

LEIS BIBLIOMÉTRICAS DE ZIPF E PONTO DE TRANSIÇÃO DE GOFFMAN:

reflexões com estudos pioneiros

Thulio Pereira Dias Gomes

 <https://orcid.org/0000-0002-2817-1253>.

✉ thuliogomes@id.uff.br.

🏢 Universidade Federal Fluminense |

ROR:  <https://ror.org/02rjhbb08> | Niterói, Brasil.

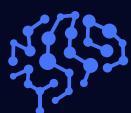
Eixo temático: Produtividade e Colaboração Científica

Modalidade: Resumo expandido

DOI: 10.22477/ix.ebbc.304

Resumo: Este trabalho propõe uma reflexão histórica e teórica sobre o fenômeno da frequência de palavras nos textos presentes em documentos científicos, como os artigos de periódicos, a partir de um experimento bibliométrico pioneiro de aplicação das leis bibliométricas de Zipf. Para isso, este estudo desenvolve uma pesquisa bibliográfica a partir do relato de Pao (1978) sobre a análise automática de textos instrumentalizada pelas leis de Zipf e do ponto de transição de Goffman. Os resultados indicam que os estudos que reverberam o experimento ampliam os quadros explicativos do campo informacional para o fenômeno da frequência de palavras em textos dos documentos científicos.

Palavras-Chave: Análise de textos. Frequência de palavras. Leis Bibliométricas de Zipf.



1 INTRODUÇÃO

George Kingsley Zipf foi um linguista interessado na modelagem matemática da frequência de palavras de um determinado texto. Zipf se dedicou aos estudos de mudanças fonéticas, da frequência de uso de fonemas e da tendência de variações em longos períodos de tempo. Em um de seus trabalhos, Zipf (1949) considerou a teoria psicológica do menor esforço para estabelecer que os seres humanos tendemos a minimizar os esforços para obter resultados exitosos. O autor aplicou a lei do menor esforço à obra *Ulisses*, de James Joyce: organizou as palavras desta obra em forma decrescente de ocorrência e concluiu que os falantes tendem a preferir palavras mais habituais em vez de palavras pouco utilizadas. Em outros termos, pelo princípio do menor esforço, os seres humanos favorecemos o comum e dificultamos o desconhecido. A tendência se explica porque geralmente é mais fácil utilizar uma palavra conhecida do que uma desconhecida. Há uma inclinação humana em preferir minimizar o esforço no uso de vocabulário a se valer da erudição para uso de palavras menos conhecidas (Urbizagástegui Alvarado; Restrepo Arango, 2011).

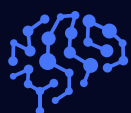
A lei de Zipf foi aplicada, revista e modificada em diversos contextos e em diferentes pesquisas, como revisa Piantadosi (2014). As aplicações da lei são, entre muitas outras, para otimização de entropia, descrições simbólicas de sistemas complexos, organização comunicativa e aperfeiçoamento de desempenho da memória. Na ciência da informação, a lei de Zipf demonstrou sua utilidade em estudos em diferentes contextos de processamento da linguagem natural para desenvolver corpora linguísticos, ontologias, taxonomias, indexação automática e outras aplicações na organização da informação e do conhecimento.

Como técnica de apoio à indexação em contextos de organização do conhecimento, as primeiras aplicações das leis bibliométricas de Zipf podem parecer ineficientes e ineficazes para atividades documentárias, pois já existem técnicas terminológicas de análise de documentos mais rigorosas e recursos de tecnologia mais sofisticados, como os mapas de *co-words*, que podem fazer as técnicas pioneiras parecerem ingênuas. Todavia, para a história das ciências e das técnicas, é complicado afirmar que um instrumento ou uma técnica estão ultrapassados, porque cada tecnologia é avançada segundo as condições de seu tempo.

As explicações sobre as leis bibliométricas de Zipf costumam enfatizar a validade matemática dos resultados. Porém, o objetivo deste trabalho é propor uma reflexão histórica e teórica sobre o fenômeno da frequência de palavras nos textos presentes nos documentos científicos, como os artigos de periódicos, a partir de um experimento bibliométrico pioneiro de aplicação das leis de Zipf. Para isso, este estudo desenvolve uma pesquisa bibliográfica a partir do relato de Pao (1978) sobre a análise automática de textos instrumentalizada pelas leis de Zipf e do ponto de transição de Goffman.

2 UM EXPERIMENTO BIBLIOMÉTRICO PIONEIRO

Miranda Lee Pao foi uma pesquisadora que atuou na *School of Library Science da Case Western*



Reserve University¹, em Cleveland, Ohio, Estados Unidos da América (EUA). Em um de seus trabalhos, a pesquisadora reportou um experimento de análise automática de textos que observa o fenômeno de transição de ocorrência de palavras em artigos científicos. O estudo, apresentado em 1977 e publicado em 1978 no *Journal of the Association for Information Science and Technology*, descreve um método para seleção de índices temáticos (*index terms*) diretamente de uma lista de frequência de palavras em artigos científicos. O procedimento visa a contribuir com a indexação que constitui-se, basicamente, de duas operações: análise de conteúdo dos documentos e tradução de tal conteúdo de acordo com uma linguagem documentária. Pao (1978) explica que a indexação automática é a mecanização desse processo, em parte ou no todo, para a redução da complexidade da análise de documentos e para a seleção de termos apropriados para um conjunto de algoritmos.

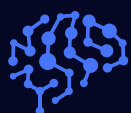
Pao (1978) aplicou as leis bibliométricas de Zipf sobre a regularidade da frequência de palavras e propõe uma abordagem para a identificação de assuntos na distribuição de palavras de alto valor semântico em determinado texto. A ideia do experimento surgiu a partir de sugestões em conversas com seu orientador, William Bill Goffman. A autora propõe o que denomina como um “experimento simples” para demonstrar a simplicidade da extração automática dos termos em relação à indexação humana.

Pao (1978) explica que, embora sejam regras totalmente diferentes, as leis bibliométricas de Zipf descrevem as duas extremidades da distribuição de palavras em um texto. A primeira lei opera condicionalmente apenas na extremidade onde encontram-se as palavras de alta frequência; enquanto a segunda lei observa a região extrema das palavras de baixa frequência. As palavras mais ocorrentes tendem a ocupar um *rank* exclusivo na distribuição das palavras, não havendo palavra mais ocorrente com a mesma frequência que outra palavra de alta ocorrência. Enquanto na outra extremidade há palavras de baixa ocorrência que têm a mesma frequência, e muitas delas ocorrem apenas uma vez no texto.

É razoável esperar uma região crítica na qual ocorre uma mudança de comportamento de palavras de alta frequência para as de baixa frequência. Pao (1978) relata uma conversa com Goffman em que o pesquisador hipotetizou que, em um *ranking* de palavras de um determinado texto, há regiões com unidades mais com maior ou menor valor semântico. Em uma primeira região, aparecem palavras de baixo valor semântico, como preposições, artigos e conjunções, que costumam ser as mais frequentes em um texto e correspondem às palavras caracterizadas na primeira lei de Zipf. As palavras de maior valor semântico estão na região intermediária do *ranking*. Na terceira região, há palavras de menor ocorrência, que tendem a ocorrer uma única vez em todo texto.

Pao (1978) analisou dois artigos de um mesmo autor. O primeiro artigo tem o título *On the geometry of libraries*, de Booth (1969), no qual foram analisadas 559 palavras diferentes, das quais 256 ocorreram uma única vez. O segundo artigo analisado foi *A law of occurrences of words of low frequency*, de Booth (1967), no qual foram analisadas 327 palavras diferentes, das quais 188 ocorreram apenas uma vez. Em ambas experiências, foi possível identificar as palavras mais significativas de cada artigo a partir da aplicação do ponto de transição para analisar cada texto. Pao (1978) conclui que, com a aplicação da lei,

¹ Case Western Reserve University é uma proeminente universidade privada dos EUA, cujas atividades remontam a 1826. A *School of Library Science* foi criada em 1903 e permaneceu em atividade até 1986, quando era nomeada *Matthew A. Baxter School of Information and Library Science*. Fizeram parte de seu corpo docente professores como Jesse Hauk Shera, John A. Rowell, Miranda Lee Pao, Tefko Saracevic e William Goffman.



é possível extrair termos para a representação de textos, tendo em vista a sua recuperação automatizada. É interessante chamar a atenção para o fato de que o trabalho de Booth (1967) foi uma aplicação pioneira das leis de Zipf na incipiente ciência da informação.

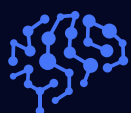
Em seguida, Pao (1978) acrescenta à sua experiência uma articulação com o ensino da observação do fenômeno da frequência de palavras, na disciplina *Introduction to Information Science*, na *School of Library Science*. A pesquisadora propôs um exercício de indexação do artigo de Booth (1969) sem o uso de um tesouro. Ela observa que nenhum dos estudantes tinha experiência anterior em indexação. Os estudantes, instruídos a não usarem mais do que oito termos, apontaram palavras que situam o artigo na temática da frequência de uso e arranjo de documentos em bibliotecas. Pao (1978) afirma que os resultados mais comuns nos índices dos estudantes estavam presentes na região de transição do artigo, o que a levou a concluir que a seleção automática de termos de uma lista de frequência de palavras é promissora para a indexação automática. A pesquisadora destaca a simplicidade de aplicação do método e reconhece que um dispositivo de controle de formas variantes das palavras traria resultados ainda melhores para a recuperação.

Curiosamente, Goffman publicou pouco sobre suas observações da aplicação das leis bibliométricas de Zipf. Todavia, suas observações foram desenvolvidas e divulgadas por seus colegas e alunos. Um exemplo disso aconteceu quando Tefko Saracevic, colega de Goffman na *Case Western Reserve University*, trouxe a experiência de análise automática de textos para as suas aulas no embrionário curso de pós-graduação em ciência da informação no Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), que reverberaram em outras pesquisas desenvolvidas no Brasil. A experiência foi replicada nas aulas e no grupo de pesquisa de Gilda Maria Braga, orientada por Saracevic no mestrado e por Goffman no doutorado e pesquisadora aposentada do IBICT. Maia (1973) e Ribeiro (1974) são dois estudos brasileiros desenvolvidos no IBICT que aplicam o ponto de transição e são publicados antes do artigo de Pao (1978) sobre a experiência na *Case Western Reserve University*. Ambos os trabalhos referem-se às aulas de Saracevic na descrição metodológica, o que sugere a utilidade das leis bibliométricas de Zipf associadas ao ponto de transição para o ensino e para a formação da primeira geração de pesquisadores brasileiros em Ciência da Informação (Gomes, 2022).

3 DO FÍSICO AO SOCIAL ESTUDO DAS PALAVRAS

O experimento bibliométrico de Pao (1978) foi uma aplicação de análise automática de textos a partir de contagem de frequência de palavras desenvolvida no período caracterizado pelo paradigma físico da ciência da informação. Segundo esse paradigma, a informação é um fenômeno objetivo e pode ser estudada cientificamente, sem considerar dimensões subjetivas do campo informacional. Nesse paradigma, excluem-se, do conceito de informação e da agenda de pesquisa, as dimensões de significação e de relação social, descartando a subjetividade e a contingência como elementos componentes da informação. Os resultados da experiência apresentados pela autora como “experimento simples”, almejam uma análise de textos sem considerar outras dimensões de significação e subjetividades dos textos e das palavras.

Outros pesquisadores, como Gilda Braga e Tefko Saracevic, ampliaram, em certa medida, as pos-



sibilidades explicativas das leis de Zipf ao aplicarem a ferramenta bibliométrica em textos estruturados em outros gêneros textuais, classificados em outras áreas do conhecimento, escritos em outros idiomas, especialmente espanhol, inglês e português. As aplicações das leis bibliométricas de Zipf passaram a contemplar outras esferas da ação humana no âmbito de contextos socioculturais concretos a partir da análise da frequência de ocorrência de palavras. Dessa forma, as aplicações dessa ferramenta foi ampliada para uma diversidade de *corpora*, com tipos de documentos variados, em idiomas diferentes e de múltiplos domínios do conhecimento. Para ilustrar essa diversidade de experiências de observação de frequência de palavras relevantes, mencionamos alguns dos artigos de periódicos que fizeram parte da revisão de literatura para esta pesquisa.

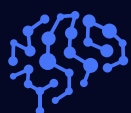
Maia (1973) analisou três artigos na área da bibliografia escritos em português e publicados em revistas brasileiras por autores reconhecidos. Ribeiro (1974) investigou uma amostra de editoriais do *Jornal do Brasil* publicados no período de 1959-1973 em língua portuguesa. Mamfrim (1991) analisou as possibilidades de indexação derivativa, mediante o ponto de transição, de artigos sobre a bibliometria escritos em português. Guedes (1994) analisou três gêneros documentais distintos na área de mecânica dos solos, em língua portuguesa: onze notas técnicas, uma revisão de literatura e um artigo de periódico. Santos (2009) aplicou o ponto de transição aos resumos das cartas do arquivo da cientista Bertha Lutz, escritos em português. Lapa e Corrêa (2010) aplicaram as leis de Zipf sobre os resumos das teses escritas em português na Universidade Federal da Paraíba (UFPB) para selecionarem os termos mais relevantes para indexação na biblioteca digital de teses da universidade. Guedes (2015) aplicou as leis bibliométricas em artigos de periódicos sobre a indústria de vinhos e da economia. Além desses exemplos publicados em periódicos científicos, Mamfrim (1991) e Braga (1996) indicam diversos trabalhos não publicados apresentados na disciplina *Bibliometria*, da Escola de Comunicação da Universidade Federal do Rio de Janeiro (UFRJ).

4 CONSIDERAÇÕES FINAIS

Pao (1978, p. 122, tradução nossa) define palavra como “uma sequência de caracteres precedida e seguida por um espaço em branco”². Essa concepção de palavra é notoriamente física e retira da análise de textos outras dimensões da palavra que poderiam contemplar os campos lexical, semântico, terminológico, conceitual, discursivo, poético, literário e até mesmo teológico. Apesar da restrição dos estudos de Miranda Pao à fisicalidade da palavra, observa-se que as pesquisas que reverberam o experimento ampliam os quadros explicativos para essas outras dimensões das palavras.

Braga (1996) reconheceu alguns desafios a serem superados em pesquisas de estruturas bibliométricas, indicando a necessidade de uma compreensão maior das variáveis envolvidas. Vinte e três anos depois, Araújo (2018) delineia um presente e indica um futuro em que as dimensões subjetiva e social da informação são consideradas nas pesquisas e nos estudos sobre os fenômenos informacionais. Com a motivação de superar tais desafios para conectar-se à análise social da informação, a discussão neste trabalho procurou religar elementos históricos e teóricos em direção a uma síntese capaz de explicar as

² Trecho original: “A word is defined as a string of characters preceded and followed by a blank space” (Pao, 1978, p. 122).



leis bibliométricas de Zipf reportadas por Pao (1978), a partir de um conjunto satisfatório, ou ao menos mais amplo, de elementos atuais para os estudos da ocorrência de palavras em textos científicos.

REFERÊNCIAS

- ARAÚJO, Carlos Alberto Ávila. Existe um pensamento informacional ibero-americano? **Logeion: filosofia da informação**, Rio de Janeiro, v. 4, n. 2, p. 31-55, 2018. DOI: <https://doi.org/10.21728/log-geion.2018v4n2.p31-55>. Disponível em: <https://revista.ibict.br/fiinf/article/view/4212>. Acesso em: 19 abr. 2024.
- BOOTH, Andrew D. A law of occurrences for words of low frequency. **Information and control**, [S. l.], v. 10, n. 4, p. 386-393, 1967. DOI: [https://doi.org/10.1016/S0019-9958\(67\)90201-X](https://doi.org/10.1016/S0019-9958(67)90201-X). Disponível em: <https://www.sciencedirect.com/science/article/pii/S001999586790201X>. Acesso em: 12 abr. 2024.
- BOOTH, Andrew D. On the geometry of libraries. **Journal of documentation**, [S. l.], v. 25, n. 1, p. 28-42, mar. 1969. DOI: <https://doi.org/10.1108/eb026463>. Disponível em: <https://www.emerald.com/insight/content/doi/10.1108/eb026463/full/html>. Acesso em: 12 abr. 2024.
- BRAGA, Gilda Maria. A representação da informação na desconstrução do contexto. **Informare: cadernos do Programa de Pós-Graduação em Ciência da Informação**, Rio de Janeiro, v. 2, n. 2, p. 53-57, jul./dez. 1996.
- GOMES, Thulio Pereira Dias. A observação da frequência de palavras relevantes em textos longos: revérberos das leis bibliométricas de Zipf e do ponto de transição de Goffman. In: SOUZA, Rosali Fernandez de; SALES, Luana; SALDANHA, Gustavo Silva (org.). **Teorias e domínios emergentes em organização do conhecimento**. Brasília, DF: IBICT, 2022. p. 61-78. Disponível em: https://ridi.ibict.br/bitstream/123456789/1241/1/SouzaSalesSaldanha_TeoriasDominiosEmergentesOrganizacaoConhecimento_2022.pdf. Acesso em: 3 maio 2024.
- GUEDES, Vânia Lisbôa da Silveira. A função de índice temático da nominalização deverbal na escrita científica: análise léxico-morfológica. **Informação e sociedade: estudos**, João Pessoa, v. 25, n. 3, p. 63-74, set./dez. 2015. Disponível em: <https://periodicos.ufpb.br/index.php/ies/article/view/25776>. Acesso em: 12 abr. 2024.
- GUEDES, Vânia Lisbôa da Silveira. Estudo de um critério para indexação automática derivativa de textos científicos e tecnológicos. **Ciência da informação**, Brasília, DF, v. 23, n. 3, p. 318-326, 1994. DOI: <https://doi.org/10.18225/ci.inf.v23i3.529>. Disponível em: <https://revista.ibict.br/ciinf/article/view/529>. Acesso em: 12 abr. 2024.
- LAPA, Reni; CORREA, Renato. Indexação automática no âmbito da ciência da informação no Brasil. **Informação e tecnologia**, [S. l.], v. 1, n. 2, p. 59-76, jul./dec. 2014. Disponível em: <https://periodicos.ufpb.br/index.php/itec/article/view/21408>. Acesso em: 12 abr. 2024.
- MAIA, Elza Lima e Silva. Comportamento bibliométrico da língua portuguesa, como veículo de representação da informação. **Ciência da informação**, Brasília, DF, v. 2, n. 2, p. 99-138, 1973. DOI: <https://doi.org/10.18225/ci.inf.v2i2.31>. Disponível em: <https://revista.ibict.br/ciinf/article/view/31/0>. Acesso em: 12 abr. 2024.



MAMFRIM, Flávia Pereira Braga. Representação de conteúdo via indexação automática em textos integrais em língua portuguesa. **Ciência da informação**, Brasília, DF, v. 20, n. 2, p. 191-203, jul./dez. 1991. DOI: <https://doi.org/10.18225/ci.inf.v20i2.355>. Disponível em: <https://revista.ibict.br/ciinf/article/view/355>. Acesso em: 12 abr. 2024.

PAO, Miranda Lee. Automatic text analysis based on transition phenomena of word occurrences. **Journal of the American Society for Information Science**, [S. l.], v. 29, n. 3, p. 121-112, 1978. DOI: <https://doi.org/10.1002/asi.4630290303>. Disponível em: <https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.4630290303>. Acesso em: 19 abr. 2024.

PIANTADOSI, Steven T. Zipf's word frequency law in natural language: a critical review and future directions. **Psychonomic bulletin & review**, [S. l.], v. 21, n. 5, p. 1112-1130, out. 2014. DOI: <https://doi.org/10.3758/s13423-014-0585-6>. Disponível em: <https://rdcu.be/dFaKh>. Acesso em: 19 abr. 2024.

RIBEIRO, Laís A. Aplicação dos métodos estatísticos e da teoria da informação e da comunicação na análise linguística: estudo da linguagem jornalística. **Ciência da informação**, Brasília, DF, v. 3, n. 2, p. 151-154, 1974. DOI: <https://doi.org/10.18225/ci.inf.v3i2.49>. Disponível em: <https://revista.ibict.br/index.php/ciinf/article/view/49>. Acesso em: 12 abr. 2024.

SANTOS, Maria José Veloso da Costa. Correspondência científica de Bertha Lutz: um estudo de aplicação da lei de Zipf e ponto de transição de Goffman em um arquivo pessoal. **Pontodeacesso**, Salvador, v. 3, n. 3, p. 317-326, 2009. Disponível em: <https://periodicos.ufba.br/index.php/revistaici/article/view/3628/2751>. Acesso em: 12 abr. 2024.

URBIZAGÁSTEGUI ALVARADO, Rubén; RESTREPO ARANGO, Cristina. La ley de Zipf y el punto de transición de Goffman en la indización automática. **Investigación Bibliotecológica**, Ciudad de México, v. 25, n. 54, p. 71-92, mayo/agosto 2011. Disponível em: https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0187-358X2011000200004. Acesso em: 12 abr. 2024.

ZIPF, George. **Human behavior and the principle of least effort**. Massachussets: Addison-Wesley Press, 1949.