional.

Quando se estuda um autor que tenha produzido obra pouco extensa, é possível fixar como *corpus* o texto integral. O *corpus* integral ou exaustivo, teòricamente melhor — por sua representatividade absoluta —, é, na realidade, um instrumento de trabalho incômodo, quando o autor estudado tem uma vasta obra — e acarreta esforços inúteis, em buscas que reafirmam uma normalização já atingida —; é òbviamente inalcançável, quando se trata de pesquisa lingüística e não estilística. Opera-se, então, com um *corpus* de amostragem.

Contudo, para que um *corpus* dêsse tipo seja considerado uma amostra válida, em relação a determinados fenômenos lingüísticos que serão objeto de análise, é necessário que preencha algumas condições que assegurem a sua *representatividade,* ou seja, a representatividade dos atos de fala nêle contidos quanto à norma — geral ou especial — que se procura estabelecer.

Se se tratar de um *corpus* de língua oral, torna-se imprescindível selecionar os informantes, os indivíduos que fornecerão os atos de fala, de acôrdo com critérios sócio-econômico-culturais — faixa etária, nível de escolaridade, nível cultural, profissão, região de origem, etc. — de modo que representem com a proporcionalidade desejada a real situação do grupo lingüístico submetido a exame. Por outro lado, as técnicas de recolho dos atos de fala que integrarão o *corpus,* da gravação enfim, devem ser tais que perturbem o menos possível a espontaneidade dêsses atos.

Se se tratar de um *corpus* de língua escrita, a representatividade será assegurada por um levantamento dos textos realizado mediante a aplicação de técnicas *aleatórias* rigorosas, (3) que permitam proceder à escolha dos trechos, sem interferência de fatôres intrínsecos, independetemente de critérios estilísticos, de ordem temática, gôsto pessoal, sem interferência apriorística dos fenômenos que justamente se pretende estudar, dando aos textos, num levantamento assim feito, uma distribuição e uma extensão no interior do *corpus* compatíveis com a sua posição e significação quantitativa na produção literária, de modo geral, ou num campo específico tomado como norma especial.

---

(3) — Referimo-nos às técnicas empregadas em estatística, para a obtenção de amostras, rigorosamente por *acaso* — no sentido matemático, evidentemente — de maneira a reduzir ao mínimo a interferência de elementos intrínsecos e até mesmo de fatôres materiais como a encardenação ou os caracteres tipográficos de partes do *corpus*. Existem inclusive tábuas de números distribuídos ao acaso, estabelecidas por computador.

Estabelecido o *corpus* de trabalho, começa a pesquisa pròpriamente dita, a busca dos dados. Claro está que só se podem pesquisar dados de certa natureza, como lembramos logo de início, ou seja, dados quantificáveis, que nos levem ao estabelecimento de relações quantitativas entre elementos do código lingüístico, que nos levem enfim a estruturas quantitativas (4) Muitas vêzes, é óbvio, o conhecimento de determinadas estruturas quantitativas nos conduz a uma melhor compreensão das estruturas qualitativas. Contudo, em estatística lingüística, partimos destas em busca daquelas, partimos de *modelos* de análise fornecidos por uma escola lingüística e percorremos o *corpus,* levantando as *ocorrências,* as *atualizações* daqueles modelos.

Procede-se, pois, ao levantamento dos fatos concretos relativos ao fenômeno lingüístico, ao aspecto estrutural, funcional ou estrutural-funcional que se deseja pesquisar Considera-se, como hipótese de trabalho, que o fenômeno estudado apresenta uma distribuição aleatória, que as variações se inscrevem numa curva da lei *normal* ou de Gauss. Prossegue-se nessa busca dos dados até que se obtenha uma curva constante. Essa constância será verificada por contra-prova, isto é, pelo exame de amostra suplementar.

Tomemos o exemplo simples, do número de palavras por verso, levantado em *Os Timbiras,* na pesquisa citada mais acima. O número de palavras por verso *varia* de um mínimo de duas palavras a um máximo de nove. Temos, pois, que o número de palavras por verso é uma *variável, x,* que pode assumir diversos valores. Em 2 e 9 temos os limites mínimo e máximo que definem o *campo da variável*. Numa amostra de 100 versos, encontramos para cada valor da variável, um certo número de versos que lhe corresponde, a que chamamos os *efetivos,* n_i, de cada valor da variável (5)

Obtida uma curva da lei normal, constante, conhecido o campo da variável e os efetivos correspondentes a cada valor da variável, estamos em condições de examinar verdadeiramente o fenômeno, do ângulo quantitativo, e procurar definir-lhe uma *norma*.

Nesse estágio da pesquisa, passa-se a trabalhar com dois *parâmetros*. Obtém-se o primeiro dêles, multiplicando-se os valores da variável *x* pelos efetivos que lhes correspondem e dividindo-se a

---

(4) — Como, por exemplo, o número de palavras por verso, a proporção de substantivos em relação às demais classes de vocábulos, a distribuição e a freqüência de determinada estrutura sintática, etc..

(5) — Numa das amostras de 100 versos (foram levantadas várias delas) tinhamos a relação variável/efetivos: 2/1; 3/2; 4/8; 5/17; 6/35; 7/27; 8/7/; 9/1.

somatória dêsses produtos pela somatória dos efetivos. Resulta dêsse cálculo a média — no caso, a média de versos por palavras, 6,5 —, que é o *parâmetro de posição*.

O parâmetro de posição assinala o epicentro do fenômeno ma. o seu conhecimento, apenas, não é suficiente para determinar como os efetivos correspondentes a cada valor da variável se dispersam em tôrno da média. Lança-se mão, pois, de um segundo parâmetro.

Se tomarmos cada verso, dos 100 considerados na amostra, veremos que o seu número de palavras não coincide com a média, mas apresenta uma diferença para mais ou menos, um *desvio* real, positivo ou negativo. Como estamos diante de um fenômeno que apresenta uma distribuição *normal,* sabemos, pelo princípio da *normalização,* que os produtos de cada desvio real pelo número de versos que o apresentam, somados, são iguais a zero Por meio de um artifício matemático, calculando-se não o produto do desvio real pelos efetivos corrspondentes mas o quadrado dêsse produto (6), pode-se obter a *variança* do fenômeno e a raiz quadrada da variança nos dá o desvio-padrão.

O *desvio-padrão* é, por conseguinte, o *parâmetro de dispersão* que nos indica como os dados reais se distribuem em tôrno da média, para um determinado fenômeno.

O desvio-padrão é um instrumento de trabalho precioso, que nos permite apreciar um desvio real, julgá-lo e definir-lhe a relevância, a significação — ou a ausência de significação diante de uma norma.

Com efeito, dada uma curva de Gauss, que descreve determinado fenômeno, estabelece-se, em estatística, que os fatos que apresentam um desvio real igual ou menor ao desvio padrão, para mais ou para menos, constituem 68% dos efetivos. Daí decorre a noção de *intervalo* Com um intervalo de um desvio padrão a menos até um desvio-padrão a mais, tomando-se por referência a média, temos uma faixa que compreende 68% das ocorrências. Com um intervalo de dois desvios-padrão a menos até dois desvios-padrão a mais, sempre tomando-se por referência a média, temos uma faixa que compreende 95% das ocorrências.

Considera-se, dêsse modo, que todo o desvio real compreendido nesse último intervalo pode ocorrer por *acaso,* que há 95

---

(6) — Já que o quadrado de um número negativo é positivo. Trata-se de um recurso que permite contornar o problema da normalização de um fenômeno para melhor poder examiná-lo em seguida.

possibilidades em 100 que assim o seja, considera-se, em suma, que se trata de um desvio *normal,* que se poderia esperar, em face da *norma* estabelecida.

Um desvio real que ultrapasse êsses limites tem, portanto, apenas 5 possibilidades em 100 de ocorrer por acaso. Como noutras ciências, em lingüística toma-se êsse critério, o de 0,05, para apreciar a significação de um desvio. Todo desvio real que tem menos de cinco possibilidades em 100 de ocorrer por acaso, é considerado, estatìsticamente, um *desvio significativo*.

Na pesquisa realizada sôbre *Os Timbiras,* efetivamente, chegou-se a um desvio-padrão de 1,3. Pudemos veri icar, então, que os versos formados de 5, 6 ou 7 palavras constituiam realmente 68% dos efetivos, que 95% dos versos continham de 4 a 8 palavras e, finalmente, que os raros versos compostos de 2, 3 ou 9 palavras apresentavam todos, invariàvelmente, um problema de natureza temática ou estilística.

Convém lembrar que o desvio estatìsticamente significativo não implica absolutamente numa apreciação de valor, no plano estético, nem é garantia de que o fato seja excepcional, no plano lingüístico. Êle permite tão sòmente separar para uma análise posterior, em têrmos qualitativos, aquêles fatos que têm maiores possibilidades de apresentar interêsse lingüístico ou estilístico.

Na pesquisa estatística e computacional, temos, bàsicamente duas "démarches" possíveis: do todo para a parte, da parte para o todo.

Conhecida uma parte, uma amostra, um *corpus* constituído de certo número de *atos de fala,* considerado como suficientemente representativo, de acôrdo com os critérios expostos acima, faz-se uma estimativa sôbre o conjunto da língua. Esta, como não pode ser examinada diretamente, é *conhecida* através de suas manifestações. Trata-se então de um estudo lingüístico.

Ao contrário, quando se conhece o todo, obras completas de um autor ou mesmo da lteratura de uma nação, pode-se analisar, à luz dêsse conjunto, uma obra particular ou até um trecho, examinando certos aspectos de suas estruturas quantitativas. Temos, nesse caso, um estudo estilístico.

Òbviamente, nem sempre é possível distinguir claramente a pesquisa lingüística da estilística, sobretudo quando se analisam fenômenos situados na faixa fronteira entre as duas disciplinas, como, por exemplo, aquêles ligados à linguagem afetiva. Assim, as duas "démarches" não são incompatíveis, antes são, complementares e o lingüista passa constantemente de uma para outra.

Num caso como noutro, são essenciais as noões de norma e desvio. Nesse terreno, têm sido comuns os mal-entendidos e as

imprecisões. Preferimos, por isso, ater-nos à sua conceituação no plano matemático, em que nos colocamos até aqui.

Como pudemos observar mais acima, o exame quantitativo de um fenômeno lingüístico qualquer, o levantamento das ocorrências de uma estrutura ou de uma função lingüística, a análise das relações quantitativas existentes entre certos elementos do código, vai determinar uma curva de Gauss, que representa a norma daquêle fenômeno. Essa curva implica numa média e num parâmetro de dispersão, o desvio-padrão.

Ora, êsses dados permitem definir, de um lado, uma *esperança matemática,* ou seja, calcular quais são as possibilidades de que o fenômeno apresente determinado comportamento em qualquer amostra examinada posteriormente, em um nôvo conjunto de *atos de fala* — de língua falada ou de obra literária —; permitem esperar — com 95 possibilidades em 100 — que a sua análise dê resultados semelhantes; e permitem, quando isso não se verifica, identificar os desvios realmente significativos e separar aquêles trechos, que, estatìsticamente, não pertencem ou não parecem pertencer ao mesmo esquema de urna, ou seja, lingüìsticamente, não representam a mesma norma.

Vê-se, pois, que a noção de norma não é absoluta mas relativa. Em primeiro lugar, o estudo de cada fenômeno dá-nos apenas a norma dêsse fenômeno. A norma lingüística, considerada como o conjunto do léxico efetivo e da sintaxe efetiva, de freqüência e distribuição regulares, resulta do conjunto das normas ligadas a cada aspecto das estruturas e funções lingüísticas, dentro do sistema. A norma não se definirá, dêsse modo, por uma curva, mas por um conjunto extremamente amplo e diversificado de curvas da lei normal, de grande complexidade.

Entretanto, mesmo uma norma assim definida é algo relativo. Com efeito, se gravarmos os atos de fala de 10.000 informantes, durante um tempo razoável, poderemos dispor de uma amostra válida, talvez, para a descrição do português falado no Brasil, desde que respeitadas as condições que garantam a representatividade do *corpus* Estaremos em condições, então, de fazer uma estimativa sôbre a *norma falada*.

O conjunto de textos literários produzidos na mesma época apresentará, inevitàvelmente, em relação àquela norma falada, um grande número de desvios, muitos dos quais estatìsticamente significativos. Poderemos então surpreender muitos dos aspectos em que aquêles textos divergem da norma falada. Alguns dêsses elementos poderão ser úteis para caracterizar os primeiros em oposição à última.

Contudo, tomados êsses textos literários em seu conjunto, e analisados no que se refere aos mesmos fenômenos examinados quando do estudo da norma falada, verifica-se que êles apresentam, no interior dêsse *corpus,* uma distribuição em curva de Gauss, constante, com parâmetros de posição e de dispersão próprios. Temos, então, que, embora apresentando desvios significativos em relação à norma falada, êles constituem, internamente, uma norma especial, a *norma literária*. Esta, por sua vez, dará elementos para o exame de um autor, em particular.

Cada autor, em relação à norma literária, apresentará em suas obras certo número de desvios significativos, que permitirão compará-lo não apenas a essa norma como também aos demais autores. Entretanto, o *corpus* constituído pelo conjunto de suas obras, mostrará, em relação aos mesmos fenômenos, curvas constantes, que definiem uma norma, *a norma de autor* Ela dará elementos para um melhor exame, um melhor conhecimento, de uma de suas obras ou mesmo de um pequeno trecho.

Cumpre fazer aqui uma observação. Uma pequena frase, mal interpretada, "o estilo é desvio" despertou esperanças descabidas e acabou levantando uma interminável polêmica. Tomar essa frase literalmente equivale a dizer "o homem é uma raposa" e verificar, depois, desencantado, que êle não tem nariz pontudo ou cauda peluda. Evidentemente, a análise dos desvios — e não "do desvio" — não permite caracterizar um estilo e nem mesmo, quando empregado isoladamente, determinar se um texto pertence ou não a certo autor. Nem a existência de tal ou qual desvio nos autoriza a um julgamento estético, ao menos no estágio atual das pesquisas. O que não se pode negar, entretanto, é que as normas especiais a que nos referimos, as relações entre elas, e o estilo de um autor apresentam, efetivamente, desvios estatìsticamente significativos, que merecem um exame objetivo, desapaixonado. Os desvios, se não explicam o estilo, podem ser, entretanto, uma de suas manifestações.

A pesquisa estatística e computacional, em lingüística, desenvolveu-se extraordinàriamente nas duas últimas décadas. Êsse desenvolvimento, entretanto, não foi uniforme. No campo da fonologia, por exemplo, basta lembrar os problemas de uma dupla transcrição, a fonológica e a dos caracteres disponíveis num computador que obriga a uma trabalhosa pré-edição e a uma "tradução" posterior dos resultados, além dos cuidados necessários para não confundir combinatória de fonemas e norma articulatória fonética, perigo sempre presente. Ainda assim, os resultados são já animadores. Outro terreno escorregadio é o da sintaxe, não apenas em razão das dificuldades de programação mas sobretudo em virtude do atraso relativo em que se encontra essa disciplina.

Assim, o campo em que mais se desenvolveu a lingüística computacional e aquêle em que apresentou resultados amplos e seguros, é o da lexicologia. O emprêgo do computador no estudo do léxico tem permitido estudos sôbre *corpus* de um milhão de palavras, por conseguinte, altamente representativo; tais estudos seriam impraticáveis sem a máquina, ou exigiriam o trabalho esterilizante de equipes numerosas por muitos anos. Dêles muito se aprendeu a respeito das estruturas quantitativas do léxico. Além do desenvolvimento da pesquisa lingüística fundamental, já começa a alongar-se a lista de aplicações práticas, na descrição ou no ensino de línguas, em informática, etc.

O ponto de partida de uma pesquisa estatística do léxico é o estabelecimento de uma *norma léxica,* uma série de regras que possibilitem determinar com segurança os limites do *vocábulo* Como não existe uma norma léxica universal nem tão pouco se conseguiu uma que se aplique de modo exautivo e sem possibilidade de êrro a uma só das línguas estudadas até hoje, o pesquisador é levado a estabelecer uma norma léxica própria, isto é, fixar os critérios segundo os quais se determinará a fronteira entre um vocábulo e outro, se poderá separar os vocábulos, quantificá-los, sem depender do traiçoeiro espaço em branco e sem desmantelar indevidamente as locuções, as léxicas compostas e complexas.

Evidentemente, uma norma léxica criteriosa é a condição *sine qua non* de qualquer trabalho estatístico e computacional, nesse campo. Contudo, visto que deve possibilitar a separação de tôdas as palavras que compõem um texto, sem exceções e sem êrros, semelhante norma léxica torna-se bastante complexa. Daí decorre a dificuldade — se não a impossibilidade — de transformá-la em programa para um computador, de modo que êste possa sòzinho aplicar os modêlos.

Contorna-se o problema com uma pré-edição, assinalando-se manualmente a separação das palavras. Assim, o *corpus* submetido ao tratamento computacional, perfurado em cartões ou fitas, já traz um sinal — delimitação de palavra — fàcilmente identificável pela máquina. Naturalmente, a separação é automática, quando se considera como palavra a unidade gráfica, separada por espaços em branco.

A norma léxica tem ainda outra função: a de indicar se os resultados obtidos na pesquisa são comparáveis aos de outra pesquisa, ou, se as normas empregadas divergem, até que ponto podem ser confrontados.

O Professor Ch. Muller distingue: *lexema,* a unidade de língua, o signo lingüístico disponível na consciência; *vocábulo,* o signo lin-

güístico atualizado em discurso, unidade de discurso: *palavra,* cada ocorrência, no texto, de um vocábulo, ou seja, unidade de texto (7).

Se tomarmos um *corpus* constituído de *x* palavras, veremos que elas não são tôdas diferentes, mas que correspondem a *y* vocábulos e que *y* é sempre menor que *x;* ou seja, que muitos vocábulos se repetem, *ocorrem* várias vêzes, que o número de vocábulos e menor que o de palavras. Dividindo-se o primeiro pelo segundo, V/P = x/y, obtemos uma fração que indica a riqueza do vocabulário daquele *corpus,* em têrmos absolutos. Tomando-se vários textos e calculando-se o quociente vocábulos/palavras de cada um dêles, pode-se classificá-los uns em relação aos outros, quanto a êsse aspecto.

A classificação dos vocábulos pela freqüência tem grande interêsse e diversas aplicações práticas. Tomando-se um texto, ou um conjunto de textos — um *corpus* —, encontramos vocábulos que nêle ocorrem uma única vez, outros que ocorrem duas vêzes, três, etc., até aquêles que atingem freqüências muito altas: V_1, V_2, V_3, V_n. Pode-se classificá-los, então, de acôrdo com a sua freqüência, estabelecendo uma hierarquia entre êles.

Os vocábulos que apresentam as freqüências mais altas e distribuição regular pelo maior número de textos (8), são aquêles que interessam à elaboração de um vocabulário fundamental. A alta freqüência dêsses vocábulos tem como conseqüência seu baixo custo de armazenamento na memória — já que êles nos são "lembrados" constantemente nas emissões dos que nos falam — e alto rendimento — já que os empregamos a todo o momento.

A noção de vocabulário fudamental — à qual se deveria ligar também a de uma gramática fundamental, baseada nas estruturas sintáticas mais importantes — não é rígida, como pensam alguns mas bastante flexível. Há um vocabulário fundamental realmente mínimo, estabelecido a partir de poucos vocábulos caracterizados pelas mais altas freqüências, e que poderíamos chamar de "vocabulário de sobrevivência", pois que é o mínimo que permite a um indivíduo em terra estranha não morrer de fome. Um segundo nível compreende vocábulos de freqüência ainda bastante alta, *de* um custo de armanezamento um pouco maior, mas que permitam uma conversação normal.

---

(7) — Assim traduzimos *lexème, vocable* e *mot,* propostos pelo Professor Ch. Muller. Cf. MULLER, *Initiation à la statistique linguistique,* Paris, Larousse, 1968, p. 136.

(8) — Referimo-nos aqui indiferentemente aos *textos* que constituem um *corpus,* tenham êles origem numa gravação de língua falada ou resultem do levantamento de trechos de língua escrita.

Não se deve esquecer aqui o vocabulário fundamental literário, obtido pelo levantamento dos vocábulos que apresentam alta freqüência e distribuição regular, encontrando-se na interseção dos conjuntos de vocabulário de um bom número de textos literários.

A organização de vocabulários especiais obedece aos mesmos princípios. Temos, por exemplo, o vocabulário fundamental técnico do francês, que está sendo elaborado pelo Centro de Informática de St. Cloud.

Os vocábulos de baixa freqüência — que ocorrem uma ou duas vêzes no texto, V_1 e V_2 — são naturalmente os mais numerosos e, embora de baixo rendimento, de acôrdo com os critérios expostos acima, têm particular interêsse. Os vocábulos de alta freqüência, como certas preposições, o artigo, certos verbos importantes, certos substantivos aparecem nas emissões de todo falante, pràticamente, e com uma distribuição pouco variável. Já os vocábulos de baixa freqüência constituem um conjunto bastante grande, uma lista extensa, que jamas coincide de um emissor para outro, de um autor para outro. Por isso, o vocabulário de baixa freqüência contribui para a caracterização de um autor ou de uma obra e constitui, por vêzes, um elemento valioso para sua identificação ou para seu reconhecimento pelo leitor.

Não nos seria possível relembrar oqui todos os trabalhos que tem sido realizados no campo da lexicologia computacional, nem mesmo citar o longo inventário de suas aplicações práticas. Faltanos o tempo, é limitado o espaço de que dispomos. Restringimo-nos, pois, a algumas de suas aplicações que têm despertado maior interêsse. Muito pouco se sabe sôbre as estruturas quantitativas do léxico, o campo é pràticamente virgem e as perspectivas de trabalho amplas.

Como dizíamos, logo de início, a estatística lingüística é, antes de tudo, uma técnica e como tal deve ser considerada. O computador, uma máquina que permite pesquisar um *corpus* extensíssimo, de maneira exaustiva, sem êrro e sem fadiga para o pesquisador Mas não se deve pedir da lingüística estatística e computacional mais do que ela pode dar.

Cabe ao lingüista a escolha das estruturas qualitativas que serão objeto de estudo, a determinação do *corpus* que será submetido a tratamento e a conveniente proposição das questões. Se falhar num dêsses aspectos de seu trabalho, os resultados serão comprometidos.

Mas não termina aí a responsabilidade do pesquisador Uma vez obtidos os dados estatísticos, inicia-se o trabalho realmente árduo e criador que é o da sua interpretação e que leva a conclusões de valor científico. Dois perigos o espreitam nessa fase: restringir ex-

cessivamente as suas conclusões, num subaproveitamento dos dados de que dispõe, ou lançar-se à aventura de afirmações que êstes de nenhuma maneira autorizam.

Contudo, entre a posição retrógrada daqueles que negam qualquer significação à estatística lingüística, e daqueles que pretendem tudo explicar, solucionar todos os problemas dêsse ponto de vista, existe uma terceira, a científica, que consiste em encarar a lingüística estatística e computacional como um instrumento válido, limitado como qualquer outro, e que, quando empregado com seriedade e objetividade, com espírito científico, enfim, trouxe importante contribuição para o desenvolvimento dos estudos lingüísticos e nos oferece hoje um vasto campo para a pesquisa.