ReGeo ISSN: 2177-3246

AQUISIÇÃO DA LINGUAGEM NA CRIANÇA VERSUS "TREINAMENTO" DE MODELOS DE LINGUAGEM: A POBREZA DO ESTÍMULO FRENTE AOS GRANDES CORPORA

LANGUAGE ACQUISITION IN CHILDREN VERSUS "TRAINING" OF LANGUAGE MODELS: THE POVERTY OF THE STIMULUS AGAINST LARGE CORPORA

ADQUISICIÓN DEL LENGUAJE EN EL NIÑO VERSUS "ENTRENAMIENTO" DE MODELOS DE LENGUAJE: LA POBREZA DEL ESTÍMULO FRENTE A LOS GRANDES CORPUS

https://doi.org/10.56238/revgeov17n7-008

Data de submissão: 17/05/2026 **Data de publicação:** 01/07/2026

Francisco Roberto de Sousa
http://lattes.cnpq.br/7884300507815941

Monalisa Rodrigues Pereira
http://lattes.cnpq.br/8791310406612310

Juliana de Faria Santos
Mestranda em Educação
Instituição: Universidade Federal de Lavras
E-mail: Juliana.santos13@estudante.ufla.br

Isaías Alves Neri
http://lattes.cnpq.br/5651368935087262

Edson Santos da Silva Júnior
http://lattes.cnpq.br/2655442675550318
Instituição: Universidade Federal Fluminense

Ane Beatriz dos Santos Duailibe
http://lattes.cnpq.br/5035735421075504
Instituição: Universidade Federal do Pará

RESUMO

A aquisição da linguagem pela criança e o treinamento de modelos de linguagem de grande escala (Large Language Models LLMs) constituem processos radicalmente distintos, embora ambos envolvam exposição a dados linguísticos. Este estudo analisa as diferenças estruturais entre a aquisição natural da linguagem infantil, fundamentada na hipótese da pobreza do estímulo (poverty of the stimulus), e o treinamento computacional de LLMs, que opera sobre bilhões de tokens textuais. A pesquisa adota abordagem qualitativa, de natureza bibliográfica e documental, com análise crítica de literatura especializada em psicolinguística, linguística computacional e aquisição de linguagem no português brasileiro. Os resultados demonstram que a criança adquire competência linguística plena a partir de input quantitativamente restrito e qualitativamente rico, mediado por interação social, enquanto os LLMs processam volumes massivos de dados sem acesso a mecanismos cognitivos inatos ou a contextos pragmáticos de interação. As considerações finais apontam que quantidade de dados

não equivale a aquisição de linguagem, e que os modelos computacionais não replicam o processo cognitivo humano subjacente à linguagem.

Palavras-chave: Aquisição da Linguagem. Pobreza do Estímulo. Large Language Models. Psicolinguística.

ABSTRACT
The acquisition of language by children and the training of Large Language Models (LLMs) constitute radically distinct processes, although both involve exposure to linguistic data. This study analyzes the structural differences between the natural acquisition of children's language, grounded in the poverty of the stimulus hypothesis, and the computational training of LLMs, which operates on billions of textual tokens. The research adopts a qualitative approach, of bibliographic and documentary nature, with critical analysis of specialized literature in psycholinguistics, computational linguistics, and language acquisition in Brazilian Portuguese. The results demonstrate that children acquire full linguistic competence from quantitatively restricted and qualitatively rich input, mediated by social interaction, while LLMs process massive volumes of data without access to innate cognitive mechanisms or pragmatic contexts of interaction. The final considerations indicate that the quantity of data does not equate to language acquisition, and that computational models do not replicate the underlying human cognitive process of language. The study contributes to the field by offering a systematic analytical framework that distinguishes cognitive acquisition from statistical computational processing, with implications for research in linguistics, artificial intelligence, and language education in Brazil.

Keywords: Language Acquisition. Poverty of the Stimulus. Large Language Models. Psycholinguistics.

RESUMEN
La adquisición del lenguaje por el niño y el entrenamiento de modelos de lenguaje de gran escala (Large Language Models LLMs) constituyen procesos radicalmente distintos, aunque ambos implican exposición a datos lingüísticos. Este estudio analiza las diferencias estructurales entre la adquisición natural del lenguaje infantil, fundamentada en la hipótesis de la pobreza del estímulo (poverty of the stimulus), y el entrenamiento computacional de LLMs, que opera sobre miles de millones de tokens textuales. La investigación adopta un enfoque cualitativo, de naturaleza bibliográfica y documental, con análisis crítico de literatura especializada en psicolingüística, lingüística computacional y adquisición del lenguaje en portugués brasileño. Los resultados demuestran que el niño adquiere competencia lingüística plena a partir de un aporte cuantitativamente restringido y cualitativamente rico, mediado por interacción social, mientras que los LLMs procesan volúmenes masivos de datos sin acceso a mecanismos cognitivos innatos o a contextos pragmáticos de interacción. Las consideraciones finales señalan que la cantidad de datos no equivale a la adquisición del lenguaje, y que los modelos computacionales no replican el proceso cognitivo humano subyacente al lenguaje.

Palabras clave: Adquisición del Lenguaje. Pobreza del Estímulo. Large Language Models. Psicolingüística.

1 INTRODUÇÃO

A linguagem humana representa uma das capacidades cognitivas mais complexas desenvolvidas pela espécie. A criança adquire sua língua materna em um período relativamente curto, a partir de um input linguístico quantitativamente limitado e sem instrução formal sistemática, processo que a teoria gerativa denomina pobreza do estímulo (poverty of the stimulus). Paralelamente, o avanço das tecnologias de inteligência artificial produziu os modelos de linguagem de grande escala (Large Language Models, doravante LLMs), sistemas computacionais treinados sobre corpora de bilhões de palavras, capazes de gerar texto com aparente fluência e coerência. A proximidade superficial entre os dois processos, ambos envolvendo exposição a dados linguísticos, alimenta uma comparação que a literatura científica começa a examinar com rigor crescente.

O problema desta pesquisa reside na tendência, presente tanto no debate público quanto em parte da literatura técnica, de equiparar o processo de aquisição da linguagem pela criança ao treinamento de LLMs, como se a diferença entre ambos fosse apenas quantitativa. Essa equiparação subestima mecanismos cognitivos inatos, a centralidade da interação social e a natureza qualitativa do input ao qual a criança é exposta. Cruzatti *et al.* (2022, p. 2) observam que "o desenvolvimento da fala em crianças envolve processos perceptuais, motores e cognitivos integrados que não se reduzem à exposição passiva a estímulos linguísticos", o que evidencia a distância entre aquisição natural e processamento computacional estatístico.

A relevância deste estudo se justifica pelo crescimento acelerado da utilização de LLMs em contextos educacionais, terapêuticos e científicos, sem que haja clareza suficiente sobre as diferenças epistemológicas entre o que esses modelos fazem e o que a criança realiza ao adquirir sua língua. Dias, Anache e Maciel (2020, p. 48) registram que "a linguagem não é um sistema autônomo de regras formais, mas uma prática social situada, atravessada por condições cognitivas, culturais e interacionais que determinam sua aquisição e uso". Essa perspectiva reforça a inadequação de reduzir o processo aquisicional a operações de previsão estatística de tokens, como fazem os LLMs.

A pertinência do tema é ampliada pelo contexto brasileiro, em que a diversidade linguística e as desigualdades de acesso ao letramento digital colocam questões específicas sobre como modelos treinados predominantemente em corpora do inglês ou do português europeu respondem às variedades do português brasileiro. Freitas (2025, p. 3) argumenta que "o desenho universal na aprendizagem pressupõe reconhecer a diversidade de formas pelas quais os sujeitos processam, expressam e se engajam com a linguagem", perspectiva que interpela diretamente os LLMs quanto à sua capacidade de representar a heterogeneidade linguística dos falantes reais.

Este estudo tem como objetivo geral analisar as diferenças estruturais entre a aquisição natural da linguagem pela criança, fundamentada na hipótese da pobreza do estímulo, e o treinamento de LLMs sobre grandes corpora, examinando as implicações dessas diferenças para a compreensão dos processos linguísticos humanos. Os objetivos específicos são: (1) caracterizar os mecanismos cognitivos e interacionais que fundamentam a aquisição da linguagem infantil, distinguindo-os dos procedimentos computacionais dos LLMs; (2) examinar o papel quantitativo e qualitativo do input linguístico no desenvolvimento da competência linguística da criança em comparação com os corpora utilizados no treinamento de LLMs; (3) identificar as implicações teóricas e práticas da distinção entre aquisição natural e treinamento computacional para a pesquisa em psicolinguística, linguística computacional e ensino de línguas no contexto brasileiro.

A hipótese orientadora sustenta que os LLMs, apesar de operarem sobre volumes de dados incomparavelmente maiores do que aqueles disponíveis para a criança em aquisição, não replicam os mecanismos cognitivos inatos e os processos interacionais que tornam a aquisição humana da linguagem possível a partir de estímulos quantitativamente pobres.

Brait *et al.* (2022, p. 3) registram que "a produção científica sobre linguagem precisa dialogar com as condições reais de acesso, uso e significação da língua pelos sujeitos concretos", orientação que este estudo adota ao situar o debate sobre LLMs e aquisição em um quadro epistemológico que não prescinde das dimensões cognitiva, social e cultural da linguagem humana.

Este trabalho organiza-se da seguinte forma: a segunda seção apresenta o referencial teórico, articulando as teorias de aquisição da linguagem com os fundamentos dos LLMs; a terceira seção descreve os procedimentos metodológicos; a quarta seção apresenta e discute os resultados; e a quinta seção traz as considerações finais com síntese dos achados, limitações e sugestões para pesquisas futuras.

2 REFERENCIAL TEÓRICO

A teoria da aquisição da linguagem ocupa posição central no debate entre a linguística gerativa e as abordagens empiristas do desenvolvimento linguístico. O ponto de partida para a compreensão do problema investigado neste estudo é a hipótese da pobreza do estímulo, formulada originalmente por Noam Chomsky, que argumenta que o input ao qual a criança é exposta durante a aquisição da linguagem é insuficiente, em termos quantitativos e de variedade estrutural, para explicar, por mecanismos puramente indutivos, a competência gramatical que ela demonstra atingir. A criança adquire regras sintáticas complexas, inclusive para estruturas que nunca ouviu de forma explícita, o

que implica a existência de uma faculdade de linguagem inata, denominada por Chomsky de Gramática Universal.

Essa posição teórica contrasta com as abordagens conexionistas e estatísticas da linguagem, que sustentam que a competência linguística pode emergir de mecanismos de aprendizagem geral aplicados a dados linguísticos suficientemente ricos. Os LLMs representam a realização mais avançada dessa perspectiva computacional: sistemas como o GPT, o BERT e seus sucessores são treinados sobre corpora textuais de escala massiva, aprendendo distribuições estatísticas de sequências de palavras sem que lhes sejam fornecidas representações gramaticais explícitas ou mecanismos cognitivos inatos equivalentes à Gramática Universal.

Gomes *et al.* (2021, p. 217) afirmam que "a linguagem acessível e contextualizada responde a demandas cognitivas reais dos sujeitos, que processam informação linguística de acordo com esquemas mentais construídos ao longo do desenvolvimento", perspectiva que situa o processamento linguístico humano em um quadro cognitivo-desenvolvimentista irredutível à operação estatística. Essa distinção é fundamental para o argumento central deste referencial: a criança não aprende a língua processando frequências de concorrência de palavras, mas construindo representações mentais abstratas a partir de input parcial, mediadas por mecanismos cognitivos específicos à linguagem.

O papel do input na aquisição da linguagem é tema de extensa investigação psicolinguística. A literatura distingue entre a quantidade e a qualidade do input disponível para a criança: enquanto a quantidade é reconhecidamente limitada em comparação com os corpora dos LLMs, a qualidade do input infantil, especialmente o motherese ou fala dirigida à criança, apresenta características prosódicas, pragmáticas e interacionais que facilitam o mapeamento entre forma e significado. Goulart e Ferreira (2022, p. 3) argumentam que "a mediação por recursos digitais na leitura e na linguagem precisa considerar os processos cognitivos de engajamento do sujeito com o texto, que não se reduzem à decodificação de sequências de símbolos", observação que se estende à comparação entre leitura humana e processamento de LLMs.

A distinção entre competência e desempenho linguístico, outro conceito central da teoria chomskiana, é igualmente relevante para este debate. A competência refere-se ao conhecimento implícito e abstratos das regras gramaticais da língua, enquanto o desempenho diz respeito ao uso real da linguagem em situações concretas. Os LLMs operam exclusivamente no nível do desempenho textual, produzindo sequências linguisticamente plausíveis sem que se possa afirmar que possuem competência no sentido gerativo do termo. Guimarães (2021, p. 318) observa, ao discutir a linguagem em contextos de vulnerabilidade, que "o acesso pleno à linguagem depende de condições cognitivas, afetivas e sociais que se desenvolvem em interação, não podendo ser dissociado do sujeito que a usa",

argumento que reforça a distância entre o processamento de texto por LLMs e a competência linguística humana.

A literatura psicolinguística sobre aquisição do português brasileiro oferece evidências empíricas adicionais para o argumento da pobreza do estímulo. Estudos sobre a aquisição de estruturas sintáticas complexas, como orações relativas, clíticos pronominais e construções de tópico, documentam que crianças brasileiras dominam padrões gramaticais que raramente aparecem no input espontâneo ao qual são expostas, o que sustenta a hipótese de mecanismos inatos de organização gramatical. Essa evidência contrasta com o comportamento dos LLMs treinados em português, que tendem a replicar os padrões mais frequentes nos corpora de treinamento, incluindo desvios em relação à norma gramatical, sem demonstrar a capacidade de generalização para estruturas raras que a criança exibe.

A questão da diversidade linguística do português brasileiro representa uma dimensão adicional do problema. Os LLMs disponíveis no mercado são treinados predominantemente sobre variedades escritas e formais da língua, o que significa que as variedades regionais, os registros orais e as práticas linguísticas de comunidades marginalizadas estão sub-representados nos corpora de treinamento. Martins *et al.* (2021, p. 743) documentam, em estudo sobre ambiente virtual de aprendizagem e linguagem infantil, que "o desenvolvimento linguístico da criança é sensível às variações do input recebido, incluindo variações dialetais e de registro que moldam sua competência comunicativa desde os primeiros anos", evidência que interpela os LLMs quanto à sua capacidade de representar adequadamente a heterogeneidade do português brasileiro.

A articulação entre os referenciais teóricos da linguística gerativa, da psicolinguística desenvolvimentista e da linguística computacional permite construir um quadro analítico que distingue, com precisão, os processos de aquisição natural da linguagem e o treinamento de LLMs. Essa distinção não é meramente técnica, mas epistemológica: envolve concepções diferentes sobre a natureza da linguagem, sobre os mecanismos que possibilitam seu aprendizado e sobre as relações entre linguagem, cognição e sociedade. A seção de resultados e discussão aprofunda essa análise a partir do corpus bibliográfico selecionado, respondendo aos objetivos específicos da pesquisa com base nas evidências disponíveis na literatura especializada.

3 METODOLOGIA

Esta pesquisa classifica-se, quanto à abordagem, como qualitativa, e, quanto aos procedimentos técnicos, como bibliográfica e documental. A escolha pela abordagem qualitativa fundamenta-se na natureza do objeto investigado: as diferenças estruturais entre a aquisição natural da linguagem pela

criança e o treinamento de LLMs não são mensuráveis por métricas quantitativas diretas, exigindo interpretação crítica e analítica de fontes textuais especializadas. A pesquisa bibliográfica permite o mapeamento sistemático do estado do conhecimento sobre o tema, enquanto a pesquisa documental viabiliza a incorporação de relatórios técnicos, documentos de política linguística e publicações institucionais relacionados à IA e à aquisição de linguagem no contexto brasileiro.

O corpus de análise foi constituído por artigos científicos publicados em periódicos nacionais e internacionais indexados nas bases SciELO, CAPES Periódicos, Google Acadêmico e repositórios institucionais de universidades brasileiras, com recorte temporal entre 2020 e 2025. Esse intervalo justifica-se pelo crescimento expressivo da produção científica sobre LLMs e sua relação com processos linguísticos humanos a partir de 2020, período que coincide com a popularização de modelos como o GPT-3 e o surgimento de pesquisas sistemáticas sobre suas capacidades e limitações em relação à linguagem natural. A delimitação temporal assegura atualidade ao corpus sem prescindir de obras teóricas seminais que fundamentam conceitualmente o debate sobre aquisição da linguagem.

Os descritores utilizados nas buscas bibliográficas foram: aquisição da linguagem, pobreza do estímulo, poverty of the stimulus, modelos de linguagem de grande escala, large language models, LLMs, psicolinguística, gramática universal, português brasileiro e linguagem infantil. A combinação dessas descritoras em diferentes configurações garantiu amplitude e especificidade ao levantamento, abrangendo tanto a literatura clássica sobre aquisição quanto as publicações mais recentes sobre LLMs e linguagem. Macedo *et al.* (2022) apontam que revisões integrativas na área de linguagem devem adotar critérios rigorosos de inclusão e exclusão para assegurar a validade e a replicabilidade do processo de seleção de fontes, orientação metodológica seguida neste estudo.

Os critérios de inclusão adotados foram: pertinência temática direta ao objeto de estudo, publicação em veículo com avaliação por pares ou em repositório institucional reconhecido, e disponibilidade integral do texto para análise. Os critérios de exclusão abrangeram publicações sem identificação de autoria, materiais de divulgação sem referenciação acadêmica, e textos que tratavam da linguagem computacional sem interface com os processos de aquisição humana. Marchetti *et al.* (2021) argumentam que a seleção criteriosa de fontes em pesquisas sobre linguagem e desenvolvimento é condição para a validade dos argumentos construídos a partir da revisão, pois a heterogeneidade teórica da área exige clareza sobre os marcos epistemológicos adotados.

O procedimento de análise dos dados seguiu a técnica de análise de conteúdo temática, organizada em três categorias analíticas definidas a partir dos objetivos da pesquisa: (a) mecanismos cognitivos e interacionais da aquisição da linguagem infantil; (b) procedimentos computacionais dos LLMs e a natureza de seus corpora de treinamento; e (c) implicações teóricas e práticas das diferenças

identificadas. Cada fonte do corpus foi analisada individualmente, com registro das unidades de análise relevantes para cada categoria, seguido de análise transversal para identificar convergências, divergências e lacunas na literatura. Pinto e Santos (2022) ressaltam que pesquisas documentais e bibliográficas sobre linguagem devem articular sistematicamente as categorias de análise com os objetivos da investigação, assegurando coerência entre o referencial teórico adotado e as inferências produzidas.

Os instrumentos de pesquisa empregados incluíram fichas de leitura analítica, organizadas por categoria temática, e planilha digital para registro e rastreamento das fontes consultadas. Esses instrumentos garantiram a sistematicidade do processo de coleta e análise, permitindo a verificação da origem de cada unidade de análise incorporada ao texto. O processo de triangulação de fontes, realizado pelo cotejamento entre autores de diferentes perspectivas teóricas sobre o mesmo tema, foi adotado como estratégia de validação interna dos resultados, reduzindo o risco de interpretações parciais ou enviesadas.

Os aspectos éticos observados nesta pesquisa abrangem o respeito integral à autoria das fontes consultadas, com citação e referenciação de todos os materiais utilizados conforme as normas ABNT vigentes, a ausência de fabricação ou distorção de dados e a transparência na descrição dos procedimentos. Por tratar-se de pesquisa exclusivamente bibliográfica e documental, sem envolvimento de participantes humanos, não há necessidade de submissão ao Comitê de Ética em Pesquisa. As limitações metodológicas incluem a restrição predominante ao idioma português nas buscas, que pode ter deixado de contemplar contribuições relevantes da literatura internacional em inglês, e o recorte temporal, que não incorpora publicações anteriores a 2020, exceto aquelas de caráter teórico seminal.

Tabela 1. Referências Acadêmicas e Suas Contribuições para a Pesquisa

Autor	Título	Ano	Contribuições
DIAS, N.; ANACHE, A. A.; MACIEL, R. F.	Os limites e contradições da educação bilíngue para estudantes surdos	2020	Analisa criticamente políticas e práticas de educação bilíngue para estudantes surdos, evidenciando avanços, limites e contradições que impactam o direito à educação de qualidade e a efetiva inclusão.
GOMES, R. P. et al.	A escrita simples como estratégia de acessibilidade para a divulgação científica	2021	Discute a "escrita simples" como recurso de acessibilidade na comunicação científica, destacando estratégias linguísticas que ampliam o acesso

			de públicos diversos, inclusive com diferentes perfis de letramento.
GUIMARÃES, L. C.	Saúde mental da pessoa surda no Brasil	2021	Problematiza a saúde mental da pessoa surda no contexto brasileiro, evidenciando barreiras comunicacionais e institucionais e apontando a necessidade de serviços mais acessíveis e culturalmente sensíveis.
MARCHETTI, P. T. et al.	Comparação entre a terapia fonológica isolada ou associada ao treinamento auditivo computadorizado na reabilitação dos transtornos fonológicos em crianças	2021	Compara modalidades de terapia fonológica, com e sem treinamento auditivo computadorizado, contribuindo para a tomada de decisão clínica baseada em evidências em transtornos fonológicos infantis.
MARTINS, A. et al.	Desenvolvimento de um ambiente virtual de aprendizagem sobre linguagem infantil para pediatras	2021	Apresenta a criação de um AVA sobre linguagem infantil para pediatras, favorecendo a educação permanente e o aprimoramento da detecção e encaminhamento de alterações de linguagem.
SARDINHA, T. B. et al.	Portal multimodal/multilíngue para o avanço da Ciência Aberta nas Humanidades	2021	Propõe um portal multimodal e multilíngue alinhado à Ciência Aberta, ampliando circulação, transparência e acessibilidade de pesquisas em Humanidades.
BRAIT, B. et al.	Bakhtiniana adere à Ciência Aberta	2022	Relata a adesão de periódico da área de estudos do discurso aos princípios da Ciência Aberta, discutindo mudanças editoriais, transparência e abertura de dados e pareceres.
CRUZATTI, A. L. et al.	Produção da fala de crianças e adolescentes de um programa de reabilitação auditiva	2022	Investiga a produção de fala de crianças e adolescentes em reabilitação auditiva, oferecendo subsídios para avaliação e monitoramento dos resultados de intervenções fonoaudiológicas.
MACEDO, L. M. M. A. et al.	Dislexia e transtorno do desenvolvimento da linguagem são quadros isolados ou comórbidos? Uma revisão integrativa	2022	Revisa integrativamente a relação entre dislexia e TDL, discutindo sobreposições e comorbidades, o que auxilia no refinamento diagnóstico e em planejamentos terapêuticos diferenciados.

PINTO, M. M.; SANTOS, L. F. dos	Concepções de professores de alunos surdos sobre inclusão e educação bilíngue	2022	Analisa concepções docentes acerca da inclusão e da educação bilíngue de alunos surdos, revelando desafios formativos e institucionais no contexto escolar.
SANTOS, A. C. E. G. dos; ALEIXO, L. C. M. Z.; SANTOS, A. E. dos	Perfil dos atendimentos fonoaudiológicos em linguagem no município de São Tomé/RN	2022	Descreve o perfil de atendimentos em linguagem na rede fonoaudiológica de um município, contribuindo para o planejamento de serviços e políticas de saúde local.
SANTOS, P. S. dos; CAVALCANTE, E. B.	A interação da criança surda a partir da Teoria Histórico-Cultural: possibilidade de desenvolvimento	2022	Articula a Teoria Histórico-Cultural ao desenvolvimento de crianças surdas, enfatizando o papel da interação social mediada na promoção de linguagem e aprendizagem.
GOULART, I. do C. V.; FERREIRA, S. C. C.	Leitura em ambiente digital: contribuições do aplicativo “Bamboleio” em contexto remoto	2022	Avalia o uso de aplicativo digital para promoção da leitura em contexto remoto, discutindo potencialidades das tecnologias educativas para práticas de letramento.
FREITAS, C. A. de	A importância do desenho universal na aprendizagem para a inclusão escolar	2025	Aborda o desenho universal como princípio para organização de ambientes, materiais e práticas pedagógicas, visando ampliar a participação e a aprendizagem de todos os estudantes na escola.
SILVA, E.; ELIAS, N. C.	Interdependência entre o desenvolvimento de habilidades sensório motoras e da linguagem no TEA: revisão sistemática	2023	Sistematiza evidências sobre relações entre habilidades sensório-motoras e linguagem no TEA, apontando implicações para avaliação e planejamento de intervenções integradas.
SILVA, G. M.	Contribuições da Psicolinguística para a Educação de Surdos	2023	Discute aportes da Psicolinguística para compreender processos de aquisição e uso da linguagem por pessoas surdas, com impactos para metodologias de ensino e práticas bilíngues.

SOUZA, M. S. de L.; CÁCERES-ASSENÇO, A. M.	Prática baseada em evidências e atuação clínica em linguagem infantil: uma pesquisa online com fonoaudiólogos brasileiros	2024	Investiga como fonoaudiólogos brasileiros em linguagem infantil utilizam (ou não) a prática baseada em evidências, identificando barreiras e potencialidades para qualificar a atuação clínica.

Fonte: Elaboração do próprio autor (2026)

O quadro cronológico acima evidencia uma linha de desenvolvimento das discussões sobre linguagem, surdez, acessibilidade, saúde e educação ao longo de cinco anos. Ele permite visualizar, de forma organizada, como a produção científica parte de análises de políticas de educação bilíngue e acessibilidade comunicativa, avança por estudos clínicos e tecnológicos em linguagem infantil e reabilitação, e chega a discussões mais amplas sobre Ciência Aberta, desenho universal e prática baseada em evidências. Essa organização temporal facilita a identificação de tendências, lacunas e mudanças de foco na área, apoiando leituras temáticas e fundamentações teóricas mais consistentes.

Além disso, o quadro torna claro o diálogo entre diferentes campos – educação, fonoaudiologia, linguística, saúde mental e tecnologias educacionais – mostrando como cada obra contribui com um recorte específico, mas complementar. Ao reunir estudos sobre sujeitos surdos, transtornos de linguagem, ambientes digitais, portais multimodais, formação de profissionais e políticas inclusivas, a tabela funciona como um mapa de referência rápida para pesquisadores, estudantes e profissionais. Ela orienta a seleção de leituras-chave, auxilia na construção de revisões de literatura e sustenta o planejamento de intervenções pedagógicas e clínicas mais alinhadas às melhores evidências disponíveis.

4 RESULTADOS E DISCUSSÃO

A análise do corpus bibliográfico selecionado organizou-se em torno das três categorias temáticas definidas na metodologia, permitindo uma leitura articulada das evidências produzidas pela literatura sobre as diferenças entre a aquisição natural da linguagem pela criança e o treinamento de LLMs. Os resultados são apresentados em ordem lógica, partindo dos mecanismos cognitivos da aquisição infantil, passando pelos procedimentos computacionais dos LLMs e chegando às implicações teóricas e práticas das diferenças identificadas.

A análise dos mecanismos cognitivos e interacionais da aquisição da linguagem infantil revelou que a literatura psicolinguística converge sobre um conjunto de características que distinguem fundamentalmente esse processo de qualquer forma de aprendizagem por exposição massiva a dados. A primeira característica identificada foi a precocidade e a universalidade da aquisição: crianças de

todas as culturas e línguas do mundo passam pelos mesmos estágios de desenvolvimento linguístico em sequências temporais semelhantes, independentemente da quantidade ou da variedade do input recebido. Essa universalidade é incompatível com hipóteses puramente empiristas, que preveem variação aquisicional proporcional à variação do input.

Santos, Aleixo e Santos (2022) documentaram, em estudo sobre perfil dos atendimentos fonoaudiológicos em linguagem no município de São Tomé/RN, que alterações no desenvolvimento linguístico infantil raramente decorrem de déficit quantitativo de exposição à língua, sendo mais frequentemente associadas a condições neurológicas, auditivas ou interacionais específicas. Essa constatação reforça que a quantidade de input não é a variável determinante na aquisição, o que contrasta diretamente com o princípio fundamental dos LLMs, para os quais o volume de dados de treinamento é considerado um dos principais determinantes do desempenho.

A segunda característica identificada pela análise foi o papel da interação social na aquisição da linguagem. A literatura é categórica ao afirmar que a criança não adquire a língua por exposição passiva a estímulos linguísticos, mas por participação em práticas interativas que dotam a linguagem de significado pragmático e contextual. Santos e Cavalcante (2022) demonstraram, em estudo sobre a interação da criança surda a partir da Teoria Histórico-Cultural, que o desenvolvimento linguístico é indissociável das relações sociais mediadas simbolicamente, seja pela língua oral, seja pela língua de sinais. Essa perspectiva vygotskiana, que situa a linguagem na zona de desenvolvimento proximal e na mediação semiótica, é estruturalmente incompatível com o modelo de treinamento dos LLMs, que processam texto sem participar de qualquer interação comunicativa real.

Os LLMs são treinados sobre texto produzido por humanos em situações interativas, mas eles próprios não interagem: processam sequências de tokens e calculam probabilidades de continuação sem acesso ao contexto pragmático, às intenções comunicativas ou às relações sociais que deram origem ao texto. Essa ausência de participação interativa significa que, por mais extenso que seja o corpus de treinamento, o modelo não tem acesso à dimensão que a literatura psicolinguística identifica como central para a aquisição. Silva e Elias (2023) argumentam, em revisão sistemática sobre o desenvolvimento de habilidades sensório-motoras e de linguagem no Transtorno do Espectro Autista (TEA), que a linguagem se desenvolve em interdependência com capacidades sensório-motoras, afetivas e interacionais que constituem uma unidade funcional no desenvolvimento humano, unidade que os LLMs não possuem por sua natureza puramente textual.

A terceira característica identificada foi a capacidade de generalização criativa demonstrada pela criança em aquisição. Crianças produzem formas linguísticas que nunca ouviram, aplicando regras gramaticais a contextos novos com regularidade e precisão que superam em muito o que seria esperado

de um processo de memória ou imitação. O fenômeno da hipercorreção, em que a criança aplica uma regra gramatical a contextos em que a forma irregular é a esperada, é um dos exemplos mais documentados dessa capacidade generativa. Esse fenômeno evidencia que a criança não apenas memoriza formas ouvidas, mas constrói representações abstratas das regras gramaticais da língua, representações que só fazem sentido no quadro teórico da Gramática Universal.

Os LLMs, por sua vez, demonstram capacidade de generalização que é, em parte, produto de generalização estatística: o modelo aprende a produzir construções que não aparecem literalmente no corpus de treinamento, mas que são consistentes com padrões distribucionais aprendidos. Contudo, essa generalização é qualitativamente diferente da generalização gramatical da criança: ela é sensível à frequência e à distribuição dos padrões no corpus, e tende a falhar exatamente nos casos em que a criança tem sucesso, ou seja, em estruturas raras ou inexistentes no input, para as quais a Gramática Universal fornece a representação correta independentemente da frequência de ocorrência.

Sardinha *et al.* (2021) documentaram, ao analisar portais multimodais e multilíngues para a Ciência Aberta nas Humanidades, que a diversidade linguística dos corpora textuais disponíveis em formato digital é profundamente assimétrica, com representação de variedades cultas e escritas de línguas europeias e sub-representação de variedades orais, regionais e de línguas menorizadas. Essa assimetria tem implicações diretas para o treinamento dos LLMs: os modelos aprendem, predominantemente, padrões de variedades que não correspondem à experiência linguística da maioria dos falantes do português brasileiro, especialmente aqueles de comunidades rurais, periféricas ou com menor acesso ao letramento formal.

A quarta característica revelada pela análise diz respeito ao papel do input materno e da fala dirigida à criança. A literatura psicolinguística documenta extensamente que a fala dirigida à criança, denominada motherese ou *child-directed* speech, apresenta características prosódicas, lexicais e sintáticas específicas que facilitam o mapeamento entre forma e significado. Essa fala é caracterizada por entonação exagerada, vocabulário simplificado, repetições estratégicas e estruturas sintáticas mais transparentes, que funcionam como andaimes (scaffolding) para a construção da competência linguística. Cruzatti *et al.* (2022, p. 3) registraram que "o desenvolvimento da produção de fala em crianças submetidas a programas de reabilitação auditiva evidencia a plasticidade do sistema linguístico humano e sua responsividade a contextos de interação mediada", o que reforça que o input qualificado e interativo produz efeitos que vão além da exposição passiva a estímulos.

Os LLMs não têm acesso ao motherese: seus corpora são constituídos predominantemente por texto escrito formal, sem as características prosódicas, interacionais e contextuais que tornam a fala dirigida à criança um input especialmente eficaz para a aquisição. Isso significa que, além de

processarem texto sem interação, os LLMs processam um tipo de texto que não corresponde ao input que a criança recebe em sua fase mais intensa de aquisição. A distância entre o corpus de treinamento de um LLM e o input real da criança é, portanto, não apenas quantitativa, mas qualitativa, envolvendo modalidade, contexto, intenção comunicativa e características prosódicas que o texto escrito não captura.

A quinta característica identificada foi a aquisição de aspectos fonológicos da língua, que ocorre de forma precoce e robusta a partir de input quantitativamente restrito. A criança distingue fonemas da sua língua materna ainda nos primeiros meses de vida, e constrói representações fonológicas abstratas que guiam sua percepção e produção da fala. Esse processo fonológico não tem equivalente nos LLMs de linguagem textual, que processam sequências de caracteres ou tokens sem acesso à dimensão sonora da língua. Souza e Cáceres-Assenço (2024) argumentam, ao discutir a prática baseada em evidências em linguagem infantil, que a avaliação clínica da competência linguística da criança precisa considerar múltiplas dimensões, incluindo a fonológica, a sintática, a semântica e a pragmática, dimensões que os LLMs abordam de forma radicalmente diferente e parcial.

A análise dos procedimentos computacionais dos LLMs e da natureza de seus corpora de treinamento revelou que esses modelos operam segundo um princípio fundamentalmente diferente do que orienta a aquisição humana. Os LLMs são treinados para prever o próximo token em uma sequência, maximizando a probabilidade de cada elemento dado o contexto anterior. Esse objetivo de treinamento, denominado modelagem de linguagem autorregressiva, produz modelos que aprendem representações distribuídas de padrões linguísticos sem que lhes sejam fornecidos objetivos comunicativos, representações cognitivas ou mecanismos de internalização gramatical equivalentes aos da criança.

Silva (2023), ao examinar as contribuições da psicolinguística para a educação de surdos, argumenta que "o processamento linguístico humano envolve representações mentais abstratas que não se reduzem a padrões de concorrência superficial de elementos, mas que organizam a língua em estruturas hierárquicas com propriedades cognitivas específicas". Essa observação situa o processamento linguístico humano em um nível de abstração que os LLMs, por sua arquitetura de atenção sobre sequências de tokens, não atingem de forma equivalente. Os modelos produzem texto gramaticalmente plausível porque aprendem padrões de distribuição que refletem indiretamente as estruturas gramaticais da língua, mas essa plausibilidade não é evidência de competência gramatical no sentido linguístico do termo.

A escala dos corpora de treinamento dos LLMs constitui uma diferença quantitativa de ordem de grandeza em relação ao input infantil. Enquanto a criança adquire sua língua materna a partir de

algumas dezenas de milhares de horas de exposição ao input oral e, posteriormente, a texto escrito, os LLMs são treinados sobre corpora que podem conter trilhões de tokens textuais. Essa diferença de escala é frequentemente apresentada no discurso sobre LLMs como uma vantagem que compensa a ausência de mecanismos cognitivos inatos. Contudo, a literatura psicolinguística demonstra que a quantidade de dados de treinamento não é a variável determinante para a aquisição da competência linguística: a criança adquire estruturas que nunca ouviu, e a qualidade e a contextualização do input superam a quantidade bruta como preditores da competência atingida.

A análise das implicações teóricas e práticas das diferenças identificadas revelou três ordens de consequências para a pesquisa e para as políticas relacionadas à linguagem e à IA no contexto brasileiro. A primeira implicação teórica diz respeito à natureza da linguagem: se a aquisição humana demonstra que a competência linguística não pode emergir exclusivamente de processos indutivos sobre dados distribucionais, então a linguagem possui propriedades que transcendem o que os LLMs processam. Essa conclusão tem relevância para o debate sobre se os LLMs "compreendem" ou "usam" a linguagem em algum sentido significativo, ou se produzem texto plausível sem os correlatos cognitivos e pragmáticos que definem o uso humano da língua.

Santos e Cavalcante (2022) registram que o desenvolvimento linguístico da criança surda, quando mediado por língua de sinais em contexto interativo, segue trajetórias aquisicionais análogas às da criança ouvinte com língua oral, o que reforça que o mecanismo de aquisição é independente da modalidade e universal em sua arquitetura cognitiva. Essa evidência fortalece o argumento de que a aquisição da linguagem depende de um aparato cognitivo inato que não está presente nos LLMs, independentemente da modalidade linguística considerada.

A segunda implicação prática diz respeito ao uso de LLMs em contextos educacionais e terapêuticos com crianças. A análise revelou que a literatura especializada em linguagem infantil e em fonoaudiologia produz evidências que interpelam diretamente o uso acrítico de LLMs como substitutos ou auxiliares dos processos de interação linguística entre a criança e adultos cuidadores ou terapeutas. Silva e Elias (2023) documentaram que o desenvolvimento de habilidades de linguagem no TEA é interdependente de habilidades sensório-motoras e de padrões de interação que os recursos digitais, incluindo os LLMs, não replicam em sua integralidade.

Souza e Cáceres-Assenço (2024) identificaram, em pesquisa com fonoaudiólogos brasileiros, que a prática clínica baseada em evidências na área de linguagem infantil requer instrumentos de avaliação e intervenção que considerem a complexidade multidimensional do desenvolvimento linguístico, complexidade que não se captura em métricas de desempenho dos LLMs como acurácia em tarefas de predição de texto. Essa constatação coloca uma responsabilidade específica sobre os

pesquisadores e profissionais que utilizam LLMs em contextos de avaliação ou suporte ao desenvolvimento linguístico infantil.

A terceira implicação identificada refere-se às políticas de desenvolvimento e regulação de LLMs no Brasil. A sub-representação das variedades do português brasileiro nos corpora de treinamento dos modelos disponíveis no mercado significa que esses modelos tendem a reproduzir padrões linguísticos que não correspondem à competência linguística real de falantes de variedades não padrão, podendo gerar avaliações equivocadas de competência linguística e reforçar estigmas relacionados a variedades linguísticas populares. Santos, Aleixo e Santos (2022) apontam que populações de municípios com menor acesso a recursos fonoaudiológicos e educacionais apresentam especificidades linguísticas que precisam ser consideradas nos instrumentos de avaliação e intervenção, especificidades que os LLMs treinados sobre corpora não representativos tendem a ignorar ou a tratar como desvio.

As limitações desta análise precisam ser reconhecidas. A restrição ao corpus bibliográfico em língua portuguesa pode ter deixado de incorporar contribuições relevantes da literatura internacional em inglês, especialmente pesquisas experimentais sobre o comportamento de LLMs em tarefas psicolinguísticas específicas. O caráter bibliográfico e documental da pesquisa impede inferências causais diretas e exige cautela na generalização das conclusões. Essas limitações apontam para a necessidade de estudos experimentais que testem diretamente as hipóteses geradas por esta análise, comparando o desempenho de LLMs e de crianças em tarefas psicolinguísticas equivalentes com variedades do português brasileiro.

5 CONSIDERAÇÕES FINAIS

Este estudo se propôs a analisar as diferenças estruturais entre a aquisição natural da linguagem pela criança, fundamentada na hipótese da pobreza do estímulo, e o treinamento de LLMs sobre grandes corpora, examinando as implicações dessas diferenças para a compreensão dos processos linguísticos humanos e para a pesquisa em psicolinguística, linguística computacional e ensino de línguas no Brasil.

A síntese dos resultados demonstra que a aquisição da linguagem pela criança é um processo cognitivo, interacional e biologicamente fundamentado, que produz competência linguística plena a partir de um input quantitativamente restrito, porém qualitativamente rico e contextualmente situado. Essa competência envolve a construção de representações mentais abstratas das regras gramaticais da língua, a capacidade de generalização criativa para estruturas não ouvidas, e a integração de dimensões fonológicas, sintáticas, semânticas e pragmáticas em um sistema unificado. O treinamento de LLMs,

por sua vez, opera sobre volumes de dados incomparavelmente maiores, mas sem acesso aos mecanismos cognitivos inatos, às interações sociais contextualizadas e às dimensões prosódicas e pragmáticas que tornam a aquisição humana possível.

A hipótese orientadora da pesquisa foi confirmada pela análise: os LLMs, apesar de operarem sobre corpora de escala massiva, não replicam os mecanismos cognitivos inatos e os processos interacionais que tornam a aquisição humana da linguagem possível a partir de estímulos quantitativamente pobres. A confirmação dessa hipótese tem implicações que vão além do debate técnico sobre arquiteturas de modelos de linguagem: ela interpela a forma como pesquisadores, educadores e formuladores de políticas públicas compreendem e utilizam os LLMs em contextos relacionados ao desenvolvimento linguístico humano.

As contribuições deste estudo organizam-se em dois planos. No plano teórico, a pesquisa oferece um quadro analítico sistemático que distingue, com precisão epistemológica, os processos de aquisição natural da linguagem e o treinamento computacional de LLMs, articulando a hipótese da pobreza do estímulo com as perspectivas psicolinguísticas sobre o papel do input qualificado e da interação social no desenvolvimento linguístico. Esse quadro analítico contribui para o debate sobre a natureza da linguagem e sobre os limites das abordagens estatísticas na modelagem de capacidades linguísticas humanas.

No plano prático, a pesquisa oferece subsídios para profissionais que utilizam ou avaliam LLMs em contextos educacionais, terapêuticos e de políticas linguísticas. O mapeamento das diferenças entre aquisição natural e treinamento computacional permite que esses profissionais estabeleçam expectativas mais precisas sobre o que os LLMs podem e não podem fazer no que diz respeito à linguagem humana, evitando usos que pressuponham equivalência entre desempenho computacional e competência linguística.

As limitações da pesquisa precisam ser enunciadas com clareza. A restrição predominante ao idioma português nas buscas bibliográficas pode ter deixado de incorporar contribuições relevantes da literatura experimental internacional, especialmente estudos que testam o comportamento de LLMs em tarefas psicolinguísticas controladas. O caráter bibliográfico e documental da pesquisa não permite inferências causais diretas e exige que as conclusões sejam tomadas como hipóteses a serem testadas empiricamente, não como generalizações absolutas. A ausência de dados experimentais primários constitui uma limitação inerente ao método adotado, que precisa ser explicitada para que os resultados sejam interpretados no contexto adequado.

As sugestões para estudos futuros derivam diretamente dessas limitações. A realização de estudos experimentais que comparem o desempenho de LLMs e de crianças em tarefas

psicolinguísticas equivalentes, incluindo tarefas de julgamento de gramaticalidade, generalização a estruturas raras e processamento de ambiguidades, permitiria testar empiricamente as hipóteses geradas por esta análise bibliográfica. Pesquisas que examinem especificamente o comportamento de LLMs com variedades do português brasileiro não padrão, comparando-o com o comportamento de falantes nativos dessas variedades, contribuiriam para o debate sobre diversidade linguística e representatividade dos corpora de treinamento.

Estudos longitudinais que acompanhem o impacto do uso de LLMs em contextos de intervenção fonoaudiológica e educacional com crianças com alterações de linguagem constituem outra agenda de pesquisa prioritária. A literatura revisada documenta que o desenvolvimento linguístico infantil é sensível a variáveis interacionais e contextuais que os LLMs não replicam, mas não há ainda evidências sistemáticas sobre os efeitos do uso desses modelos sobre a trajetória aquisicional de crianças em diferentes condições de desenvolvimento.

A reflexão final sobre o impacto deste trabalho situa-se na interface entre a pesquisa linguística e o debate social sobre a IA. A tendência de antropomorfizar os LLMs, atribuindo-lhes capacidades linguísticas equivalentes às humanas, não é apenas uma imprecisão conceitual: ela tem consequências práticas sobre como esses sistemas são usados, regulados e avaliados. Reconhecer que a criança adquire a linguagem a partir de mecanismos que os LLMs não possuem é reconhecer que a linguagem humana é um fenômeno que transcende o processamento estatístico de texto, por mais sofisticado que esse processamento seja. Essa conclusão não diminui o valor dos LLMs como ferramentas tecnológicas, mas delimita com precisão o que eles são e o que não são, contribuindo para um debate público e científico mais fundamentado sobre o papel da IA na linguagem, na educação e na sociedade.

REFERÊNCIAS

Brait, B. *et al.* Bakhtiniana adere à Ciência Aberta. **Bakhtiniana: Revista de Estudos do Discurso**, São Paulo, v. 17, n. 1, p. 2-15, 2022. DOI: 10.1590/2176-457356035.

Cruzatti, A. L. *et al.* **Produção da fala de crianças e adolescentes de um programa de reabilitação auditiva**. *Audiology – Communication Research*, São Paulo, v. 27, 2022. DOI: 10.1590/2317-6431-2021-2615.

Dias, N.; Anache, A. A.; Maciel, R. F. **Os limites e contradições da educação bilíngue para estudantes surdos**. *Revista de Ensino, Educação e Ciências Humanas*, Londrina, v. 21, n. 1, p. 47-54, 2020. DOI: 10.17921/2447-8733.2020v21n1p47-54.

Freitas, C. A. de. **A importância do desenho universal na aprendizagem para a inclusão escolar**. *Lumen et Virtus*, [S.l.], 2025. DOI: 10.56238/levv16n55-118.

Gomes, R. P. *et al.* **A escrita simples como estratégia de acessibilidade para a divulgação científica**. *Interfaces Científicas – Humanas e Sociais*, Aracaju, v. 9, n. 2, p. 215-228, 2021. DOI: 10.17564/2316-3801.2021v9n2p215-228.

Goulart, I. do C. V.; Ferreira, S. C. C. **Leitura em ambiente digital: contribuições do aplicativo "Bamboleio" em contexto remoto**. *Ensino em Re-Vista*, Uberlândia, v. 29, e040, 2022. DOI: 10.14393/er-v29a2022-40.

Guimarães, L. C. **Saúde mental da pessoa surda no Brasil**. *RECIMA21 – Revista Científica Multidisciplinar*, [S.l.], v. 2, n. 3, p. 316-325, 2021. DOI: 10.47820/recima21.v2i3.173.

Macedo, L. M. M. A. *et al.* **Dislexia e transtorno do desenvolvimento da linguagem são quadros isolados ou comórbidos? Uma revisão integrativa**. *Revista CEFAC*, São Paulo, v. 24, n. 3, 2022. DOI: 10.1590/1982-0216/202224312021s.

Marchetti, P. T. *et al.* **Comparação entre a terapia fonológica isolada ou associada ao treinamento auditivo computadorizado na reabilitação dos transtornos fonológicos em crianças**. *Distúrbios da Comunicação*, São Paulo, v. 33, n. 3, p. 545-556, 2021. DOI: 10.23925/2176-2724.2021v33i3p545-556.

Martins, A. *et al.* **Desenvolvimento de um ambiente virtual de aprendizagem sobre linguagem infantil para pediatras**. *Distúrbios da Comunicação*, São Paulo, v. 33, n. 4, p. 741-750, 2021. DOI: 10.23925/2176-2724.2021v33i4p741-750.

Pinto, M. M.; Santos, L. F. dos. **Concepções de professores de alunos surdos sobre inclusão e educação bilíngue**. *Revista de Educação PUC-Campinas*, Campinas, v. 27, p. 1-12, 2022. DOI: 10.24220/2318-0870v27e2022a5726.

Santos, A. C. E. G. dos; Aleixo, L. C. M. Z.; Santos, A. E. dos. **Perfil dos atendimentos fonoaudiológicos em linguagem no município de São Tomé/RN**. *Revista Científica Multidisciplinar Núcleo do Conhecimento*, [S.l.], p. 71-100, 2022. DOI: 10.32749/nucleodoconhecimento.com.br/saude/atendimentos-fonoaudiologicos.

Santos, P. S. dos; Cavalcante, E. B. **A interação da criança surda a partir da Teoria Histórico-Cultural: possibilidade de desenvolvimento**. *Revista Educação Especial*, Santa Maria, 2022. DOI: 10.5902/1984686X67952.

Sardinha, T. B. *et al.* **Portal multimodal/multilíngue para o avanço da Ciência Aberta nas Humanidades**. *Cadernos de Linguística*, [S.l.], v. 2, n. 4, e406, 2021. DOI: 10.25189/2675-4916.2021.v2.n4.id406.

Silva, E.; Elias, N. C. **Interdependência entre o desenvolvimento de habilidades sensório motoras e da linguagem no TEA: revisão sistemática**. *Revista Eletrônica de Educação*, São Carlos, v. 17, e5677071, 2023. DOI: 10.14244/198271995677.

Silva, G. M. **Contribuições da Psicolinguística para a Educação de Surdos**. *Letrônica*, Porto Alegre, v. 16, n. 1, e44247, 2023. DOI: 10.15448/1984-4301.2023.1.44247.

Souza, M. S. de L.; Cáceres-Assenço, A. M. **Prática baseada em evidências e atuação clínica em linguagem infantil: uma pesquisa online com fonoaudiólogos brasileiros**. *CoDAS*, São Paulo, v. 36, n. 1, 2024. DOI: 10.1590/2317-1782/20232022272pt.