

INTELIGÊNCIA ARTIFICIAL GENERATIVA, DISCURSO E AUTORIA: UM PANORAMA BIBLIOMÉTRICO DA PRODUÇÃO CIENTÍFICA INTERNACIONAL (2020- 2026)

**GENERATIVE ARTIFICIAL INTELLIGENCE, DISCOURSE, AND AUTHORSHIP:
A BIBLIOMETRIC PANORAMA OF INTERNATIONAL SCIENTIFIC
PRODUCTION (2020-2026)**

Ciências Humanas, Linguística & Letras e Artes, Ciências Sociais
Aplicadas

• 07/10/2026

REGISTRO DOI: [10.70773/revistatopicos/791354143](https://doi.org/10.70773/revistatopicos/791354143)

Francisco Alves de Freitas Neto¹

Carlos Henrique Medeiro de Souza²

Fabio Machado de Oliveira³

Leonard Barreto Moreira⁴

Camila Pinheiro Mascarenhas de Freitas⁵

Luiz Claudio Tavares Silva⁶

RESUMO

O avanço acelerado das Inteligências Artificiais Generativas (IAGs) tem provocado profundas inquietações nos ambientes educacionais e acadêmicos, tensionando as propriedades constitutivas da escrita e os parâmetros tradicionais da autoria. O presente artigo tem como objetivo geral mapear e analisar a produção científica internacional, no recorte longitudinal de 2020 a 2026, acerca dos tensionamentos provocados pelas IAGs no discurso e na reconfiguração da função-autor. A sustentação teórica articula os conceitos de função-autor de Foucault, tecnodiscursos de Paveau e cena enunciativa de Maingueneau. Metodologicamente, realizou-se um estudo bibliométrico quantitativo e exploratório na base *Web of Science*, com tratamento de metadados via rotinas em *Python*. O corpus consolidado de 667 publicações foi submetido à aplicação das Leis de Bradford, Lotka e Zipf, além da projeção de grafos relacionais via *NetworkX*. Os resultados revelam a liderança de produção entre China e Estados Unidos; a delimitação de um núcleo de Bradford de 40 periódicos que concentram 177 artigos (taxa de dispersão de 12,025); e forte assimetria na produtividade autoral aferida pela Lei de Lotka ($R^2 = 0,859$; expoente de 3,321). A validação da Lei de Zipf ($R^2 = 0,960$; expoente de -1,150) destacou o termo *stance* entre os mais frequentes, evidenciando o escrutínio sobre o ethos discursivo. A rede de coocorrência confirmou a centralidade dos modelos de linguagem associados à análise do discurso e à integridade acadêmica no ensino superior. Conclui-se que o campo converge para a emergência de modelos de autoria distribuída, demandando abordagens pedagógicas voltadas à recuperação da agência humana e ao exercício de uma autoria crítica.

Palavras-chave: Inteligência Artificial Generativa; Análise do Discurso; Função-Autor; Bibliometria; Escrita Acadêmica.

ABSTRACT

The rapid advancement of Generative Artificial Intelligence (GenAI) has raised profound concerns across educational and academic settings, straining the core properties of writing and traditional authorship parameters. This paper aims to map and analyze international scientific production from 2020 to 2026 regarding the disruptions caused by GenAI in discourse properties and the reconfiguration of the author-function. The theoretical framework integrates Foucault's author-function, Paveau's technodiscourse, and Maingueneau's enunciative scene. Methodologically, a quantitative and exploratory longitudinal bibliometric study was conducted on the Web of Science database, with metadata computationally processed using Python routines. The consolidated corpus of 667 publications was examined through Bradford's, Lotka's, and Zipf's Laws, alongside relational network mapping via NetworkX. Results show research leadership shared between China and the United States; a Bradford core of 40 journals accounting for 177 articles (dispersion rate of 12.025); and pronounced asymmetry in author productivity under Lotka's Law ($R^2 = 0.859$; exponent of 3.321). Zipf's Law validation ($R^2 = 0.960$; exponent of -1.150) highlighted the term stance among the most frequent words, demonstrating substantial scholarly scrutiny over ethos and discursive positioning. The co-occurrence network confirmed the role of language models as hubs linking discourse analysis to academic writing and integrity in higher education. The study concludes that the literature converges toward distributed authorship models, calling for pedagogical practices designed to reclaim human agency and foster critical authorship.

Keywords: Generative Artificial Intelligence; Discourse Analysis; Author-Function; Bibliometrics; Academic Writing.

1. INTRODUÇÃO

Os professores do Instituto Federal Fluminense (IFF) – campus Itaperuna têm manifestado inquietações a respeito da influência das Inteligências Artificiais Generativas (IAGs) nos trabalhos escritos entregues pelos estudantes. Eles relataram que, em certas situações, os textos exibem uma organização pouco direta, abordagens muito semelhantes e indícios de paráfrases de obras clássicas, além de dificuldades para contextualizar e interpretar questões mais complexas. Na área de computação, em particular, docentes dos cursos de Bacharelado em Sistemas de Informação e Técnico em Informática notam uma uniformidade nas estratégias utilizadas ao apresentarem os algoritmos para solucionar os problemas propostos, o que levanta a suspeita de possíveis cópias ou de forte intervenção de modelos de IAGs nas atividades de programação. Diante desse cenário, esses professores associam esses sinais ao uso de ferramentas de escrita com auxílio de inteligência artificial, como o ChatGPT da OpenAI. Embora não exista um consenso definitivo, há inquietações sobre os efeitos dessas tecnologias nos processos de avaliação e no desenvolvimento cognitivo, uma vez que o uso descontrolado pode ocultar o verdadeiro nível de conhecimento e entendimento dos alunos.

Por outro lado, a comunidade científica tem reconhecido o potencial disruptivo das tecnologias de IAGs nos processos criativos e cognitivos, especialmente no que diz respeito à produção e interpretação de textos. Para investigar esse tensionamento, foi realizada uma pesquisa exploratória por meio de um levantamento sistemático da literatura na base de dados de periódicos Web of Science (WOS), buscando responder à seguinte questão: Como a produção científica internacional tem mapeado as reconfigurações

da função-autor e os tensionamentos nas práticas discursivas diante da adoção de modelos generativos?

Para responder a essa questão central, o presente artigo tem como objetivo geral mapear e analisar a produção científica internacional no período de 2020 a 2026 acerca dos tensionamentos provocados pelas Inteligências Artificiais Generativas nas propriedades do discurso e na reconfiguração da função-autor. A partir deste propósito primário, delineiam-se os seguintes objetivos específicos: (i) identificar os principais polos de pesquisa, periódicos e pesquisadores que lideram as discussões na intersecção entre modelos generativos e textualidade; (ii) examinar as redes de coocorrência de palavras-chave e co-citação para rastrear como categorias discursivas (a exemplo de cena enunciativa e ethos) têm se articulado com os limites computacionais dessas ferramentas na literatura; e (iii) investigar de que modo os estudos recentes fundamentam o deslocamento da agência humana em direção a modelos de autoria distribuída; refletindo sobre as implicações práticas dessa transição para os processos de ensino-aprendizagem e produção acadêmica.

Para alcançar os objetivos delineados, o percurso metodológico adotado caracteriza-se por um levantamento sistemático da literatura ancorado na base de dados *Web of Science* (WoS). A extração do corpus analítico foi orientada pela seguinte equação de busca: **ALL=((“large language model*” OR “generative AI” OR ChatGPT) AND (“discourse analysis” OR “author function” OR ethos OR enunciation OR “authorial voice” OR “authorial identity” OR stance))**. O conjunto de metadados recuperado foi submetido a um processamento quantitativo fundamentado nas três leis empíricas clássicas da bibliometria. Especificamente,

aplicou-se a Lei de Bradford para mapear a dispersão do conhecimento e identificar os periódicos nucleares do campo; a Lei de Lotka para avaliar a produtividade científica e os padrões de consolidação dos pesquisadores que investigam essa intersecção; e a Lei de Zipf para analisar a frequência de palavras-chave e termos em resumos (Grácio et al; 2020); A articulação conjunta dessas métricas estruturará a análise das redes de coocorrência; permitindo revelar não apenas a base intelectual da área; mas também as frentes de pesquisa emergentes que tensionam a configuração da autoria diante das tecnologias generativas.

Com esta pesquisa, pretende-se não só esclarecer como a ciência tem interpretado os tensionamentos discursivos e a reconfiguração da função-autor diante das tecnologias generativas, mas também apresentar um diagnóstico sistematizado do estado da arte sobre o assunto. Ao cartografar a literatura internacional e evidenciar as frentes teóricas em ascensão, busca-se disponibilizar fundamentos acadêmicos consistentes que apoiem a comunidade educacional na leitura das ambiguidades da produção textual contemporânea. Nesse sentido, o estudo almeja abrir caminho para a reflexão sobre o uso ético e integrado dessas ferramentas no ensino, contribuindo para a elaboração de estratégias didáticas capazes de distinguir a simples reprodução algorítmica de uma escrita genuína, com ênfase na promoção de uma autoria crítica e na recuperação da agência humana no processo de construção de sentidos.

2. FUNDAMENTAÇÃO TEÓRICA OU REVISÃO DA LITERATURA

Para dar sustentação teórica a esta pesquisa, considerando seu caráter interdisciplinar, esta seção foi organizada em duas partes, cada uma voltada a distintos campos do saber. Na primeira

subseção, são apresentados os conceitos centrais da textualidade, com ênfase na análise de corpora discursivos e na noção de função autoral. A segunda subseção aborda os fundamentos da bibliometria, contemplando a definição e o emprego das Leis de Bradford, Lotka e Zipf, conforme discutido por Spinak (1996), já mencionadas na Introdução. Por fim, procura-se articular esses eixos, evidenciando de que modo a perspectiva bibliométrica pode contribuir para examinar as relações da escrita autoral com IAGs e como a produção acadêmica tem estudado esse tema.

2.1. Análise do Discurso e Função Autoral

A compreensão da autoria e do texto constitui um desafio epistemológico que transcende a figura empírica do indivíduo que escreve. Para Foucault (1992), a figura do autor não deve ser vista apenas como o indivíduo que cria uma obra, mas como uma função discursiva dentro de um sistema de produção e circulação de discursos. A “função autor” não é uma identidade natural ou universal; é, antes, uma construção histórica que estabelece um conjunto de regulações e expectativas, determinando a maneira como um texto é recebido, interpretado e validado em uma determinada cultura.

Ao aplicar essa perspectiva aos ecossistemas digitais atuais, torna-se necessário ampliar a própria noção de texto. O texto configura-se como uma relação cultural complexa que envolve o autor, o leitor e o contexto tecnológico e social em que a comunicação ocorre, requerendo metodologias como as propostas por Paveau (2019) para a análise dos tecnodiscursos. Quando a produção textual é mediada por Inteligências Artificiais Generativas (IAGs), a materialidade

discursiva e a intencionalidade enunciativa passam por fricções significativas.

Para aprofundar a dinâmica dessa interação, a abordagem teórica de Maingueneau (2024) é fundamental, especialmente pela análise da cena de enunciação e pelos conceitos de ethos, pathos, arqui-enunciador e meta-enunciador. No contexto algorítmico, o Modelo de Linguagem de Larga Escala (LLM) atua de forma eficaz na construção do pathos, assumindo a função de arqui-enunciador ao operar como um reproduutor marcadamente polifônico das vozes encontradas em seus dados de treino.

No entanto, apesar dessa hipercompetência retórica, a máquina não dispõe de “vivência enunciativa”. Essa falta inerente a impede de ocupar o lugar de meta-enunciador, isto é, de adotar uma distância e formular um posicionamento crítico em relação ao próprio discurso. Tal impossibilidade de projetar um ethos genuíno se manifesta na limitação semântica dos textos produzidos. Em consequência, a incorporação dessas ferramentas reconfigura a agência do sujeito humano, deslocando-o da posição de produtor soberano e originário de sentido para a de um “diretor de cena” inserido em um processo de autoria distribuída.

O presente estudo bibliométrico, portanto, foca em como a academia tem observado esta interferência enunciativa, mapeando os temas (frequência de palavras) que emergem dessas análises.

2.2. Conceitos de Bibliometria

Conforme salientado por Grácio et al. (2020), a bibliometria consolidou-se como uma área de estudo fundamental para compreender a dinâmica e avaliar a comunicação científica. O autor

indica que esta disciplina, por meio de métodos quantitativos, subsidia a tomada de decisão e contribui para a qualificação da pesquisa ao produzir material de referência bibliográfica de elevado padrão. Dessa forma, o emprego da bibliometria na construção do corpus deste trabalho oferece um procedimento objetivo para identificar lacunas informacionais, tendências de publicação e a atualização de autores e termos de indexação no contexto das Inteligências Artificiais Generativas.

A operacionalização desta análise fundamenta-se em três leis essenciais: a Lei de Bradford, relacionada à dispersão e produtividade dos periódicos; a Lei de Lotka, voltada à produtividade dos autores; e a Lei de Zipf, que analisa a frequência de ocorrência das palavras (Ferreira, 2010). Como este estudo pretende consolidar uma base para a consulta acadêmica, torna-se fundamental reconhecer as fontes de maior impacto. Nesse cenário, a Lei de Bradford desempenha um papel central. Conhecida como Lei de Dispersão, é o instrumento utilizado para mensurar a produtividade de periódicos e reconhecer tanto o núcleo quanto as zonas de dispersão de um tema específico. Segundo essa lei, os estudos iniciais sobre um assunto emergente tendem a se concentrar em um número restrito de revistas especializadas. Caso o campo continue em desenvolvimento, consolida-se um núcleo de revistas altamente produtivas, enquanto outros periódicos periféricos dão origem a novas áreas de dispersão (Lousada et al., 2012).

Além de mapear as fontes de referência, a compreensão do avanço científico exige a análise da relevância e da consolidação dos pesquisadores que publicam sobre o tema. Para essa finalidade, recorre-se à Lei de Lotka, ou Lei do Quadrado Inverso. Este princípio explica como a produtividade autoral se distribui com base em um

modelo de distribuição de tamanho-frequência. Investigada em 1926, a lei afirma que a quantidade de pesquisadores que realiza n contribuições numa área, corresponde aproximadamente, a $1/n^2$ do total daqueles que produzem apenas um trabalho, sendo que a parcela de autores com publicação única aproxima-se geralmente de 60% (Melo Ribeiro, 2017). A aplicação desta lei permitirá identificar os autores que lideram e estabilizam as discussões teóricas na área.

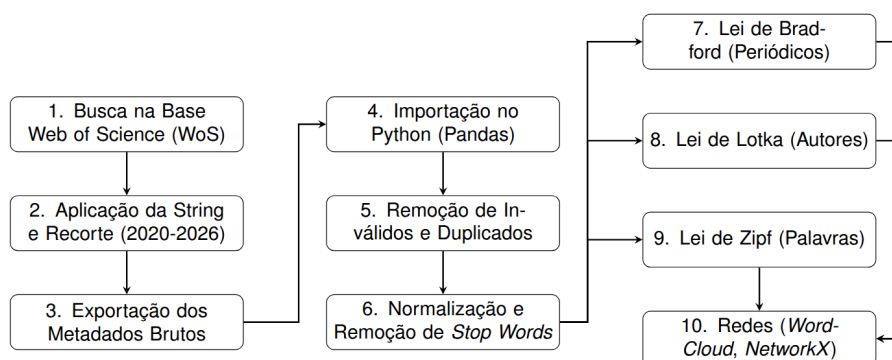
Por fim, para que este mapeamento seja mais útil do que a catalogação de autores e revistas e consiga adentrar na estrutura do discurso científico, aplica-se a Lei de Zipf. Proposta em 1949 e conhecida como “lei do menor esforço”, a Lei de Zipf investiga a frequência com que as palavras ocorrem em um texto ou conjunto documental, ordenando-as em um ranking de ocorrências. Ampliada posteriormente para contemplar o compartilhamento de posições por palavras de baixa frequência (Grácio et al., 2020), a Lei de Zipf revela-se indispensável para este estudo. Será por meio de sua aplicação que as palavras-chave mais recorrentes e as redes de coocorrência temáticas serão isoladas; permitindo rastrear o vocabulário e a formação dos campos semânticos que estruturam o debate contemporâneo.

3. METODOLOGIA

A presente pesquisa configura-se, conforme as tipologias metodológicas apresentadas por Gil (2002) e Medeiros et al. (2010), como um estudo bibliométrico de caráter exploratório e descritivo, amparado por uma abordagem quantitativa. Além disso, a investigação estrutura-se a partir de um levantamento longitudinal que abrange o período de 2020 a 2026, visando acompanhar o desenvolvimento e a evolução temporal desse campo. Esse

procedimento busca mapear o estado da arte e organizar a produção científica internacional acerca dos impactos das Inteligências Artificiais Generativas na autoria e nas propriedades do discurso. Para garantir o rigor e a reprodutibilidade do estudo, o delineamento metodológico foi organizado em três etapas centrais: coleta de dados, processamento e higienização, e análise de redes bibliométricas. Todo esse trajeto metodológico está descrito na Figura 1.

Figura1. Fluxograma das etapas metodológicas da pesquisa bibliométrica.



Fonte: Autoria própria

3.1. Coleta de Dados e Estratégia de Busca

A seleção do corpus ocorreu por meio de um levantamento sistemático na base de dados *Web of Science* (WoS), utilizando a equação de busca descrita na Tabela 1, com a exportação dos metadados para um arquivo de planilha eletrônica. Após a importação inicial, o conjunto de dados foi restrito exclusivamente aos artigos publicados até o ano de 2026, realizando-se a exclusão das publicações datadas de 2027.

Parâmetro / Etapa	Quantitativo
Base de Dados: Web of Science (WoS)	–
Recorte Temporal Inicial: 2020–2026	–
Equação de Busca: ALL=(("large language model*" OR "generative AI" OR ChatGPT) AND ("discourse analysis" OR "author function" OR ethos OR enunciation OR "authorial voice" OR "authorial identity" OR stance))	–
Total de publicações brutas recuperadas	807
Total após exclusão de registros futuros (> 2026)	802
Total de publicações consolidadas após higienização (títulos únicos, com resumo e palavras-chave)	667

Tabela 1. Parâmetros de busca e quantitativo de publicações na *Web of Science*

3.2. Processamento e Higienização dos Dados

A fase de tratamento de dados foi executada inteiramente em ambiente computacional por meio da linguagem *Python*, empregando a biblioteca *Pandas*. Aplicaram-se filtros sistemáticos para excluir artigos inválidos, removendo, em sequência, os registros que apresentavam ausência de título, palavras-chave e resumo (*.abstract*). Na etapa seguinte, todos os títulos foram convertidos para minúsculas a fim de garantir a verificação e a remoção precisa de documentos duplicados. Para viabilizar a análise semântica estrutural, procedeu-se à normalização dos textos extraídos dos resumos, títulos e palavras-chave; este procedimento englobou rotinas em *.loop* para a supressão de pontuações, parênteses e caracteres especiais, além da filtragem de uma lista de *stop words* (como conjunções, artigos e preposições da língua inglesa).

3.3. Análise Bibliométrica e Visualização de Redes

A extração das métricas bibliométricas foi conduzida por meio de procedimentos estatísticos, com suporte das bibliotecas *Numpy* e *Scipy*. A Lei de Bradford foi aplicada a partir do fracionamento do conjunto total de artigos em três partes de tamanho equivalente, possibilitando identificar o núcleo inicial composto pelas fontes (periódicos) mais produtivas e representativas do campo (Lousada et

al., 2012). A verificação da Lei de Lotka ocorreu via teste estatístico com regressão linear sobre a distribuição de produtividade dos autores, relacionando o logaritmo na base 10 do número de publicações ao logaritmo do número de pesquisadores, de modo a estimar o coeficiente angular. De forma semelhante, a Lei de Zipf foi examinada por regressão linear, correlacionando o logaritmo do *rank* com o logaritmo da frequência de ocorrência das palavras consolidadas no corpus. Por fim, o mapeamento relacional do campo discursivo foi apresentado visualmente por meio da geração de nuvens de palavras (*WordCloud*) e de uma rede de coocorrência de palavras-chave; construída com a biblioteca *NetworkX*, que definiu arestas a partir de combinações pareadas de termos observados em um mesmo artigo.

Por fim, em consonância com as práticas de ciência aberta e visando garantir total transparência e reprodutibilidade dos procedimentos metodológicos adotados, todo o código-fonte desenvolvido para esta investigação está publicamente disponível. Os *scripts* em *Python*, que abrangem as rotinas de higienização dos metadados, a aplicação estatística das leis bibliométricas e a geração dos grafos relacionais, podem ser consultados e replicados por meio do repositório hospedado no *GitHub*, acessível no seguinte endereço eletrônico:

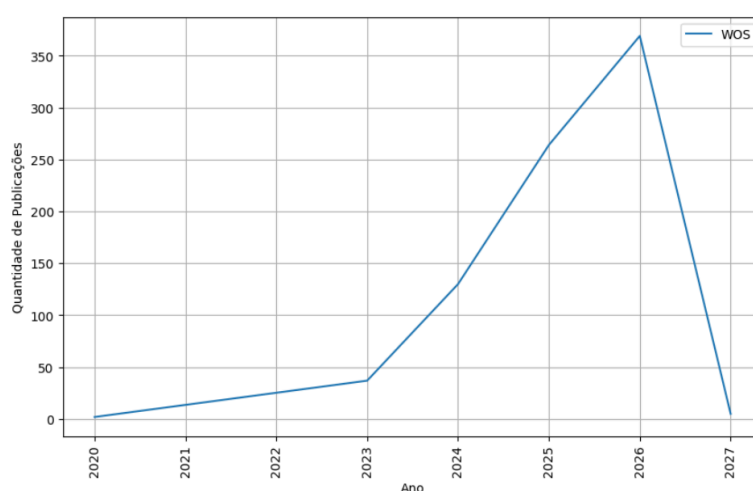
https://github.com/franciscofreitasneto/Bibliometria_Inteligencia_Artificial_Generativa_Discurso_Autoria.

4. ANÁLISE DOS RESULTADOS

A execução da estratégia de busca e a consequente exportação dos metadados da base *Web of Science* resultou em um conjunto bruto inicial de 807 publicações. Durante a análise exploratória preliminar,

focada na distribuição cronológica dos trabalhos, constatou-se uma curva de crescimento do interesse científico a partir de 2023. Contudo, o mapeamento identificou a presença de cinco artigos indexados com ano de publicação em 2027 (vide Figura 2). Considerando que a coleta de dados foi realizada em 2026, esses registros correspondem a publicações em acesso antecipado (*early access*) ou edições futuras. A manutenção desses dados causava uma distorção visual ao plotar o gráfico de tendências, sugerindo erroneamente uma queda abrupta na produção científica recente. Para assegurar a integridade do levantamento longitudinal e respeitar o recorte temporal metodológico (2020-2026), procedeu-se à eliminação desses registros futuros, reduzindo o corpus para 802 documentos. Após esta etapa, o processamento continuou com a eliminação de artigos repetidos e registros sem resumo ou palavras-chave; consolidando uma base final tratada de 667 publicações para as análises bibliométricas.

Figura 2. Quantidade de publicações por ano na base *Web of Science*

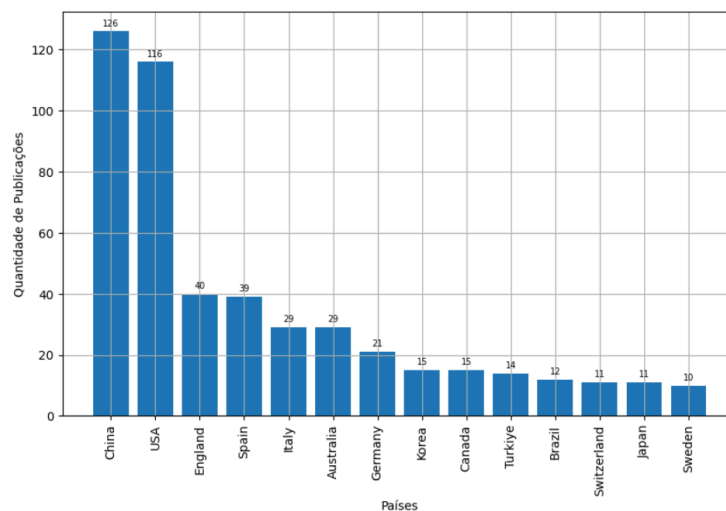


Fonte: Autoria própria

4.1. Panorama Geral da Produção Científica

A caracterização inicial do corpus bibliométrico, composto por 667 publicações consolidadas, revela uma concentração geográfica e disciplinar. O mapeamento das afiliações institucionais indica que a China e os Estados Unidos lideram a produção científica na intersecção entre Inteligência Artificial Generativa e análise discursiva/autoral, com 126 e 116 artigos publicados, respectivamente. Em um patamar intermediário de produtividade, destacam-se países europeus e a Austrália, sendo a Inglaterra a que mais publicou, com 40 registros, seguida pela Espanha com 39, pela Itália com 29 e pela Austrália, também com 29 publicações, conforme descrito na Figura 3.

Figura 3. Quantidade de Publicações por Localização (Mínimo de 10 Publicações)

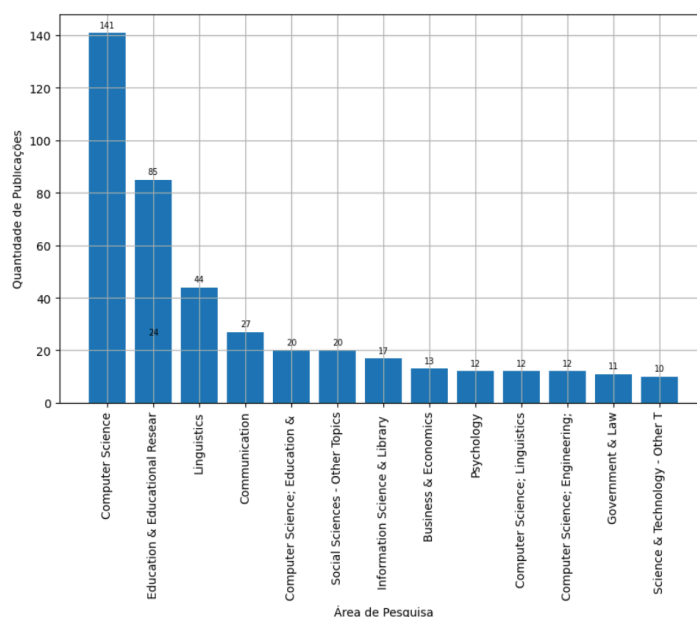


Fonte: Autoria própria

A análise das áreas de pesquisa (*Research Areas*) confirma a natureza interdisciplinar das investigações sobre o tema (Vide Figura 4). A área de Ciência da Computação predomina, com 141 publicações. Imediatamente a seguir, destacam-se a Educação e Pesquisa Educacional (*Education & Educational Research*) com 85 registros, e a Linguística (*Linguistics*) com 44 publicações. A Comunicação (*Communication*) surge em quarto lugar, com 27

trabalhos. Essa análise disciplinar reflete as dimensões abordadas na presente tese: a infraestrutura algorítmica computacional, o impacto direto nos ecossistemas de ensino-aprendizagem e as tensões estruturais nas textualidades e no discurso.

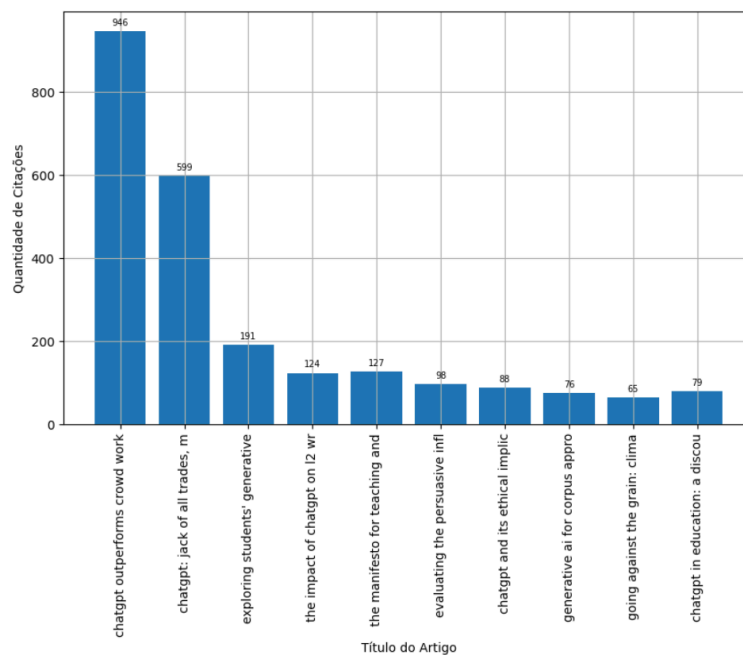
Figura 4. Quantidade de Publicações por Área de Pesquisa (Mínimo de 10 Publicações)



Fonte: Autoria própria

Quanto ao impacto e à recepção acadêmica, avaliados pelo volume de citações, a literatura de monografia mostra uma assimilação de trabalhos recentes. Destaca-se o artigo intitulado "*ChatGPT outperforms crowd workers for text-annotation tasks*", que acumula 946 citações globais (*All Databases*) e 804 no núcleo principal da WoS (*WoS Core*). O segundo trabalho mais referenciado, "*ChatGPT: Jack of all trades, master of none*", soma 599 citações globais. Outros estudos voltados para o impacto direto, como "*Exploring Students' Generative AI-Assisted Writing Processes: Perceptions and Experiences from Native and Nonnative English Speakers*" (191 citações) e "*The Impact of ChatGPT on L2 Writing Skill Development and Learners' Perceptions*" (124 citações), também figuram no topo da base intelectual do campo (Vide Figura 5).

Figura 5. Quantidade de Citações dos Artigos



Fonte: Autoria própria

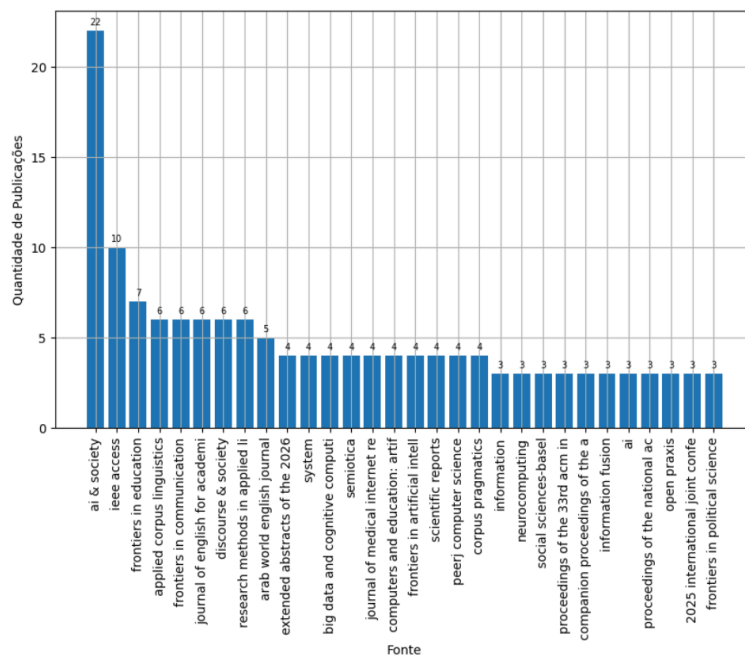
De modo geral, o panorama revela um campo de pesquisa em expansão, impulsionado predominantemente por potências tecnológicas mundiais e abrangendo áreas que vão desde a engenharia de algoritmos até as ciências da linguagem e a educação. Com este cenário macroestrutural estabelecido, faz-se necessário um aprofundamento nas dinâmicas de disseminação e nas estruturas internas dessa produção. Para tanto, as subseções seguintes aplicam as leis bibliométricas clássicas para desvelar o comportamento de concentração dos periódicos difusores, a produtividade dos autores e a rede semântica que fundamenta o atual discurso científico.

4.2. Lei de Bradford: Dispersão e Núcleo de Periódicos

A distribuição documental entre os veículos de divulgação indica que o periódico *AI & Society* é a principal fonte difusora da temática, liderando o *ranking* absoluto com 22 artigos. Na sequência, periódicos técnicos e educacionais dominam o topo da lista, com o

IEEE Access contabilizando 10 publicações, seguido por *Frontiers in Education* (7 publicações) e *Applied Corpus Linguistics* (6 publicações). O comportamento quantitativo dessa distribuição evidencia uma dinâmica clássica de comunicação científica, caracterizada por um pequeno núcleo de periódicos altamente produtivos e uma queda contínua de fontes dispersas, conforme apresentado na Figura 6.

Figura 6. Quantidade de Publicações por Periódicos: 30 maiores fontes de publicações



Fonte: Autoria própria

Para a aplicação formal da Lei de Bradford, o volume total de 667 publicações, provenientes de 481 periódicos distintos, foi fracionado em três zonas de dispersão, com o objetivo de agrupar aproximadamente um terço da produção em cada estrato (Lousada et al., 2012; Brookes, 1969; Vanti, 2002; Spinak, 1996). A análise estatística revelou a formação de um núcleo de Bradford, composto por apenas 40 periódicos de impacto, que reúnem, juntos, 177 artigos relevantes sobre o tema. O comportamento cumulativo que delimita este núcleo central está detalhado na Tabela 2.

Tabela 2. Distribuição do Núcleo de Bradford das fontes de publicação

	Qtd.Publicações	Qtd.Fontes	Total_Fontes	Total_Publicacoes	Log_Total_Fontes
0	22	1	1	22	0.000000
1	10	1	2	32	0.301030
2	7	1	3	39	0.477121
3	6	5	8	69	0.903090
4	5	1	9	74	0.954243
5	4	10	19	114	1.278754
6	3	21	40	177	1.602060

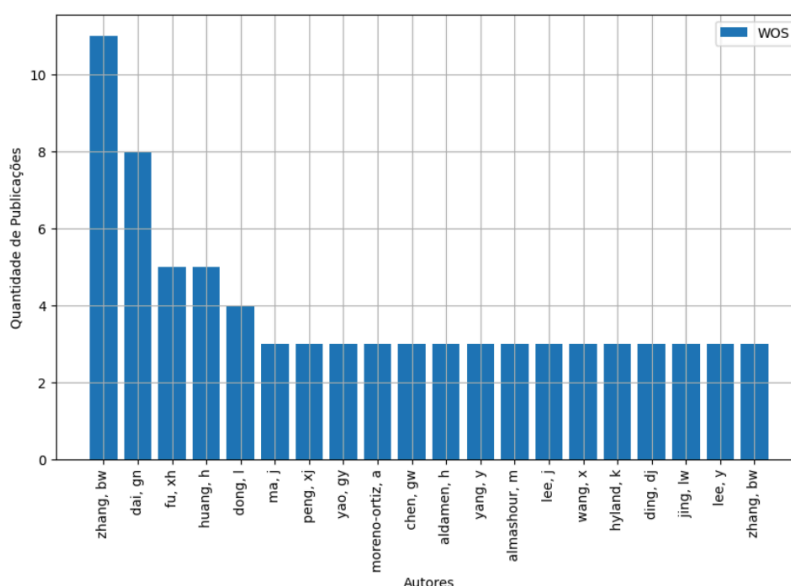
A partir desses valores, calculou-se a taxa de Bradford (razão entre o total de periódicos da amostra e o número de periódicos do núcleo), que resultou em 12,025. Esse valor indica que, para cada periódico pertencente ao núcleo de maior relevância, existem aproximadamente 12 outros periódicos de menor produtividade dispersos nas zonas periféricas, evidenciando que a literatura sobre a interseção entre modelos de linguagem, análise do discurso e autoria ainda carece de um conjunto consolidado de veículos de publicação dedicados ao tema, encontrando-se, em vez disso, pulverizada por um número expressivo de fontes ocasionais. Esse dado é coerente com o caráter interdisciplinar e emergente do objeto investigado nesta pesquisa, reforçando a necessidade de uma revisão da literatura que não se limite ao núcleo de periódicos mais produtivos, sob o risco de excluir contribuições relevantes publicadas isoladamente em veículos específicos.

4.3. Lei de Lotka: Produtividade e Consolidação Autoral

A análise da autoria, fundamental para compreender os polos de consolidação teórica do campo, mapeou um total de 2.282 participações autorais na base de dados higienizada. O ranqueamento absoluto revelou a presença de pesquisadores produtivos na vanguarda das discussões, destacando Zhang, B. W., com 11 publicações, e Dai, G. N., com 8 trabalhos, seguidos por Fu, X. H. e Huang, H., ambos com 5 contribuições acadêmicas (Vide Figura

7). Esta concentração inicial sugere um padrão de forte desigualdade na produção científica, o qual foi submetido à validação estatística por meio da Lei de Lotka.

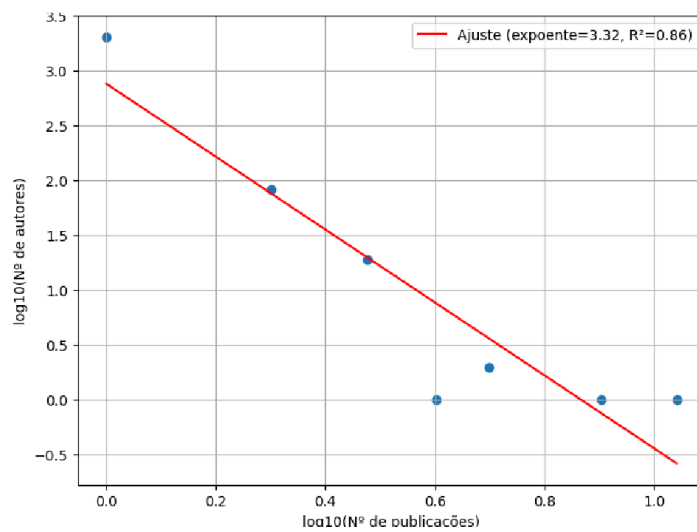
Figura 7. Quantidade de Publicações por Autor: 20 autores mais relevantes



Fonte: Autoria própria.

Para aferir a conformidade dessa distribuição de produtividade com o modelo teórico clássico, conhecido como Lei do Quadrado Inverso, foi aplicada uma regressão linear (Egghe; Rousseau, 1990; Urbizagastegui Alvarado, 2008). O cálculo correlacionou o logaritmo da quantidade de publicações ao logaritmo do número correspondente de autores. O modelo estatístico gerado apresentou um coeficiente de determinação (R^2) de 0,859. Esse alto valor de R^2 indica um grau satisfatório de ajuste dos dados empíricos à curva teórica, confirmando que a literatura sobre Inteligência Artificial Generativa e discurso segue rigorosamente uma distribuição de lei de potência, conforme apresentado na Figura 8⁷.

Figura 8. Verificação da Lei de Lotka



Fonte: Autoria própria

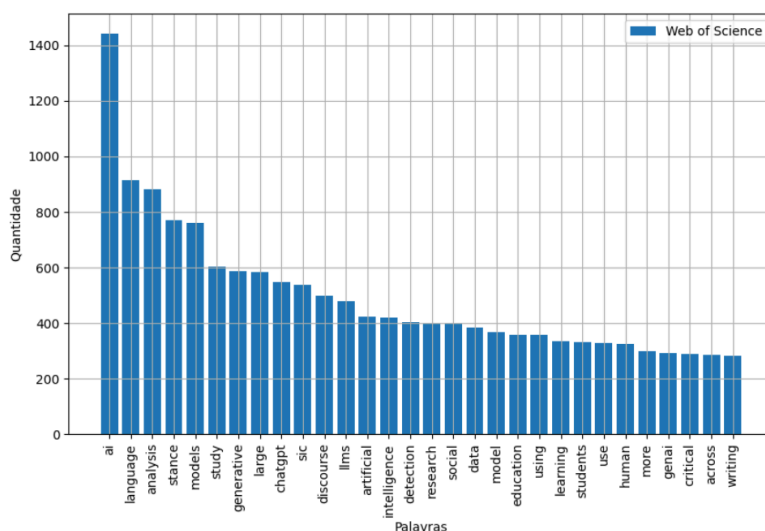
O coeficiente angular (expoente de Lotka) obtido na regressão foi de 3,321. Embora esteja acima do parâmetro clássico de Lotka (como citado por Lotka (1926) e Coile (1977) deve ser aproximadamente 2), esse valor elevado evidencia um padrão de forte assimetria na área: a grande maioria dos pesquisadores participou de modo pontual, com apenas uma publicação, possivelmente motivada pelo caráter recente e emergente do tema. Em contrapartida, um grupo reduzido e altamente produtivo sustenta e consolida os fundamentos metodológicos e epistemológicos do campo. A partir do referencial adotado, observa-se que a configuração da “função-autor” de liderança nesse debate acadêmico continua restrita a um núcleo seletivo de investigadores.

4.4. Lei de Zipf e Estruturação do Campo Semântico

Para realizar o estudo discursivo da literatura produzida, aplicou-se a Lei de Zipf, que quantificou a frequência de ocorrência do vocabulário consolidado a partir dos títulos, resumos e palavras-chave das publicações; A agregação desses três campos resultou em um léxico central que estrutura o debate; destacando os termos *ai* (1.441 ocorrências), *language* (913), *analysis* (882), *stance* (770) e *models* (762). A incidência do termo *stance*

(posicionamento/tomada de posição) entre os cinco termos mais frequentes é relevante para os objetivos desta pesquisa, pois atesta a centralidade das categorias enunciativas e das investigações sobre o ethos discursivo neste corpus sobre inteligência artificial (Vide Figura 9).

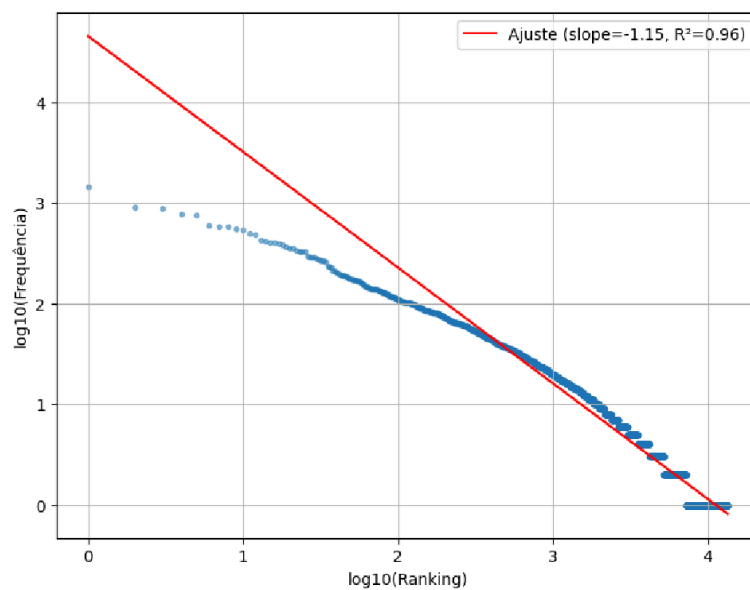
Figura 9. Quantidade de palavras: Títulos + Resumos + Palavras-Chave: As 50 palavras mais frequentes.



Fonte: Autoria própria

A aderência empírica deste vocabulário à “lei do menor esforço” foi testada estatisticamente por meio de uma regressão linear, correlacionando o logaritmo da ordem da série (*rank*) com o logaritmo da frequência total das palavras (Zipf, 1949). O modelo estatístico obteve um coeficiente de determinação (R^2) de 0,960, e o coeficiente angular da reta (expoente de Zipf) foi estimado em -1,150, um valor muito próximo ao previsto teoricamente (-1) para distribuições que seguem a Lei de Zipf (Vide Figura 10) (Mandelbrot, 1953; Booth, 1967).

Figura 10. Verificação da Lei de Zipf



Fonte: Autoria própria

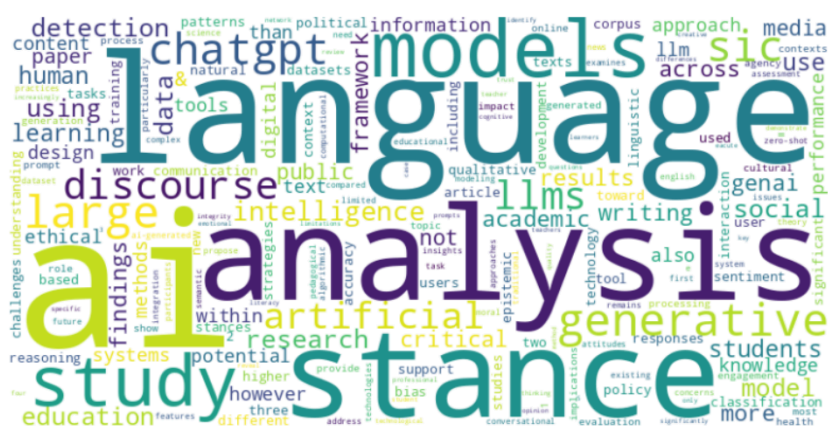
Esse resultado atesta que o corpus constitui uma amostra válida e representativa de linguagem natural, corroborando a adequação metodológica da coleta e do tratamento de dados realizados para as análises subsequentes. Confirma-se, assim, o padrão característico esperado por essa lei: um núcleo reduzido de termos de alta frequência (*ai, language, analysis, stance, models*) concentra grande parte das ocorrências, enquanto uma extensa cauda de termos de baixa frequência fornece as especificidades temáticas e contextuais do campo.

A estruturação subsequente da rede de coocorrência de palavras-chave extrai exatamente as correlações entre esses nós de alta frequência; O grafo relacional gerado (com restrição de grau mínimo de 5 coocorrências) evidencia a convergência disciplinar; interligando diretamente componentes da infraestrutura técnica (como *large language models* e *chatgpt*) aos pilares da teoria da linguagem, tais como a análise do discurso e a análise crítica do discurso. Essa teia semântica sugere, de forma exploratória, uma tendência da produção científica recente em aproximar a discussão sobre a natureza generativa dos algoritmos às propriedades tradicionalmente atribuídas à função autoral.

4.5. Visualização Semântica e Redes de Coocorrência

A visualização das métricas de frequência se apresenta por meio da projeção da nuvem de palavras (*WordCloud*) e do grafo relacional. A nuvem de palavras, elaborada a partir da totalização dos vocábulos extraídos de títulos, resumos e palavras-chave; ilustra o conjunto terminológico que caracteriza este domínio de pesquisa; Em destaque na imagem; observa-se que termos relacionados à engenharia computacional e à tecnologia (tais como *models*, *ai*, *generative*, *chatgpt*) fundem-se com as categorias essenciais das ciências da linguagem (como *language*, *analysis*, *stance*, *discourse*) (Vide Figura 11).

Figura 11. Nuvem de palavras construída a partir da frequência absoluta dos termos.

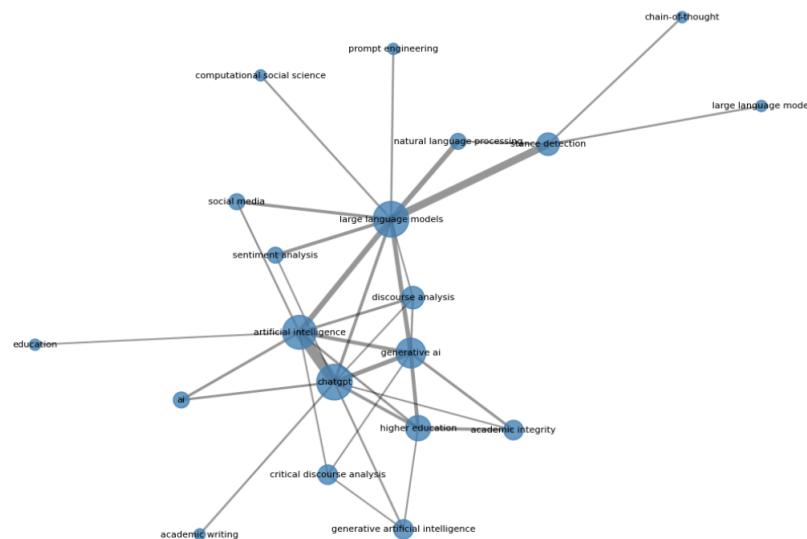


Fonte: Autoria própria

Além da contagem isolada, a análise topológica das palavras-chave dos autores permitiu mapear a forma como esses conceitos se articulam dentro das publicações; Utilizando a biblioteca *NetworkX*, foi construída uma rede de coocorrência parametrizada, com um limite mínimo de 5 interações conjuntas por artigo. O grafo resultante é composto por 20 nós (*nodes*) e 36 arestas (*edges*),

revelando a arquitetura estrutural da literatura de fronteira (vide Figura 12).

Figura 12. Rede de coocorrência de palavras-chave (grau mínimo de 5 interações).



Fonte: Autoria própria

A análise cuidadosa das conexões do grafo revela que termos tecnológicos de grande alcance (como *large language models*, *chatgpt* e *artificial intelligence*) funcionam como importantes *hubs* de agregação. A partir desses núcleos centrais, partem vínculos expressivos em direção a abordagens de exame textual, destacadas pelos nós *discourse analysis* e *critical discourse analysis*. Além disso, observa-se elevada densidade de arestas associando esses algoritmos a campos de aplicação educacional e a preocupações éticas, representados pelos nós *higher education*, *academic writing* e *academic integrity*.

Esta configuração de rede corrobora a hipótese basilar da pesquisa: a comunidade científica internacional está mobilizando ferramentas da análise do discurso não apenas para investigar a materialidade da IA, mas, primordialmente, para avaliar os tensionamentos que a

escrita mediada por algoritmos exerce sobre a autoria, a originalidade e a integridade nos ambientes de ensino superior.

5. CONCLUSÃO

O presente estudo foi motivado pelas inquietações percebidas no contexto educacional diante da padronização e da redução da densidade autoral em produções mediadas por Inteligências Artificiais Generativas (IAGs), o que exigiu uma investigação sistemática sobre como a literatura internacional tem compreendido esse fenômeno. No decorrer do artigo, delineou-se um percurso analítico que integrou a fundamentação discursiva (sustentada na função-autor foucaultiana, nos tecnodiscursos de Paveau e na cenografia enunciativa de Maingueneau) ao rigor quantitativo da bibliometria. A partir da análise longitudinal de 667 publicações indexadas na base *Web of Science* entre 2020 e 2026, o trabalho alcançou integralmente seu objetivo geral de mapear e examinar a produção científica internacional sobre os tensionamentos gerados pelas IAGs nas propriedades do discurso e na reconfiguração da autoria.

Quanto aos objetivos específicos, cada etapa do delineamento metodológico proporcionou respostas fundamentadas empiricamente. Em relação ao primeiro objetivo específico, identificar os principais polos de pesquisa, periódicos e pesquisadores que lideram as discussões, os resultados evidenciaram uma liderança geográfica compartilhada entre a China (126 artigos) e os Estados Unidos (116 artigos), seguidos por países europeus e pela Austrália.

A aplicação da Lei de Bradford (Grácio, 2020) demonstrou a existência de um núcleo restrito formado por 40 periódicos, destacando-se de maneira significativa o *AI & Society*, que apresenta 22 publicações, sendo responsável pela veiculação de 177 artigos na área.) A Taxa de Bradford calculada em 12,025 atestou a elevada dispersão periférica da literatura, reflexo direto do caráter interdisciplinar e emergente do tema. Paralelamente, a aplicação da Lei de Lotka (Lotka, 1926) revelou uma acentuada assimetria na produtividade autoral ($R^2 = 0,859$; expoente angular de 3,321), confirmando que, embora autores como Zhang e Dai se destaquem na vanguarda da produtividade, a maior parte dos pesquisadores apresenta uma atuação transitória e pontual sobre o tema.

Quanto ao segundo objetivo específico, examinar as redes de coocorrência e vocabulário para rastrear a articulação entre categorias discursivas e limites computacionais, a verificação da Lei de Zipf (Zipf, 1949) confirmou a consistência estrutural do corpus linguístico, com um expressivo ajuste empírico ($R^2 = 0,960$; expoente de -1,150). A presença do termo *stance* (posicionamento enunciativo) entre as cinco palavras mais frequentes, ao lado de *ai*, *language*, *analysis* e *models*, comprova que a investigação acadêmica não se restringe à mecânica do processamento de linguagem natural, mas investiga diretamente as marcas de subjetividade e a projeção do *ethos* discursivo. A rede de coocorrência de palavras-chave; composta por 20 nós e 36 arestas; ratificou essa convergência ao posicionar ferramentas como o *ChatGPT* e modelos de linguagem como nós agregadores, diretamente vinculados às abordagens de análise do discurso e à análise crítica do discurso.

Por fim, no que diz respeito ao terceiro objetivo específico, investigar como as pesquisas fundamentam o deslocamento da agência humana e refletem sobre a autoria distribuída no ensino e na aprendizagem, a topologia relacional revelou adensamentos expressivos em torno dos nós <*higher education*>, <*academic writing*> e <*academic integrity*>. Os achados sugerem que a literatura internacional já reconhece esse deslocamento, fornecendo um terreno empírico para o aprofundamento teórico proposto nesta tese, que interpreta tal deslocamento à luz dos conceitos de arqui-enunciador e meta-enunciador de Maingueneau.

Em suma, este estudo conclui que a produção científica internacional reconhece a emergência de modelos de autoria compartilhada e distribuída, nos quais as tensões discursivas não anulam a função-autor, mas exigem sua redefinição epistemológica e pedagógica. Diante disso, desmistificar tanto o determinismo tecnológico quanto o pânico moral mostra-se indispensável para que as instituições educacionais desenvolvam práticas formativas orientadas à recuperação da agência humana e ao exercício de uma autoria crítica em contextos saturados por tecnologias generativas.

REFERÊNCIAS BIBLIOGRÁFICAS

BOOTH, A. D. A 'law' of occurrences for words of low frequency. **Information and Control**, Elsevier, v. 10, n. 4, p. 386–393, 1967. DOI: 10.1016/S0019-9958(67)90374-2.

BROOKES, B. C. Bradford's law and the bibliography of science. **Nature**, Nature Publishing Group, v. 224, n. 5223, p. 953–956, 1969.

COILE, R. C. Lotka's frequency distribution of scientific productivity. **Journal of the American Society for Information Science**, Wiley

Online Library, v. 28, n. 6, p. 366–370, 1977. DOI: 10.1002/asi.4630280610.

EGGHE, L.; ROUSSEAU, R. **Introduction to informetrics: quantitative methods in library, documentation and information science**. Amsterdam: Elsevier Science Publishers, 1990.

FERREIRA, A. G. C. Bibliometria na avaliação de periódicos científicos. **DataGramaZero-Revista de Ciência da Informação**, v. 11, n. 3, p. 1–9, 2010.

FOUCAULT, M. **O que é um autor?** Lisboa: Vega, 1992.

GIL, A. C. **Como elaborar projetos de pesquisa**. São Paulo: Atlas, 2002. v. 4.

GRÁCIO, M. C. C. **Análises relacionais de citação para a identificação de domínios científicos: uma aplicação no campo dos Estudos Métricos da Informação no Brasil**. [S. l.]: Editora UNESP, 2020.

GRÁCIO, M. C. C.; MARTÍNEZ-ÁVILA, D.; OLIVEIRA, E. F. T. d.; ROSAS, F. S. **Tópicos da bibliometria para bibliotecas universitárias**. [S. l.]: Editora UNESP, 2020.

LOTKA, A. J. The frequency distribution of scientific productivity. **Journal of the Washington Academy of Sciences**, v. 16, n. 12, p. 317–323, 1926.

LOUSADA, M.; GARCIA, C. L. S.; WOIDA, L. M.; DAL'EVEDOVE, P.; GARCIA, R.; VALENTIM, M. L. P. Produção científica sobre gestão do conhecimento e gestão da informação no âmbito da ciência da

informação: uma aplicação da Lei de Bradford. In: FACULTAD DE COMUNICACIÓN Y DOCUMENTACIÓN Y SERVICIO DE PUBLICACIONES DE LA . . ., 2. **ANALES de documentación**. [S. l.: s. n.], 2012. v. 15.

MAINGUENEAU, D. **Análise de textos de comunicação**. [S. l.]: Cortez Editora, 2024.

MANDELBROT, B. An informational theory of the statistical structure of language. In: JACKSON, W. (ed.). **Communication Theory**. London: Butterworths, 1953. p. 486–502.

MEDEIROS, C. H.; KAUARK, F. d. S.; MANHÃES, F. C. **Metodologia de Pesquisa: Um guia prático**. [S. l.]: Via Litterarum, Itabuna - BA, 2010. v. 1.

MELO RIBEIRO, H. C. Bibliometria: quinze anos de análise da produção acadêmica em periódicos brasileiros. **Biblios**, Los autores, n. 69, p. 1–20, 2017.

PAVEAU, M.-A. **Análise do Discurso Digital: Dicionário das formas e das práticas**. Campinas, SP: Pontes Editores, 2019. Tradução de Julia Lourenço Costa e Roberto Leiser Baronas.

SPINAK, E. **Diccionario enciclopédico de bibliometría, cienciometría e informetría**. Caracas: UNESCO, 1996.

URBIZAGASTEGUI ALVARADO, R. A produtividade dos autores na literatura de enfermagem: uma aplicação da Lei de Lotka. **Revista Brasileira de Enfermagem**, SciELO Brasil, v. 61, n. 4, p. 433–441, 2008.

VANTI, N. A. Da bibliometria à webometria: uma exploração conceitual dos mecanismos de representação e análise da informação na rede mundial de computadores. **Ciência da Informação**, v. 31, n. 2, p. 152–162, 2002.

ZIPF, G. K. **Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology**. Cambridge, MA: Addison-Wesley Press, 1949.

¹ Doutorando em Cognição e Linguagem – UENF

² Doutor em Comunicação - UFRJ

³ Pós-Doutor em Cognição e Linguagem – UENF

⁴ Doutor em Modelagem Computacional - UERJ

⁵ Mestre em Cognição e Linguagem - UENF

⁶ Doutorando em Cognição e Linguagem - UENF

⁷ Ressalta-se que a regressão da Lei de Lotka foi calculada sobre as 7 classes discretas de produtividade identificadas no corpus ($n=7$, referentes aos estratos de 1, 2, 3, 4, 5, 8 e 11 publicações). Aplicando-se a penalização pelos graus de liberdade decorrente do número reduzido de classes, obtém-se um coeficiente de determinação ajustado de $R^2_{adj} = 0,831$ (frente ao $R^2 = 0,859$), preservando um expressivo patamar de covariância explicada. Cumpre pontuar que, no objetivo buscado nesta descrição bibliométrica, a linearização em escala logarítmica cumpre uma função estritamente descritiva de

ajuste de curva voltada à estimação paramétrica do expoente estrutural α , não se caracterizando como modelagem inferencial ou preditiva. Por essa razão, métricas de erro de predição (como MSE, RMSE ou MAE) e diagnósticos complexos de resíduos não se aplicam ao escopo exploratório desta mensuração.