SEMIÓTICA, INFORMAÇÃO E LINGUAGENS NATURAIS

Jorge de Albuquerque Vieira
(Pontifícia Universidade Católica de São Paulo)
jorgeavi451@hotmail.com

ABSTRACT: *This work apply concepts of Information Theory as utilized in Mathematical Linguistics as a method for deal with general texts, related to the spoken and writen human activity as those obtained from fisiological indexes associated to this activity.*

KEYWORDS: *Information; grammar; entropy; redundancy.*

1. Introdução

Nosso trabalho pretende trabalhar a base conceitual e ontológica da aplicação da Semiótica na análise de sinais provindos da atividade científica, através da construção de signos de alta complexidade e explorando as noções de *sistema organizado* e *sistema de informação*. Em particular, é nosso interesse trabalhar os aspectos de organização de sistemas naturais, aspectos estes não muito discerníveis pelas técnicas usuais de tratamento de dados e sinais científicos.

Tem sido notada, na literatura, a crescente necessidade de desenvolvimento de técnicas que consigam captar a complexidade de sistemas naturais. Do ponto de vista metodológico, isso significa a necessidade de ir além dos temas da ordem, regularidade e simetria e atingir aqueles de complexidade, organização e evolução; do ponto de vista semiótico, isto implica o dimensionamento de signos (fortemente indiciais) que consigam captar estes últimos aspectos, já que os índices convencionais não o conseguem.

Acreditamos assim que a citada necessidade repousa na crescente imposição de parâmetros de complexidade (principalmente integralidade e organização, ver por exemplo, Denbigh, 1975) sobre aqueles típicos no uso convencional (ordem e periodicidade, por exemplo Bloonfield, 1976), na busca de um maior entendimento nos sistemas estudados pelas várias ciências.

A base ontológica que é adotada em nosso trabalho é aquela sugerida por Mario Bunge (1977), ou seja, a Teoria Geral de Sistemas construída como uma Ontologia Científica. Ao longo do texto estaremos utilizando os conceitos de sistema e parâmetros sistêmicos, segundo esse autor e outros, como Kenneth Denbigh e Avanir Uyemov (1975), Edgar Morin (1986), etc. Nesse contexto, a *complexidade* será considerada um parâmetro livre e não a integralidade, como proposto por Denbigh (1981). O enfoque semiótico segue a proposta de Anderson et al. (1984), por ser coerente com a Ontologia de Mario Bunge em sua visão sistêmica (Bunge, 1979).

2. Séries Temporais, Sinais e Sistemas

Chama-se *Série Temporal* uma série de medidas considerada em sua sucessão natural de obtenção no tempo, sendo muitas vezes as medidas tomadas a intervalos regulares de tempo (Bloonfield, 1976: 1). Uma série temporal, codificada ou não, é um sistema. Pela definição devida à Uyemov (1975: 96) temos que um sistema é um agregado ou conjunto de coisas ou elementos tão relacionados que conseguem a partilha de propriedades, que se tornam comuns à esses elementos:

$$(m)S =_{df} [R(m)]P$$

Na expressão, o agregado (m) é equivalente a um alfabeto $A = \{x_i\}$. O sistema é unidimensional em sua representação, na forma de uma cadeia de signos. Tais signos são conectados por um conjunto R de relações, o que é admitido *a priori* para a série (a análise clássica de séries temporais admite que os termos das mesmas são dependentes entre si). Do ponto de vista de um processo estocástico, o conjunto R equivale a uma distribuição de probabilidades condicionais da forma

$$p(x_i / x_a x_b x_c \ldots)$$

onde, segundo a Teoria da Informação, a sequência $x_a x_b x_c \ldots$ indica a *faixa de influências intersimbólicas* ou seja, o efeito que signos anteriores têm sobre o signo "presente" x_i (Goldman, 1968:17). Tal efeito indiretamente exprime o "vigor" da gramática envolvida, a qual em Teoria da Informação é expressa por p, para os signos individuais e também para os arranjos por eles formados.

Na modelagem clássica de séries temporais, se temos um processo AR (*Auto Regressive*), este é mais gramatical que um MA (*Moving Average*). Do ponto de vista sistêmico, o primeiro apresenta o conjunto R de relações mais ativo, atuante, ou seja, uma coesão maior do que aquela para MA. O domínio da faixa de influências intersimbólicas é o domínio da coesão, no caso uma forma de conectividade. Pelo critério de Denbigh, (1975: 87), podemos ter R com relações a favor, indiferentes ou contra formas de coesão e funções. No nosso caso, vemos que só lidamos com relações ou conexões do primeiro tipo, embora a idéia de função como uma propriedade característica de um subsistema da cadeia sígnica (como um "pico", por exemplo) seja uma propriedade partilhada por um conjunto de estados da série que não surgirá em outros subsistemas; nesse caso, formas de inibição deverão surgir associadas aos parâmetros que fazem o ajuste linear nos modelos clássicos, de modo que certos estados "permitam" o "pico" e outros não. Mas quanto à série mesma (e não o seu modelo) podemos dizer que todos os seus estados são coesos em algum nível.

Se a série e seu texto codificado possuem N termos, o número de relações, termo a termo, será N-1. Do ponto de vista da estocasticidade, os signos considerados isolados terão probabilidades de ocorrência individuais. Isso ainda não caracteriza relação. Mas a partir dos arranjos dos signos tomados 2 a 2 (a redundância de 2a. ordem), a idéia de relação faz sentido. A construção de ordens de redundância leva ao estabelecimento de relações mais complexas e pro-

fundas, envolvendo grupos ou subsistemas de signos que aliás, é o que a modelagem clássica ARIMA (a fusão dos processos AR e MA e mais ainda uma possível tendência, denotada por I) propõe, linearmente.

Se o par ordenado G = < A,R > exprime gramática, ele equivale à R(m), na definição de Uyemov. Assim,

$$(m)S =_{df} [\ G\ ]\ P$$

onde P significa "propriedades partilhadas". Os vínculos R levam a subsistemas de signos que têm uma P característica deles, enquanto entidade coletivas. Nas linguagens humanas naturais, essas "ilhas" conectas são o suporte da semântica. Da mesma maneira, nos sinais factuais elas representam características de mensagens emitidas pelas fontes. A aplicação de R e a consequente geração de P, com o surgimento dos subsistemas, indica a integralidade. Seja M o conjunto de todas as mensagens ou subsistemas gerados pela gramática em sua integralidade. O texto total é assim formado por M. Temos agora o par ordenado

$$< G , M > = L$$

onde L é uma linguagem. Na definição,

$$(m)S =_{df} L$$

ou seja, o sistema de signos é uma linguagem (uma gramática associada à propriedades coletivas gera mensagens com integralidade, tal que [G]P é equivalente à <G,M>).

3. Gramática e Integralidade

Adotamos a integralidade como índice de gramaticalidade. O vigor de uma gramática é expresso quantitativamente por

$$R = 1 - [\ S_r / S_{max}\ ]$$

que é a redundância. O termo S_r / S_{max} é a entropia relativa, ou o quanto a entropia real afasta-se da condição de entropia máxima. Do ponto de vista da conectividade ou R, temos para os sistemas uma outra relação, c_r / c_{max} , ou seja, a razão entre o número de conexões que ocorrem realmente e aquelas em número máximo, que seriam possíveis se não houvessem as restrições gramaticais. Essa relação é chamada em Biologia, de *conectância*. Sistemas estáveis apresentam um valor "mediano" de conectância, assim como linguagens naturais apresentam um valor "mediano" de entropia. Conectância está associada à regra gramatical, à estrutura, ao número de relações. Já a entropia fala indiretamente de heterogeneidade imposta pela gramática, fala de integralidade.

A diferença básica entre c_r e S_r é que o primeiro lida com número de relações (estrutura) e o segundo com graus de liberdade "pesados" pela distribuição de probabilidades p. O primeiro é um número de relações, o segundo é uma

característica coletiva e de comparação interna entre signos relacionados. O valor de c_r não esclarece sobre o vigor gramatical nem sobre a integralidade no sistema. Podemos ter dois conjuntos conectados pelo mesmo número de conexões, mas com graus de integralidade diferentes. Ou seja, um termo indica o *quanto* o sistema é conecto, o outro *como*.

Suponhamos um alfabeto fixo: nesse caso, R só varia com S. E

$$S = -\sum_{i=1}^{n} p_i \log p_i$$

onde S (a entropia; não confundir com "sistema", para o que utilizamos a mesma notação) cresce na medida em que a distribuição dos p_i tende à homogeneidade e os arranjos sígnicos tornam-se mais e mais independentes; S cai quando surge uma heterogeneidade na distribuição, favorecendo como mais prováveis (ou complementarmente, como mais raros) certos arranjos de signos.

Como S e R são propriedades de *ensemble*, dois ou mais sinais podem apresentar uma mesma evolução de redundância, devida, para cada um, a signos ou arranjos diferentes. Assim, *toda a análise de R deve ser acompanhada da identificação dos signos que agem sobre R*. A natureza dos textos muda e os signos podem alterar seus papéis de acordo com as circunstâncias. São esses papeis que demarcam o domínio da semântica.

Em processos estocásticos ruidosos, observamos uma queda na redundância com o aumento do comprimento do subsistema de signos (ou com a ordem de redundância). Quanto maior a ordem, maior o tamanho da "palavra" estudada, maior o alcance ao longo da faixa de influências intersimbólicas, mais fraca torna-se a gramática e as palavras mais independentes estatisticamente (Goldman, 1968:17). Isso sugere processos sem integralidades ou pelo menos com as mesmas fracas, tal que a função memória (ou faixa de influências) é progressivamente perdida. Já em sinais com alguma forma de organização (que em nível mais baixo pode se manifestar como ordenação ou a "arrumação" em senso comum), a redundância pode ter valor máximo em outra ordem que não a primeira. Simulações mostram que textos progressivamente arrumados apresentam crescimento de redundância em ordens diferentes da primeira.

Mais uma vez, para podermos avaliar o real significado da evolução de R, temos que analisar os arranjos sígnicos permitidos e suas distribuições, fugindo ao domínio do *ensemble*. Na distribuição de redundâncias podemos encontrar máximos em uma ou algumas ordens *n*. Isso indica que estes podem estar sendo gerados pela associação de signos muito frequentes ou uma mescla de signos frequentes com alguns mais raros. Assim, a redundância de 1a. ordem exprime, por exemplo, o domínio de um signo diante dos demais. Mas este poder pode ser amplificado quando o signo em questão ocorre com um outro também frequente ou pode ser atenuado quando o arranjo envolve o signo frequente com um raro. É importante lembrar que, do ponto de vista de uma gramática, *não ficamos restritos aos arranjos entre signos frequentes, como se estes fossem os mais gramaticais. Arranjos que envolvem signos de pesos estatísticos diversos podem ser, inclusive, os mais importantes quanto ao processo estudado.*

Vemos assim que a evolução de R pode ser usada no estudo de um sinal particular, mas quando pretendemos comparar sinais, cuidados devem ser tomados: valores idênticos de redundância não esclarecem quanto aos signos particulares que a estão gerando.

Grandes subcadeias homogêneas em meio a várias heterogêneas elevarão R e a manterão por várias ordens. São homogeneidades destacando-se na diversidade. É um caso diferente daquele em que tudo é heterogêneo. Da mesma forma, a Lei de Zipf em linguagens naturais e leis de formação de mensagens, também exponenciais, como encontramos em Goldman, Shannon, etc., podem caracterizar quantitativamente textos e cadeias.

Podemos estabelecer a chamada *evolução em Redundâncias*, visível por meio de um gráfico onde no eixo das ordenadas são colocados os números inteiros naturais que denotam a ordem de redundância e no eixo coordenado os valores numéricos das redundâncias para essas ordens. O gráfico é discreto mas, para facilitar a visualização, podemos unir os pontos obtidos por segmentos de reta. O que notamos, como ocorre para processos estocásticos, é que a redundância cai progressivamente com a ordem. E cada uma destas ordens refere-se a um tipo de arranjo de signos. Assim, a ordem dois descreve os arranjos dos signos tomados dois a dois, e assim por diante.

Uma redundância de primeira ordem é obtida quando os signos que compõem o texto codificado são considerados isolados. Calculando-se a distribuição de frequências de ocorrência dos mesmos, calculamos a entropia real e a ideal (esta última sendo o logarítmo do número que expressa o tamanho do alfabeto). Daí, calculamos a redundância, dita então de primeira ordem.

Uma redundância de segunda ordem é obtida quando os signos são tomados agora 2 a 2, segundo suas ocorrências no texto. Passamos a ter assim um novo alfabeto, novos signos, que são os pares que ocorrem - a gramaticalidade começa a manifestar-se no fato de que nem todos os arranjos tomados 2 a 2 possíveis a partir do primeiro alfabeto ocorrem. Surgem restrições, alguns arranjos ocorrendo e outros não. O alfabeto de segunda ordem é sempre menor do que aquele previsto pela Matemática Combinatorial. Os cálculos de entropias e redundâncias são então repetidos.

O processo prossegue, com a sucessiva construção de ordens mais elevadas, com a geração de "palavras" de comprimento maior. Essa construção evidencia como as redundâncias evoluem, logo como se comporta o alcance ou vigor gramatical no sinal.

Quando ocorre uma determinada queda no valor da redundância entre duas ordens consecutivas, isso é uma medida quantitativa da perda do "vigor" gramatical que naturalmente ocorre na construção de "palavras" progressivamente maiores. Em alguns casos de organização associada a algum critério forte de ordenação, pode ocorrer um crescimento de redundância para uma ordem que não é a primeira, gerando uma diferença negativa.

4. Conclusões

O método acima descrito permite trabalhar textos gerais, e não somente os gerados pelas linguagens naturais humanas, como sistemas organizados (ad-

mitindo-se gramaticalidade, como expressa por ordens de redundância, como sendo um critério de organização). Características de conjunto (*ensemble*) podem ser então quantificadas, como uma forma de entropia e a redundância implicada; por outro lado, tais características partilhadas podem ser analisadas pelo estudo da ocorrência de signos ou subsistemas de signos que formam os arranjos sígnicos considerados na construção das redundâncias de ordem superior. A presente discussão apoia-se na chamada *faixa de influências intersimbólicas* , um sintoma ou índice tanto do conceito de gramaticalidade utilizado quanto de uma *função memória* do texto enquanto sistema (Bunge, 1979: 161).

REFERÊNCIAS BIBLIOGRÁFICAS

ANDERSON, Myrdene.; DEELY, John; KRAMPEN, Martin; RANSDELL, Joseph; SEBEOK, Thomas; VON UEXKULL, Thure. (1984). A Semiotic Perspective on the Sciences: Steps toward a New Paradigm. *Semiotica 52-1/2, 7-47.*

BLOONFIELD, Peter. (1976). *Fourier Analysis of Time Series: an Introduction*. New York: John Wiley and Sons.

BUNGE, Mario (1977). *Treatise on Basic Philosophy – Vol 3.* Dordrecht: D. Reidel Publ. Co.

_____ (1979). *Treatise on Basic Philosophy – Vol 4.* Dordrecht: D. Reidel Publ. Co.

DENBIGH, Kenneth. (1975) *A Non-Conserved Function for Organized Systems.* Em *Entropy and Information in Science and Philosophy*, KUBAT, Libor; ZEMAN, Jiri. (Eds). Praga: Elsevier Sci. Publ. Co., 83-91.

GOLDMAN, Stanford. (1968) *Information Theory.* New York: Dover Publ. Inc.

MORIN, E. (1986) *O Método – Vol. III: O Conhecimento do Conhecimento.* Mira-Sintra: Publicações Europa-América.

UYEMOV, Avanir (1975) *Problem of Direction of Time and the Laws of System's Development.* Em *Entropy and Information in Science and Philosophy.* Kubat, L.; Zeman, J. (Eds.). 93-102. Praga: Elsevier Sc. Publ. Co.

VIEIRA, Jorge de Albuquerque (2000) "Organização e Sistemas". *Informática na Educação: Teoria e Prática / Programa de Pós-Graduação em Informática na Educação – vol. 3, n. 1.* Porto Alegre, UFRGS, 11-24.