# CORRESPONDÊNCIA CIENTÍFICA DE BERTHA LUTZ: UM ESTUDO DE APLICAÇÃO DA LEI DE ZIPF E PONTO DE TRANSIÇÃO DE GOFFMAN EM UM ARQUIVO PESSOAL

**Resumo**

Identifica a viabilidade de aplicação das leis de Zipf e Ponto de Transição de Goffman em um arquivo pessoal, o de Bertha Maria Júlia Lutz (1894-1976), com vistas a encontrar palavras com alto conteúdo semântico para sua indexação. Filha do cientista brasileiro Adolpho Lutz, Bertha foi cientista formada pela Sorbonne, feminista, deputada federal e professora emérita da UFRJ. Especializou-se em anfíbios anuros e exerceu seu trabalho no Museu Nacional/UFRJ e no Instituto Oswaldo Cruz. Zipf foi linguista e formulou duas leis baseadas na observação empírica e na análise de frequência de ocorrência de palavras em um texto *suficientemente* longo de artigos técnico-científicos. Essas leis foram complementadas por diversos estudos, enriquecidas pelo Ponto de Transição de Goffman e relacionam-se com a representação da informação e maior precisão na sua recuperação. A aplicação das leis foi realizada na correspondência científica e confirmou-se a viabilidade de seu emprego na indexação de arquivos.

**Palavras-chave**: Lei de Zipf. Ponto de Transição de Goffman. Correspondência científica. Lutz, Bertha

**Maria José V. C. Santos**
Professora do Curso de Biblioteconomia e Gestão de Unidades de Informação (CBG) – Faculdade de Administração e Ciências Contábeis/UFRJ.
Mestre em Ciência da Informação (IBICT/UFRJ).
maze@mn.ufrj.br

**BERTHA LUTZ´S SCIENTIFIC CORRESPONDENCE: A STUDY OF APLICATION OF THE ZIPF LAW AND THE GOFFMAN TRANSITION POINT IN A PERSONAL ARCHIVE**

**Abstract**

The present paper verifies the possibility of applying the Zipf´s laws ando also the Transition formula of Goffman in a personal archive of the Bertha Maria Júlia Lutz (1894-1976), in order to find words with a high semantic content for its indexing. Daughter of the scientistdAAdolpho Lutz, Bertha scientist was formed by the Sorbonne, feminist, Congresswoman and emeritus professor of UFRJ. Bertha was Adolpho Lutz´s daughter, a famous brazilian scientist. Graduated from Sorbonne, feminist, she was a federal deputie and an emeritus professor of Universidade Federal do Rio de Janeiro (UFRJ). She specialized herself in the studies of amphibians and she worked at Museu Nacional / UFRJ and Instituto Oswaldo Cruz. Zipf was a linguist and formulated two laws based on empirical observation and also in the analysis of the words frequency of occurrence in technical and scientific articles long enough to allow the analysis. These laws were enhanced by several studies, including Transition Formula of Goffman and they regard the information indexing and also they allow better precision in the retrieval. The application of these laws was also used in the scientific communication and confirmed its usability in achives indexing.

**Key-words:** Zipf Law. Goffman Transition Point. Scientific Correspondence. Lutz, Bertha

## 1 INTRODUÇÃO

As cartas tiveram um papel basilar para a comunicação na ciência. Utilizadas para a transmissão de conhecimento e difusão de ideias, eram trocadas entre pesquisadores e cientistas para relatar ideias originais para serem discutidas, opinadas e avaliadas pelos pares.

Bazerman (2006) acredita que as cartas contribuíram no surgimento de gêneros distintos. O primeiro artigo científico, segundo ele, emergiu da correspondência do alemão Henry Oldenburg com outros estudiosos. Oldenburg foi o primeiro editor do *Philosophical Transactions of the Royal Society of London* que, junto com o *Journal des Savants*[1] publicado na França, são os primeiros periódicos científicos que surgiram, ambos publicados no ano de 1665. Os primeiros números do *Philosophical Transactions of the Royal Society of London* foram editados sob a forma de resumo dessa correspondência e das reuniões da Royal Society.

A rede de comunicação de correspondentes é conhecida pela expressão colégio invisível utilizada pela primeira vez, de acordo com Merton (1968), por Boyle e reutilizada por Solla Price referindo-se a um grupo de pesquisadores que mantinha correspondência entre si. (MUELLER, 1994).

No campo da Ciência da Informação, a bibliometria é definida, segundo Tarapanoff (AMARAL, 2006), como o estudo de aspectos quantitativos da produção, distribuição e uso da informação registrada, a partir de modelos matemáticos, para o estabelecimento de critérios e tomadas de decisões na área estudada. É um importante instrumento para estabelecer indicadores em uma determinada área do conhecimento porque apresenta os aspectos quantitativos da produção, disseminação e uso da informação científica registrada.

[1] Grafia antiga: *Journal des Sçavants*

A bibliometria é a ciência que apresenta um conjunto de leis e princípios empíricos, baseados na observação, utilizando métodos matemáticos e estatísticos para investigar, avaliar e quantificar os processos de comunicação escrita. Dentre as diversas leis existentes e mais utilizadas destacam-se: a lei de Bradford que estima a relevância de periódicos em determinada área; lei de Lotka que estima a relevância de autores em determinada área; os estudos de obsolescência e vida-média da literatura científica, e as leis de Zipf relacionadas à frequência de ocorrência de palavras em um determinado texto, essa última, enriquecida com o Ponto de Transição de Goffman, utilizada na realização dessa pesquisa com o objetivo principal de verificar a viabilidade da utilização de ferramentas das leis e princípios da bibliometria, relacionados ao estudo estatístico e matemático da frequência de ocorrência das palavras em uma correspondência científica de um arquivo privado.

Entende-se por arquivo, a reunião de documentos contendo, segundo Belloto (2004?), informações sobre o estabelecimento, a competência, as atribuições, as funções, as operações e as atuações levadas a efeito, por uma entidade pública ou privada, no decorrer de sua existência. Existem, portanto, arquivos pertencentes a entidades coletivas públicas e privadas e arquivos pertencentes a pessoas físicas, os arquivos pessoais ou privados.

Espera-se identificar a possibilidade de aplicação das leis de Zipf e Ponto de Transição de Goffman nas cartas do arquivo privado de Bertha Lutz com vistas a identificar palavras com alto conteúdo semântico para facilitar a indexação dos assuntos tratados na correspondência e com isso, contribuir para melhor precisão na recuperação da informação por diferentes usuários.

## 2 LEIS DE ZIPF E PONTO DE TRANSIÇÃO DE GOFFMAN

A primeira lei de Zipf está relacionada às palavras de alta frequência em um texto e a segunda às de baixa frequência, e foram formuladas a partir da observação empírica e da análise de frequência de ocorrência de palavras em um texto suficientemente longo.

A primeira lei de Zipf está assim formulada: "o produto da ordem de série(r) de uma palavra, pela sua frequência (f) é aproximadamente constante (c).

$$r \times f = c$$

A segunda lei enuncia que "em um texto, várias palavras de baixa frequência de ocorrência têm a mesma frequência".

Essas leis foram complementadas por diversos estudos, destacando-se a modificação proposta por Booth para a segunda lei, representada matematicamente da seguinte forma:

$$\frac{I_1}{I_n} = \frac{n\,(n+1)}{2}$$

Onde $I_1$ é o número de palavras que têm frequência 1; $I_n$ é o número de palavras que têm frequência n; e o número 2 representa a constante válida para a língua inglesa.

Goffman, a partir da segunda lei de Zipf modificada por Booth, admitiu que havia uma região na lista de palavras, localizada entre as palavras de alta frequência e as de baixa frequência, região essa a qual ele denominou de "ponto de transição" (ponto T), com probabilidade de concentrar as palavras de alto conteúdo semântico, ou seja, palavras mais significativas e representativas do conteúdo temático ou intelectual de um texto.

O Ponto T de Goffman é representado matematicamente pela expressão:

$$n = \frac{-1 + \sqrt{1 + 8I_1}}{2}$$

Onde n representa o ponto T; $I_1$ é o número de palavras que tem frequência 1.

As leis de Zipf enriquecidas com o ponto T de Goffman, relacionam-se diretamente com a representação da informação e maior precisão na sua recuperação.

Nessa pesquisa propõe-se realizar um estudo de viabilidade de indexação temática nas cartas do arquivo pessoal de Bertha Lutz. Mostra a aplicação da lei de Zipf enriquecida com o cálculo do Ponto de Transição de Goffman em uma amostra representada por resumos do conjunto de 100 cartas trocadas entre a cientista Bertha Lutz e seus correspondentes, de forma a localizar palavras de alto conteúdo semântico para sua indexação.

## 3 BERTHA LUTZ: CIENTISTA, FEMINISTA, DEPUTADA FEDERAL

Bertha Maria Júlia Lutz nasceu em São Paulo no ano de 1894 e morreu no Rio de Janeiro em 1976. É nacionalmente conhecida por seu trabalho científico e sua atuação como líder feminista. Filha da enfermeira inglesa Amy Fowler e do cientista pioneiro da medicina tropical no Brasil, Adolpho Lutz, Bertha formou-se em biologia em Paris, pela Sorbonne, especializou-se em anfíbios anuros, mas exerceu seu trabalho também em outras especialidades da biologia, tendo trabalhado em instituições de renome como, o Museu Nacional da Universidade Federal do Rio de Janeiro (UFRJ), onde ingressou, por concurso público, e o Instituto Oswaldo Cruz, ambos no Rio de Janeiro. Foi a segunda mulher a ingressar no serviço público brasileiro.

Parte do acervo documental de Bertha Lutz está custodiado na Seção de Memória e Arquivo (Semear) do Museu Nacional da UFRJ. Outra parcela dessa documentação também pode ser encontrada no fundo[2] Federação Brasileira para o Progresso Feminino, entidade da

[2] Conjunto de documentos organicamente produzido e/ou acumulado e utilizado por um indivíduo, família ou entidade coletiva no decurso das suas atividades e funções. (CONSELHO INTERNACIONAL DE ARQUIVOS, 2001, p. 5).

qual Bertha Lutz foi fundadora e presidente e que se encontra sob a guarda do Arquivo Nacional.

O arquivo pessoal de Bertha Lutz encontra-se organizado com as seguintes séries documentais:[3] Feminismo; Documentos Pessoais; Produção Científica – aí incluída a correspondência; Conselho Federal das Expedições Artísticas e Científicas; Conselho Federal Florestal, e Adolpho Lutz.

Para o presente estudo foi selecionada a série Produção Científica, especificamente a correspondência científica.

## 4 APLICAÇÃO DA LEI DE ZIPF E PONTO DE TRANSIÇÃO DE GOFFMAN

As leis de Zipf enriquecidas com o Ponto de Transição de Goffman foram aplicados nos resumos dos conteúdos de 100 cartas do arquivo pessoal de Bertha Lutz, considerados nesse estudo como um texto "suficientemente longo", de acordo com premissa de Zipf para aplicação das leis.

A partir daí, foi produzida uma listagem de palavras acompanhadas de respectivas frequências de aparecimento no texto, que permitiu delimitar a região de concentração de palavras de alto conteúdo semântico, o ponto T de Goffman, para a indexação da correspondência científica de Bertha Lutz.

A metodologia de aplicação consistiu das seguintes etapas:

- Etapa 1 – Delimitação da amostra - foram selecionadas na correspondência científica do fundo Bertha Lutz, Série Produção Científica, os resumos das 100 (cem) primeiras cartas dessa série. Esses resumos foram consolidados em um único texto,

[3] "Subdivisão de um fundo documental [...] por resultarem de um mesmo processo de acumulação, ou de uma mesma atividade." (CONSELHO INTERNACIONAL DE ARQUIVOS, 2001, p. 5).

considerado nesse estudo "suficientemente longo" para ser analisado segundo Zipf e Goffman;

➢Etapa 2 – Contagem das palavras - foi utilizado o *software* contador de palavras *Rank Words* 2.0.[4]

➢Etapa 3 – Listagem e ordenação das palavras - o *software* produziu um quadro em 3 colunas assim distribuídas: palavras, frequência em ordem decrescente, e o *rank* das palavras*;*

➢Etapa 4 – Aplicação da fórmula do Ponto de Transição de Goffman – foi aplicada a fórmula do Ponto de T de Goffman para identificar a frequência, onde ocorre a transição das palavras de alta frequência para as palavras com baixa frequência de ocorrência;

➢Etapa 5 – Delimitação da região de Transição de Goffman – foi identificada na listagem, a faixa com potencial para representar o conteúdo temático do texto, palavras que deverão ser utilizadas para a indexação e que deverão oferecer maior precisão na recuperação da informação.

## 4 RESULTADOS

Os resultados serão analisados por meio do quadro gerado pelo *software Rank Words*[5] e pela aplicação da fórmula do Ponto de Transição de Goffman a partir dos resultados desse quadro.

Esse quadro apresenta 3 colunas contendo as seguintes informações: palavras, frequência em ordem decrescente, e o *rank* das palavras

[4]Disponível em: <http://download.cnet.com/Rank-Words3000-2279_4-10909564.html>.

[5] Ferramenta encontrada na Web, no site CNET DOWLOAD, utilizada para identificar todas as palavras usadas em um texto, listadas de acordo com a sua frequência.

Encontrou-se no quadro, o total de 1902 palavras, que foram retiradas do conteúdo do texto dos resumos das 100 cartas do arquivo privado de Bertha Lutz. Desse total observou-se que 618 palavras são palavras distintas que equivalem a 32,5% do total de palavras. Essas 618 palavras foram repetidas de uma (1) a 237 vezes. O índice médio de repetição de cada palavra no texto do conteúdo das cartas é de 3,7 vezes.

O número de palavras que ocorreram uma única vez no texto, ou seja, o número de palavras com frequência 1, $I_1$ da fórmula matemática para o cálculo do ponto T abaixo, é de 406 palavras.

O Ponto T de Goffman foi calculado aplicando-se a fórmula matemática a seguir:

$$n = \frac{-1 + \sqrt{1 + 8I_1}}{2}$$

Aplicando-se a fórmula aos resultados encontrados no quadro do *Rank Words*, tem-se:

$$N = \frac{-1 + \sqrt{1 + (8 \times 406)}}{2}$$

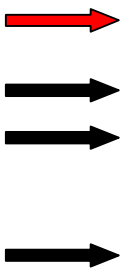
$$n = \frac{-1 + \sqrt{1 + 3248}}{2} = \frac{-1 + \sqrt{3249}}{2} = \frac{-1 + 57}{2} = \frac{56}{2} = 28$$

Segundo os cálculos com a aplicação da fórmula do Ponto de Transição de Goffman, encontrou-se o valor 28, valor que se refere à frequência de palavras, ou seja, o Ponto "T" localiza-se na frequência 28 do quadro de distribuição de palavras do *software Raking Words*. Esta frequência 28 está associada ao termo Lutz, considerado de alto conteúdo semântico em relação à correspondência, já que a titular do arquivo chama-se Bertha Lutz.

Na margem de frequências entre 21 a 37, considerou-se a região onde se concentram as palavras de alto conteúdo semântico (palavras-chave, descritores), tais como: pedido, informações e sapos – palavras também relacionadas às atividades da cientista, já que ela era especialista em anfíbios anuros, classe de animais onde estão incluídos os sapos.

No extrato do quadro gerado pelo *software* contador de palavras, a seguir, pode-se observar a delimitação da zona de concentração de palavras de alto conteúdo semântico.

**Quadro 1 - Extrato do quadro do *software Rank Words***



| Palavras | Frequência | Rank |
|---|---|---|
| **Envio** | 37 | 5 |
| **sobre** | 33 | 6 |
| **Lutz** | **27** | 7 |
| **para** | 26 | 8 |
| **Pedido** | 25 | 9 |
| **informações** | 25 | 10 |
| **da** | 25 | 11 |
| **em** | 22 | 12 |
| **sapos** | 21 | 13 |

**Fonte**: Dados da pesquisa.

## 5 CONCLUSÕES

As leis de Zipf enriquecidas com o Ponto de Transição de Goffman foram aplicadas no conteúdo de 100 cartas contidas no arquivo pessoal de Bertha Lutz para verificar a viabilidade de aplicabilidade dessas leis em documentos de arquivo. Verificou-se que o objetivo principal da pesquisa foi alcançado considerando-se válida a aplicação das referidas leis a documentos de arquivo, uma vez que os resultados apontaram uma zona de concentração de palavras de alto conteúdo semântico que podem ser utilizadas na indexação temática da correspondência da cientista, conforme formulação das leis.

As palavras que se encontravam na zona de concentração do ponto T, tais como Lutz – nome da titular do arquivo; pedido, informações e sapos – animal estudado por Bertha Lutz, representam realmente o conteúdo das cartas.

A título de experiência, aplicou-se as leis a um número menor de cartas (50 cartas) e pode-se constatar também, que não foram registradas mudanças significativas no resultado da pesquisa, apresentando também palavras de alto conteúdo semântico.

Recomenda-se que outros estudos sejam realizados com outros tipos de documentos arquivísticos de fundos específicos, para confirmar a viabilidade de aplicação das leis em material de arquivo.

Espera-se, com esse estudo, contribuir para a pesquisa na área de bibliometria e obter, em estudos futuros, resultados que apontem para um maior grau de precisão na indexação de material de arquivo, precisamente em correspondências científicas.

**Artigo submetido em 18/10/2009 e aceito para publicação em 09/12/2009.**

REFERÊNCIAS

**AMARAL, R. M. et al. Criação de indicadores sobre o serviço de comutação bibliográfica da BCo/UFSCar em 2004-2005, através de análise bibliométrica automatizada. In: SEMINÁRIO NACIONAL DE BIBLIOTECAS UNIVERSITÁRIAS, 14., 2006. Salvador.** Anais... **Salvador: SNBU, 2006.**

BAZERMAN, C. Cartas e a base social de gêneros diferenciados. In _____. **Gêneros textuais, tipificação e interação**. 2. ed. São Paulo: Cortez, 2006.

CONSELHO INTERNACIONAL DE ARQUIVOS. **ISAD (G)**: norma geral internacional de descrição arquivística. 2. ed. – Rio de Janeiro: Arquivo Nacional, 2001.

BELLOTO, H. L. Prefácio: Inventário dos acervos das escolas técnicas estaduais do estado de São Paulo. In: MORAES, C. S. V.; ALVES, J. F. (Org.). **Contribuição à pesquisa do ensino técnico no estado de São Paulo:** inventário de fontes documentais. São Paulo: Centro Paula Souza, [2004?]

MUELLER. S. P. M. A ciência, o sistema de comunicação científica e a literatura científica. In: CAMPELLO, B. S.; CENDON, B. V.; KREMER, J. M. (Org.). **Fontes de informação para pesquisadores e profissionais**. Belo Horizonte: Editora da UFMG, 2000.