

# A LÍNGUA DA CIÊNCIA BRASILEIRA: uma análise de 12 milhões de publicações da Plataforma Lattes

**Giulia Magalhães Ferreira**

<https://orcid.org/0009-0004-7822-284X>  
[giulia.m@aluno.ufabc.edu.br](mailto:giulia.m@aluno.ufabc.edu.br)  
Universidade Federal do ABC (UFABC)

**Jesús P. Mena-Chalco**

<https://orcid.org/0000-0001-7509-5532>  
[jesus.mena@ufabc.edu.br](mailto:jesus.mena@ufabc.edu.br)  
Universidade Federal do ABC (UFABC)

## Resumo:

A predominância do inglês como língua franca da ciência tem sido amplamente documentada em bases bibliográficas internacionais. No entanto, ainda há pouca evidência sobre a distribuição linguística da produção registrada em bases curriculares nacionais. Este estudo analisa 12,1 milhões de publicações da Plataforma Lattes, identificando o idioma predominante a partir dos títulos. Os resultados mostram forte presença do português (73,17%), seguido pelo inglês (22,54%) e pelo espanhol (2,25%). Os achados sugerem que a produção científica brasileira registrada nos currículos mantém expressiva orientação nacional ou regional, enquanto o inglês aparece como componente relevante da internacionalização científica. A análise evidencia o potencial da Plataforma Lattes como fonte complementar para compreender dimensões linguísticas da circulação da ciência brasileira.

**Palavras-Chave:** Idioma. Internacionalização. Comunicação científica. Plataforma Lattes.

## EIXO TEMÁTICO:

Bases e Fontes de Dados

## MODALIDADE:

Resumo expandido

**DOI:** 10.22477/x.ebbc.1191

## COMO CITAR ESTE ARTIGO:

FERREIRA, Giulia Magalhães; MENA-CHALCO, Jesús P. A língua da ciência brasileira: uma análise de 12 milhões de publicações da Plataforma Lattes. In: EBBC - Encontro Brasileiro de Bibliometria e Cientometria. 10. **Anais [...]**. Curitiba, 2026. DOI: 10.22477/x.ebbc.1191. Disponível em: <https://doi.org/10.22477/x.ebbc.1191>.



## 1 INTRODUÇÃO

O conhecimento científico é produzido em diferentes regiões do mundo e reflete diversidades culturais, geográficas, institucionais e linguísticas. Embora o inglês ocupe atualmente a posição de língua franca da ciência, essa predominância é relativamente recente. Durante a Antiguidade, o grego e, posteriormente, o latim desempenharam papel central na circulação do conhecimento no Ocidente. A partir do Renascimento, o latim perdeu espaço para línguas nacionais europeias, como francês, alemão e italiano (Gordin, 2015). Após a Primeira Guerra Mundial e, sobretudo, após a Segunda Guerra, o inglês consolidou sua posição dominante na ciência internacional, impulsionado pela influência científica, econômica e política dos EUA (Montgomery, 2013). Atualmente, essa centralidade é particularmente evidente nas ciências naturais, nas quais o inglês predomina amplamente como idioma de publicação (Ammon, 2012).

Na bibliometria, a questão linguística também se relaciona aos limites das fontes utilizadas para representar a produção científica. Bases internacionais, como WoS e Scopus, são fundamentais para estudos de comunicação científica e avaliação da pesquisa, mas apresentam diferenças de cobertura e vieses disciplinares, regionais e linguísticos (Pranckutė, 2021). Estudos recentes também mostram que a cobertura de publicações em OpenAlex, Scopus e Web of Science varia conforme idioma, país, área do conhecimento e modelo de acesso aberto, com menor visibilidade relativa de periódicos voltados a comunidades nacionais ou regionais (Simard *et al.*, 2024).

A SciELO tem papel central na ampliação do acesso aberto dos periódicos latino-americanos (Packer *et al.*, 2014). O inglês como estratégia de internacionalização não produz efeitos homogêneos. Cintra, Silva e Furnival (2020), ao analisarem

periódicos da SciELO em Ciências Sociais Aplicadas, observaram crescimento da publicação em inglês, mas verificaram que, em parte expressiva dos periódicos, artigos em português ou espanhol tem maior média de citações no SciELO Citation Index.

Nesse contexto, a Plataforma Lattes (PL) constitui uma fonte relevante para observar dimensões da produção científica brasileira que podem permanecer menos visíveis em bases internacionais. Por reunir CVs de pesquisadores, professores e estudantes, a PL registra diferentes tipos de produção. Além disso, sistemas como o scriptLattes permitem extrair, organizar e deduplicar informações registradas nos CVs, possibilitando análises cientométricas (Mena-Chalco; Cesar Jr., 2009).

Apesar dessa abrangência, ainda são escassos estudos que investiguem, em larga escala, a distribuição linguística da produção brasileira. Assim, este trabalho pergunta: qual é a distribuição dos idiomas predominantes nos títulos das publicações registradas na PL e como essa distribuição varia segundo o tipo de produção científica ao longo do tempo? Diante disso, o objetivo deste estudo é analisar a distribuição de idiomas em mais de 12 milhões de registros bibliográficos deduplicados, tratando o idioma predominante dos títulos como um indicador parcial, mas relevante, dos diferentes circuitos de comunicação da ciência brasileira.

## 2 MATERIAL E MÉTODOS

Este estudo possui natureza quantitativa e exploratória, situando-se no campo da cientometria e da análise de produção científica.

### 2.1 COLETA E ESTRUTURAÇÃO DE DADOS

A base utilizada neste estudo foi construída a partir da coleta de Cvs Lattes realiza-

da em fevereiro de 2026. Ao todo, foram obtidos 9.549.171 CVs. Foram extraídas as produções declaradas considerando quatro tipos documentais: (i) artigos em periódicos, (ii) artigos publicados em eventos, (iii) livros e (iv) capítulos de livros.

A extração e a estruturação dos dados foram realizadas por meio de rotinas automatizadas em Python, voltadas à seleção dos campos bibliográficos. A PL foi adotada por constituir uma das mais abrangentes bases nacionais. Seu uso permite observar tipos de produção e circuitos de comunicação que podem não estar plenamente representados em bases bibliográficas internacionais, cuja cobertura é marcada por critérios seletivos de indexação e por vieses disciplinares, regionais e linguísticos (Lane, 2010; Packer *et al.*, 2014; Pranckutė, 2021; Simard *et al.*, 2024).

## 2.2 DEDUPLICAÇÃO DE PUBLICAÇÕES

Uma mesma publicação pode aparecer em múltiplos CVs devido à coautoria. Para evitar a contagem duplicada desses registros, foi aplicado um processo de deduplicação baseado no algoritmo do scriptLattes, proposta por Mena-Chalco e Cesar Jr. (2009). O procedimento identifica registros potencialmente equivalentes de uma mesma publicação e os consolida em um único registro, utilizando critérios baseados na similaridade textual entre títulos.

Esse processo permite transformar a base curricular em uma base de publicações únicas, mais adequada para análises agregadas da produção científica. Ainda assim, por se tratar de dados autodeclarados e heterogêneos, variações na grafia dos títulos, erros de preenchimento e inconsistências nos metadados podem produzir falsos positivos ou falsos negativos na deduplicação. Portanto, os resultados devem ser interpretados em escala agregada, considerando que o procedimento reduz a duplicidade, mas não elimina todos os ruídos da base. Após a deduplicação, a base final analisada contém 12.145.551 publicações, no período 2000 e

2025.

## 2.3 IDENTIFICAÇÃO AUTOMÁTICA DE IDIOMA

A identificação do idioma foi realizada a partir dos títulos das publicações registrados nos CVs. Portanto, o procedimento adotado identifica o idioma predominante do título, e não necessariamente o idioma integral do documento. Essa distinção é importante porque títulos científicos são textos curtos e podem conter siglas, nomes próprios, termos técnicos, expressões em língua estrangeira ou construções multilíngues, o que aumenta a ambiguidade da classificação.

Para reduzir esse problema, foi adotado um procedimento híbrido, combinando classificadores estatísticos e heurísticas complementares. Inicialmente, os títulos foram submetidos a classificadores automáticos de idioma, baseados em modelos de n-grams e em probabilidades de ocorrência de caracteres. Em seguida, foram aplicadas regras adicionais para tratar casos ambíguos, considerando padrões lexicais recorrentes em títulos científicos e características específicas de idiomas latinos. Essa etapa é importante em contextos multilíngues, nos quais português, espanhol, francês e italiano apresentam elevada similaridade lexical.

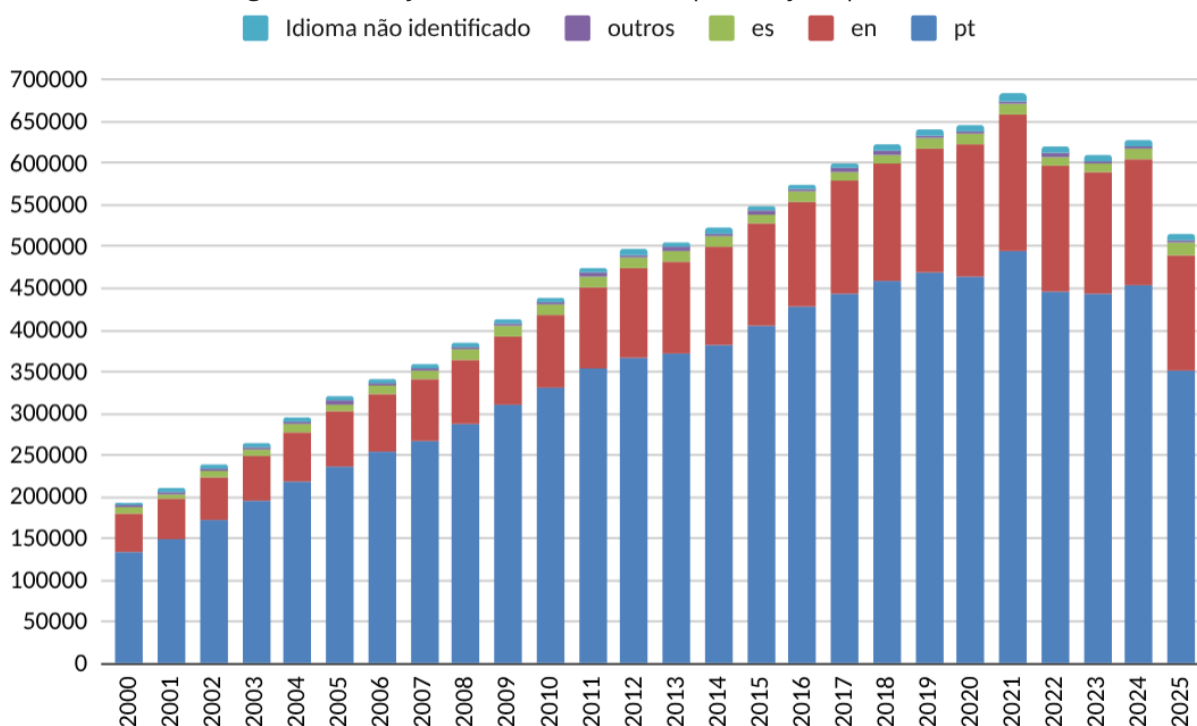
Os resultados dos classificadores foram integrados em um processo de decisão para atribuir a cada registro um idioma predominante. Quando não foi possível identificar o idioma com nível suficiente de confiança, o registro foi classificado como idioma não identificado. Ainda assim, o procedimento pode estar sujeito a imprecisões, especialmente em títulos muito curtos, com vocabulário técnico internacionalizado ou com mistura de idiomas. Os resultados devem ser interpretados em escala agregada, como uma aproximação da distribuição linguística da produção registrada na PL.

## 3 RESULTADOS E DISCUSSÃO

A análise revelou forte predominância do português nos títulos das publicações registradas na PL. Do total de 12.145.551 publicações, 73,17% apresentam títulos em português, enquanto 22,54% estão em inglês e 2,25% em espanhol. Outros idiomas representam menos de 1%. Esses resultados indicam que, quando observada a partir de uma base nacional, a produção brasileira apresenta uma configuração linguística distinta daquela usualmente capturada por bases internacionais, nas quais o inglês tende a predominar (Pranckutė, 2021; Simard *et al.*, 2024).

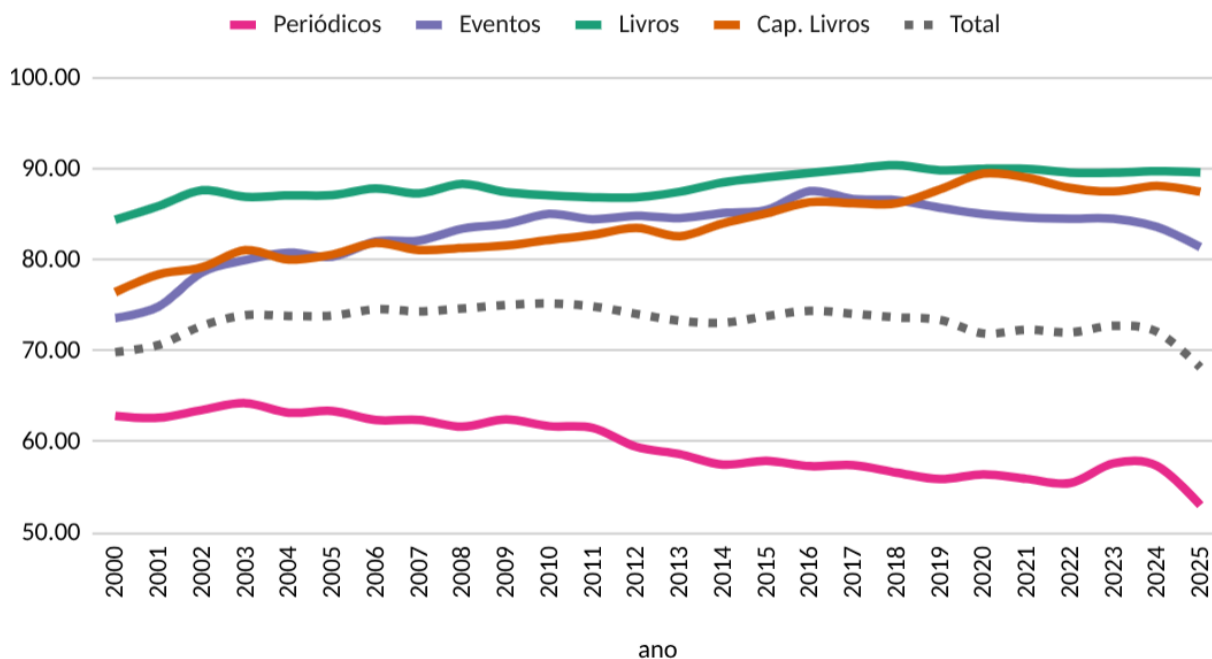
A Figura 1 apresenta a evolução do número total de publicações registradas entre 2000 e 2025. Observa-se crescimento expressivo da produção ao longo do período analisado, passando de aproximadamente 193 mil publicações em 2000 para mais de 680 mil em 2021. Esse crescimento evidencia tanto a expansão da produção registrada nos CVs quanto à consolidação da Plataforma Lattes como fonte relevante para estudos cientométricos. Ainda assim, esses números devem ser interpretados como produção registrada nos CVs, considerando que a base depende do preenchimento e da atualização realizados dos próprios usuários.

Figura 1-Evolução do volume total de publicações por idioma.



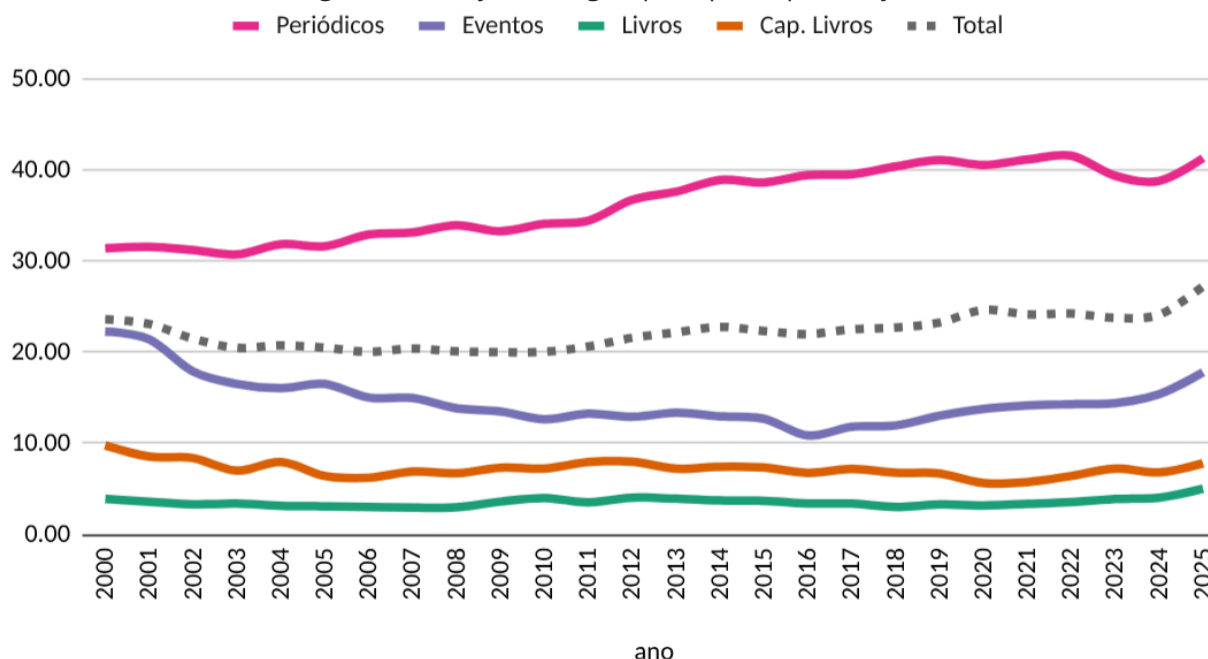
A Figura 2 mostra a evolução da proporção de títulos em português segundo o tipo de publicação. Observa-se predominância do português em livros, cap. de livros e artigos em eventos, frequentemente acima de 80% das publicações. Nos artigos de periódicos, entretanto, a participação do português é significativamente menor e apresenta tendência de redução ao longo do tempo, variando entre aproximadamente 53% e 64% da produção. Esse resultado sugere que os diferentes tipos documentais cumprem papéis distintos nos circuitos de comunicação científica, com maior presença do português em formatos tradicionalmente associados à circulação nacional, à formação acadêmica e à comunicação com comunidades locais ou regionais.

Figura 2 - Evolução do português por tipo de publicação.



A Figura 3 apresenta a evolução da proporção de títulos em inglês. Os resultados indicam que o inglês está concentrado principalmente em artigos de periódicos, nos quais sua participação cresce ao longo do período e atinge valores próximos de 40% das publicações. Nos trabalhos em eventos, livros e capítulos de livros, a presença do inglês permanece significativamente menor, geralmente abaixo de 20%. Esse padrão é compatível com a literatura que associa os periódicos científicos às estratégias de internacionalização e maior visibilidade em bases bibliográficas, embora o uso do inglês, isoladamente, não determine necessariamente maior impacto ou circulação, como observado por Cintra, Silva e Furnival (2020) no contexto da SciELO Brasil.

Figura 3-Evolução do inglês por tipo de publicação.



Os resultados revelam que a distribuição linguística da produção registrada na PL não é homogênea. O idioma predominante dos títulos varia de forma expressiva segundo o tipo documental, indicando a coexistência de diferentes circuitos de comunicação científica. Artigos de periódicos apresentam maior internacionalização linguística, expressa pela presença crescente do inglês. Por outro lado, livros, capítulos e trabalhos em eventos permanecem majoritariamente vinculados ao português evidenciando a importância desse idioma na circulação nacional.

Esse padrão sugere que o português não deve ser interpretado apenas como ausência de internacionalização, mas também como parte constitutiva da comunicação científica brasileira. Assim, a análise da PL oferece uma perspectiva complementar às bases internacionais, permitindo observar dimensões da produção científica que podem ser sub-representadas.

## 4 CONCLUSÃO

Este trabalho analisou a distribuição linguística de mais de 12 milhões de publi-

cações registradas na PL, considerando o idioma predominante dos títulos. Os resultados mostram que o português permanece como o principal idioma da produção registrada nos CVs Lattes, representando 73,17% das publicações, seguido pelo inglês, com 22,54%, e pelo espanhol, com 2,25%. Esses achados sugerem que a produção registrada na PL combina diferentes circuitos de comunicação: um mais associado à circulação internacional, no qual o inglês tem presença relevante, especialmente em artigos de periódicos, e outro nacional ou regional, no qual o português mantém papel central, sobretudo em livros, capítulos e trabalhos em eventos.

As limitações do estudo decorrem principalmente do uso de dados autodeclarados nos currículos, das possíveis inconsistências do processo de deduplicação e da identificação automática de idioma a partir de títulos curtos. Ainda assim, a escala da base analisada oferece um panorama relevante da dimensão linguística da ciência brasileira e reforça o potencial da PL como fonte complementar às bases bibliográficas internacionais. Estudos futuros podem aprofundar a análise por áreas instituições, tipos documentais e outros indicadores de internacionalização científica.

## REFERÊNCIAS

AMMON, Ulrich. Linguistic inequality and its effects on participation in scientific discourse and on global knowledge accumulation. **Applied Linguistics Review**, v. 3, n. 2, p. 333–355, 2012. DOI: 10.1515/applirev-2012-0016. Acesso em: 9 mar. 2026.

CINTRA, Paulo R. ; SILVA, Marco Donizete P.; FURNIVAL, Ariadne C. Uso do inglês como estratégia de internacionalização da produção científica em Ciências Sociais Aplicadas: estudo de caso na SciELO Brasil. **Em Questão**, Porto Alegre, v. 26, n. 1, p. 17-41, jan./abr. 2020. Acesso em: 9 mar. 2026.

GORDIN, Michael D. Scientific Babel: How science was done before and after global English. Chicago: University of Chicago Press, 2015.  
LANE, Julia. Let's make science metrics more scientific. **Nature**, v. 464, p. 488–489, 2010. Acesso em: 15 abr. 2025.

MENA-CHALCO, Jesús P.; CESAR JR., Roberto Marcondes. scriptLattes: an open-source knowledge extraction system from the Lattes platform. **Journal of the Brazilian Computer Society**, v. 15, n. 4, p. 31–39, 2009. DOI: 10.1007/BF03194511. Acesso em: 15 abr. 2025.

MONTGOMERY, Scott L. **Does science need a global language? English and the future of research**. Chicago: University of Chicago Press, 2013.

PACKER, Abel L. *et al.* SciELO: 15 anos de acesso aberto. **Ciência & Saúde Coletiva**, v. 19, n. 7, 2014. Acesso em: 9 mar. 2026.

PRANCKUTĖ, Raminta. Web of Science (WoS) and Scopus: the titans of bibliographic information in today's academic world. **Publications**, v. 9, n. 1, art. 12, 2021. DOI: 10.3390/publications9010012. Acesso em: 9 mar. 2026.

SIMARD, Marc-André; BASSON, Isabel; HARE, Madelaine; LARIVIÈRE, Vincent; MONGEON, Philippe. The open access coverage of OpenAlex, Scopus and Web of Science. **arXiv**, 2024. DOI: 10.48550/arXiv.2404.01985. Acesso em: 9 mar. 2026.