

Estado da publicação: O preprint foi publicado em outro meio.

DOI do preprint publicado: <https://doi.org/10.25189/2675-4916.2026.v7.n3.id925>

# IA generativa e construção de sentido: abordagem semiótica a partir do Sul Global

Letícia Moraes, Silvia Maria de Sousa

<https://doi.org/10.1590/SciELOPreprints.13920>

Submetido em: 2025-10-29

Postado em: 2025-11-10 (versão 1)

(AAAA-MM-DD)

## **IA generativa e construção de sentido: abordagem semiótica a partir do Sul Global**

Letícia Moraes, Departamento de Língua Portuguesa e Linguística, Universidade Federal da Paraíba, Brasil <https://orcid.org/0000-0002-4642-5974>

Silvia Maria de Sousa, Departamento de Ciências da Linguagem, Universidade Federal Fluminense, Brasil. <https://orcid.org/0000-0003-4069-5778>

### **Resumo:**

A popularização dos modelos de Inteligência Artificial (IA), especialmente da IA generativa (IAGen), a partir da década de 2010, somada à circulação algorítmica de dados, tem provocado transformações em diversas esferas da sociedade. Nesse contexto, o artigo investiga os processos de geração de significação em modelos de IA generativa, buscando delimitar, sob uma perspectiva semiótica, o que vem a ser o *big data* e o processo de dataficação. Mobilizam-se, entre outros, os conceitos de *práticas semióticas* e *formas de vida* (Fontanille 2008; 2015), com o intuito de descrever como o processo intangível e dinâmico da dataficação agencia cenas práticas que, ao serem estabilizadas no *big data*, favorecem determinadas formas de vida. Considerando o controle de atores concentrados no denominado Norte Global sobre as tecnologias difundidas globalmente e o modo como se dão as operações entre big data, dataficação, algoritmos e IA, é possível explicar, por exemplo, a intensificação de preconceitos e estereótipos o denominado viés. Para adensar a discussão semiótica da IA a partir do Sul Global e do problema da sub-representação de dados, o estudo toma como exemplo textos gerados pelo Sabiá, modelo de IA desenvolvido no Brasil, comparando-o a um modelo globalizado, o ChatGPT. Os resultados ajudam a compreender como diferenças na constituição de datasets repercutem na geração de sentido. A pesquisa defende que a semiótica esteja implicada na compreensão e na construção de IAs mais éticas e culturalmente plurais, com vistas a reduzir as assimetrias estruturais entre Norte e Sul.

Palavras-chave: *Inteligência Artificial Generativa; Dataficação; Sul Global; Práticas Semióticas; Formas de vida;*

## **Generative AI and Meaning-Making: A Semiotic Approach from the Global South**

### **Abstract:**

The widespread adoption of Artificial Intelligence (AI) models—particularly Generative

AI (GenAI)—since the 2010s, along with the algorithmic circulation of data, has reshaped multiple dimensions of social life. This article examines the processes of construction of meaning in generative AI through the lens of discursive semiotics, focusing on how *Big Data* and *datafication* operate as semiotic regimes. Drawing upon the concepts of *semiotic practices* and *forms of life* (Fontanille 2008; 2015), the analysis describes how the intangible and dynamic process of datafication configures *practical scenes* that, once stabilized within Big Data, privilege particular *forms of life*. Given the concentration of technological control in the Global North, as well as the operative interactions among Big Data, datafication, algorithms, and AI, one can account, for instance, for the intensification of prejudices and stereotypes—the so-called bias. To expand the semiotic debate on AI from a Global South perspective—especially in light of the structural underrepresentation of data—the study compares texts generated by *Sabiá*, a Brazilian AI model, with those produced by the globally disseminated model ChatGPT. The findings demonstrate how variations in dataset constitution reverberate in processes of construction of meaning. The study argues that semiotics must be implicated in both the comprehension and the construction of more ethical and culturally plural AIs, in order to reduce structural asymmetries between North and South.

**Keywords:** *Generative AI; Datafication; Global South; Discursive Semiotics; Semiotic Practices; Forms of Life*

### **Considerações iniciais**

As investigações contemporâneas no campo da computação social e das humanidades digitais vêm considerando o *big data* como um complexo repositório de artefatos culturais produzidos, coletados e organizados a partir de ações humanas cotidianas. Essa compreensão, presente nos trabalhos das pesquisadoras da Microsoft Research, Danah Boyd e Kate Crawford (2012), desloca o enfoque meramente técnico-operacional comumente atribuído aos dados para uma abordagem que reconhece sua constituição sociocultural e política. Por meio dessa concepção, os dados deixam de ser tomados como entidades brutas ou neutras, ao pôr em destaque que eles resultam de escolhas metodológicas e de decisões institucionais que refletem e reproduzem valores culturais e estruturas de poder e conteúdos ideologicamente determinados.

Não são raras as pesquisas que assumem como objeto de análise ou como mote

de discussão os efeitos pragmáticos e discursivos do uso de dados e algoritmos na cultura contemporânea. No artigo "Datafication, dataism and dataveillance: *big data* between scientific paradigm and ideology", José van Dijck (2014), chama a atenção para a necessidade de uma crítica inter/transdisciplinar aos sistemas de big data que tenha um olhar para além da engenharia e da economia. Em diálogo com essas mesmas preocupações, mas através de uma perspectiva ético-política, a geógrafa Dra. Louise Amoore (2020), da Durham University, introduz a expressão "Cloud ethics", que intitula o seu livro mais conhecido. Para ela:

a ética da nuvem está atenta a como os algoritmos passam a configurar os atributos de nós mesmos e dos outros, e ao que significa viver em um mundo onde tais atributos são ativamente inferidos e utilizados como base para ações.(Amoore 2020, p. 4)<sup>1</sup>.

Inspirada pelos escritos de filósofos(as) como Michel Foucault e Judith Butler, Amoore propõe uma reflexão sobre como os algoritmos produzem axiologias identitárias de si e dos outros, participando da constituição da subjetividade, da alteridade e do julgamento moral em sociedades dataficadas. Interrogações como "quem deve ser responsabilizado se um ataque de drone, operado por algoritmos, resultar na morte de civis?" ou "como responder eticamente se uma cirurgia conduzida por inteligência artificial levar ao óbito de um paciente?" são formuladas pela autora para expor de forma contundente a necessidade de uma reflexão sobre os princípios éticos que estão em jogo num mundo mediado por sistemas de inteligência artificial, algoritmos e dados.

Nessa breve reunião inicial de reflexões interdisciplinares, cujo objetivo é demonstrar como as transformações ocorridas em diversas esferas da sociedade encontram eco em múltiplos campos do conhecimento, retomemos, ainda, o pensamento do antropólogo Nick Seaver (2018) ao sugerir que os dados sejam abordados pelos antropólogos como objetos etnográficos, à semelhança de qualquer artefato cultural complexo, o que requer atenção à multiplicidade de vozes, às disputas discursivas e às práticas implicadas em sua constituição. À luz das contribuições de Boyd e Crawford (2012), van Dijck (2014), Amoore (2020) e do próprio Seaver (2018), é possível afirmar que os dados não apenas espelham a realidade cultural, mas constituem parte integrante

---

<sup>1</sup> Original: "Cloud ethics is attuned to how algorithms come to figure the attributes of ourselves and of others, and what it means to live in a world where such attributes are actively inferred and acted upon."

da semiosfera. Longe de propor discussões estritamente a respeito da natureza ontológica dos dados, tais reflexões revelam os modos como os sentidos passam a ser gerados e a disputar novos espaços (concretos, virtuais e simbólicos) de circulação em sociedades dataficadas. De um ponto de vista semiótico, nota-se que por meio das tecnologias digitais - e dos objetos gerados por elas - passam a ganhar sentido dadas ações individuais, condutas éticas, organizações coletivas e, mesmo, serem criadas configurações axiológicas de gosto estético.

Essa discussão nos encaminha para pôr em questão a natureza epistemológica dos dados sob a ótica de uma semiótica "implicada".<sup>2</sup> Para nós, semioticistas, o que são, afinal, os dados? Como apreendê-los a partir de um enquadramento conceitual próprio à disciplina? Em que nível de pertinência analítico alocá-los? De que maneira lógica de produção, estocagem, organização e circulação de dados influencia na geração de sentidos? Para tentar responder a essas perguntas, o artigo é dividido em três partes. A primeira abarca uma discussão semiótica do conceito de dataficação, estabelecendo uma relação com os conceitos de *práticas e formas de vida* (Fontanille 2008; 2015) e é seguida pelo exame da problemática do enviesamento de dados no Sul Global. Busca-se dar um contorno semiótico robusto para esses dois aspectos. Por fim, a terceira parte do artigo, compreende uma breve análise exemplificativa com base em respostas geradas por modelo de IA regional desenvolvido no Brasil em comparação com respostas do ChatGPT.

## 1. Big data e dataficação

Compreendidos como extensas e complexas coleções de objetos semióticos manifestados, no plano da expressão, por meio de múltiplas linguagens e códigos — verbais, verbovisuais, visuais, matemáticos, sincréticos, entre outros, os dados não constituem realidades tangíveis do mundo natural. Diferente das pedras, eles não são elementos brutos: esses objetos são sempre "algo" modelado. Ao qualificá-los como objetos semióticos, recorreremos ao pensamento de Marrone (2008, s.p), para quem um objeto semiótico é "um simulacro, qualquer coisa que deve ser preparada". Assim, para que um artefato seja reconhecido como um dado é indispensável a instauração de uma "análise", conforme apontou Hjelmslev (2010 [1975]). Foi ainda por meio do pensamento hjelmsleviano que a semiótica atentou para a necessidade de recortar e inscrever a

---

<sup>2</sup> Sobre a semiótica "implicada", cf. (Aldama et al. 2021); e (Schwartzmann; Silva 2022).

semiótica-objeto dentro de um sistema de significação, pois assim é possível a descrição formal das relações internas estabelecidas por ele.

Para apreender os dados semioticamente, bem como delimitar como a sua produção e circulação produzem sentidos, abordaremos a questão da dataficação, sobretudo a partir dos estudos de Jacques Fontanille (2008; 2015) a respeito das *práticas e formas de vida*. Para localizar tais conceitos na epistemologia da semiótica, retomemos a explicação de Sousa (2023):

[...] em suas primeiras formalizações, a semiótica estava interessada em analisar o *fazer* dos sujeitos narrativos observando como as modalidades do tipo *querer, dever saber e poder* os conduziam na direção de um objeto, denominado teoricamente *objeto de valor*. Com o avanço da teoria e o enfrentamento de novos desafios, a análise passou a incorporar a dimensão cognitiva e sensível dos discursos, analisando o efeito das modalidades sobre o *ser* dos sujeitos. As categorias metodológicas passaram a integrar, cada vez mais, a análise das graduações, investigando *paixões*, operações da *práxis enunciativa* e, mais recentemente, *modos de interação e práticas semióticas*. (p. 66- grifo da autora)

Passemos a observar o processo da dataficação à luz da noção de práticas semióticas.

### 1.1 A dataficação: *big data* e práticas semióticas

O processo de modelagem de dados faz parte de um processo mais amplo, conhecido como dataficação. Essa expressão cunhada, em 2013, por Mayer-Schoenberger da Universidade de Oxford e pelo jornalista Kenneth Cukier, ambos militantes no campo da governança e regulação da internet, designa uma sequência finita de ações executáveis — algoritmos — que avaliam o desempenho de dados previamente coletados e tomam decisões a partir de regras estabelecidas (Mayer-Schoenberger; Cukier 2013, p. 28). Tal definição técnica, advinda da área da computação, tende a silenciar aspectos fundamentais para as teorias das ciências humanas. Ao figurativizar uma suposta neutralidade tecnológica, mascaram-se os vieses implicados em tais processos técnicos, especialmente quando examinados sob a perspectiva epistemológica do "Sul Global"<sup>3</sup>. É possível interrogar o caráter técnico ao perguntar como as decisões são tomadas e como se dá o estabelecimento de regras. Ou ainda que valores axiológicos as orientam e as

---

<sup>3</sup> A divisão entre Sul e Norte Global tem suas raízes na geopolítica e na economia com as expressões "Terceiro mundo" (Alfred Sauvy, 1952) e "países em desenvolvimento". Entre as décadas de 1980 e 1990, a expressão "Global south" começou a aparecer nos debates das Nações Unidas, bem como a ideia de divisão entre norte e sul, como aparece no Relatório Brandt (North-south: A programme for Survival, 1980) para enfatizar as desigualdades entre ambos os eixos.

sustentam. Tais reflexões são essenciais quando temos a intenção de compreender a geração e a circulação de dados.

A dataficação não pode ser reduzida a um mero processo técnico de codificação da realidade: ela é, antes, um dispositivo semiótico que converte práticas sociais em dados quantificáveis para alimentar ecossistemas algorítmicos cuja lógica e infraestrutura permanecem, em grande medida, sob controle de atores concentrados no denominado Norte Global (Mejias; Couldry, 2024). Ressaltamos que não apenas as regras de processamento, mas os próprios valores discursivos concretizados no conteúdo dos dados - compostos por toda sorte de gêneros textuais diversos, imagens, áudios, códigos - influenciam na formatação de formas de vida - que deveriam ser locais - mas que passam a ser instauradas/atravessadas por um processo de modelização que achata configurações de sentidos próprios a cada cultura, contribuindo para a normalização de epistemologias coloniais, intensificando a marginalização saberes e experiências não hegemônicas.

No âmbito dos estudos de uma semiótica “implicada” (Aldama *et al.*, 2021)<sup>4</sup>, definimos a dataficação como arranjo sintagmático composto de diferentes cenas práticas que, em um primeiro momento, convertem certas ações e gestos humanos em dados quantificáveis, para, em seguida, serem processados por algoritmos de sistemas de inteligência artificial, finalmente, usados para dar sentido às *práticas e formas de vidas* (Fontanille, 2008; 2015) de sujeitos inseridos nessa mesma sociedade datafificada, de onde dados foram originalmente extraídos. O aspecto circular da dataficação, impiedoso nele mesmo, torna-se ainda mais problemático quando se expõe que os dados originais dizem respeito a uma pequena parcela do mundo, mas alcançam uma escala global após a dataficação.

Em *Pratiques sémiotiques*, Fontanille (2008, p. 4) propõe uma hierarquização dos níveis de pertinência<sup>5</sup> da análise semiótica e descreve as cenas práticas como arranjos sintagmáticos de comportamentos e ações humanas, dotados de um plano da expressão e de um plano do conteúdo. As práticas, observa o autor, são “cursos de ação” (2019, p. 251) abertos, cuja significação ocorre no desenrolar sintagmático entre um começo e um fim. Diferentemente de objetos semióticos inscritos em suportes estáveis, como uma fotografia, práticas semióticas produzem sentido *ao vivo*, isto é, enquanto se realizam,

---

<sup>4</sup> Mantemos o termo em parênteses, pois acreditamos que toda teoria semiótica é, ou deveria ser, implicada.

<sup>5</sup> Jacques Fontanille propõe uma ampliação da imanência semiótica, por meio de seis níveis: (1) Signos, (2) Textos-enunciados, (3) Objetos, (4) Cenas práticas, (5) Estratégias, (6) Formas de vida. Cada nível seria descrito enquanto instância formal, evocando um tipo específico de experiência e sendo composto por uma instância material.

permitindo que o próprio movimento prático seja um elemento gerador de sentido e, portanto, a significação passa a ser dinâmica.

O *big data* constitui, assim, um desdobramento, uma resultante, digamos, mais estabilizada desse processo mais intangível e dinâmico da dataficação (Moraes, 2024), ao fornecer a matéria-prima para que sistemas de inteligência artificial façam previsões baseadas em cálculos matemáticos e estatísticos (Jay Agrawal; Goldfarb, 2018). Assim, não é difícil notar que o *big data*, ainda que caracterizado por materialidade codificada e processo enunciativo não transparente, concretize linhas isotópicas ancoradas em vieses ideológicos colonialistas e cosmovisões do Norte.

A dataficação, reconfigura o campo de discurso ao converter cenas práticas, cursos de ações diversas - comprar, gesticular, circular, acessar, pesquisar - em objetos semióticos manipuláveis. Isso possibilita a predição de ações e comportamentos tanto quanto a reorganização das configurações de sentidos que emergem das interações humanas. Nessa dinâmica, o dado perde sua condição de vestígio arquivado, ou seja, testemunho de um fazer passado, e assume a função de programação, atuando como suporte material na modelização probabilística de futuros possíveis dentro de um quadro de leitura realizado por sistemas de IA. Esse processo de passagem de atos vividos a dados, incide na alteração de dinâmicas estabelecidas entre os modos de existência, já que na passagem da virtualidade à atualização, incide uma espécie de enunciação algorítmica, co-enunciada por actantes humanos e não humanos.<sup>6</sup> Como trata-se de processo circular, veloz, contínuo, não transparente e de alcance tentacular é cada vez mais difícil precisar o que é característico dos dados de entrada (dados puros – cenas vividas) e de saída (datafizados), sem mencionar a dificuldade crescente em delimitar os papéis e posições dos sujeitos sociais no decorrer desse processo. Essa espécie de nebulosa, que embaça a dataficação, está na base dos problemas relacionados à transparência, responsabilidade e ética em IA. Iniciativas mundiais trabalham na formulação de políticas em favor de uma “IA explicável” e da manutenção da dignidade humana em face do uso da IA (Coeckelbergh 2023, p.143), com as quais a semiótica deve contribuir.

A compreensão do processo de dataficação permite-nos definir, com maior precisão, o *big data* como um grande conjunto de coleções de objetos semióticos gerados pelo desencadeamento de práticas humanas, cuja circulação e apropriação obedecem às lógicas estruturais de poder, como demonstraremos na segunda seção deste artigo.

---

<sup>6</sup> D’Armenio et al. (2024) fazem uso da expressão co-enunciação para designar o modo de funcionamento dos modelos de IA generativa, salientando a participação humana e maquina na produção da enunciação.

Importa reconhecer que boa parte dos objetos semióticos gerados por dataficação inicia o processo de sentido *in actu*, no desenrolar de uma cena prática: esse é o caso de uma transação monetária realizada usando cartão de crédito no ato da compra de um produto no supermercado. Tal cena prática, embora não estabilizada integralmente em suporte físico, gera objetos semióticos passíveis de estabilização (a nota fiscal eletrônica, a lista de itens comprados por indivíduo, o CPF informado<sup>7</sup>, etc.), que passam a integrar grandes coleções de objetos semióticos, portadores de novo plano de expressão e de conteúdo: o *big data*. Tais coleções são caracterizadas pela incomensurabilidade, heterogeneidade e sobreposição de linguagens (verbal, áudio, visual), de gêneros, tipos textuais, assuntos, áreas do conhecimento, além de restrições de acesso e sigilo imposto pelas empresas detentoras de banco de dados. Isso contribui para a perda de noção de origens autorais, fiscais, temporais, organizacionais e para a dificuldade em localizar a assunção enunciativa, afetando a identificação sobre quem fala e sob que condições.

Salientamos a estreita articulação entre a geração de dados, inicialmente extraídos de ações humanas e hoje já gerados automaticamente por IAs, e o processo de dataficação porque, do ponto de vista semiótico, este último opera como uma espécie de textualização (Greimas; Courtés 2008 [1979], p. 504) das ações e dos comportamentos humanos inscritos nas cenas práticas e, cada vez mais, emulados por IAs. Antes da consolidação das tecnologias de inteligência artificial associadas ao *big data*, muitos gestos e traços da experiência cotidiana não eram sistematicamente registrados em suportes estáveis, o que limitava a possibilidade de captá-los e analisá-los em tempo real, extraíndo, assim, os padrões recorrentes, que são essenciais para os modelos preditivos que antecipam possíveis comportamentos de modos coletivos da organização social.

A passagem de um suporte não (ou menos) estável para um outro mais estabilizado ocorre pelo processo de dataficação. O suporte, nesse caso, cumpre um papel mediador decisivo: ele permite compreender a dataficação como um tipo de textualização, isto é, “um conjunto de procedimentos — chamados a se organizarem numa sintaxe textual — que visam à constituição de um contínuo discursivo, anteriormente à manifestação do discurso nesta ou naquela semiótica” (Greimas; Courtés 2008 [1979], p. 503). No caso da dataficação, esse contínuo se realiza na modulação de ações em dados quantificáveis, de maneira que os gestos humanos, inscritos em uma cena prática, antes

---

<sup>7</sup> No Brasil, é comum que os supermercados e outros comércios peçam o número de documento de identidade nacional do comprador.

efêmeros, passem a existir como objeto semiótico manipulável.

Na última década, a expansão do uso do *big data* não se traduziu apenas na intensificação de sua aplicação para fins de predição de comportamentos e mapeamento de *formas de vida*, mas também para uma reconfiguração profunda de sua gestão e estocagem. Dados que, anteriormente, encontravam-se sob controle majoritário de entidades estatais — nacionais ou regionais — passaram progressivamente a ser administrados por corporações privadas transnacionais. Paralelamente, observa-se uma ampliação substancial do escopo de conteúdos armazenados: para além de registos administrativos e demográficos, passaram a integrar essas bases informações capazes de revelar preferências pessoais de consumo, trajetos cotidianos, posicionamentos ideológicos e políticos, redes de contatos sociais, para citar algumas.

Essa mudança assume contornos ainda mais problemáticos quando analisada a partir de uma perspectiva crítica. Práticas locais são textualizadas em dados que alimentam modelos construídos sob a égide de axiologias hegemônicas do Norte. Inseridas em um processo mais amplo, o de dataficação, esses modelos operam como mecanismos de padronização e normatização das práticas e *formas de vida*, impondo recortes de pertinência que frequentemente desconsideram — ou até mesmo apagam — as especificidades culturais, históricas e políticas das populações de onde os dados foram extraídos, confirmando o processo de perda das origens a que nos referimos.

Assim, mais do que incidir sobre condutas individuais isoladas, importa-nos compreender de que maneira esse processo intervém na configuração de formas de vida, reorientando práticas, produzindo sentidos e instituindo novos regimes de organização coletiva nas sociedades dataficadas.

### *1. 2. A dataficação: do big data às formas de vida*

Uma *forma de vida* designa o tipo mais abrangente de semiose já identificado, definida como modos coletivos de organização social e cultural que portam e produzem sentido. Elas são formas sintagmáticas da vida (Fontanille 2015, p. 43) e possibilitam a apreensão de uma experiência elementar, genérica e coletiva: a própria vida — ou o curso que ela assume (Fontanille 2017, p. 73). No último nível da análise semiótica, as formas de vida permitem a compreensão da experiência das práticas (Fontanille, 2008), estas estruturadas por um *ethos* e por normas sociais, e nas quais múltiplos objetos semióticos convergem, de maneira orquestrada, para gerar o sentido da vida. Trata-se, portanto, de

uma articulação semiótica entre ethos, normas e práticas que estruturam uma semiosfera.

A existência dos sujeitos sociais só é possível no interior das *formas de vida*. Ao incorporar um quadro axiológico, o sujeito reconhece-se nos valores que estruturam a semiosfera, constituindo-se como tal. É nesse espaço coletivo que “os sujeitos e os grupos sociais são os protagonistas das interações e das trocas socioculturais” (Fontanille 2015, p. 247). Longe de reduzirem a simples padrões comportamentais, as *formas de vida* configuram modos de ver, ler, organizar e habitar o mundo. Segundo Fontanille Zilberberg (2001 [1998]), em *Tensão e Significação*, no capítulo dedicado às formas de vida, “toda manifestação de uma forma de vida é então considerada como um condensado da forma de vida inteira” (Fontanille; Zilberberg, 2001 [1998], p. 204) e sua apreensão, no entanto, só se realiza pela relação, pois é no confronto e na coexistência com outras formas de vida que elas se tornam compreensíveis (Fontanille, 2015, p. 59).

Embora possam sustentar narrativas identitárias, as *formas de vida* não se deixam analisar no plano estritamente individual, uma vez que sua natureza é essencialmente coletiva. Elas delineiam e sustentam um percurso existencial para o sujeito no interior do corpo social, conferindo-lhes congruência para a existência sintagmática, ao mesmo tempo em que articulam um sistema de valores inerente ao sentido da vida do sujeito. Ao relacionarmos esse conceito com a discussão da dataficação, podemos compreender como a análise de dados, realizada por modelos de inteligência artificial, algoritmos e estatísticas, facilita que as empresas detentoras de dados identifiquem os sentidos inscritos e suas saliências nas organizações coletivas, de um lado, e, de outro, possam agir influenciando os esquemas de seleção de conteúdos categoriais dominantes, a discursivização (as figuras e temas) e valorização que esquematizam determinadas *formas de vida*.

No que diz respeito à análise, diferentes actantes econômicos coletivos mobilizam o *big data* para orientar decisões: instituições financeiras estimam a probabilidade de inadimplência de indivíduos ou grupos; redes sociais veiculam anúncios direcionados a partir de perfis comportamentais; operadoras de planos de saúde ajustam mensalidades com base em históricos de consumo médico; e plataformas de streaming, como Spotify, Netflix ou Amazon Prime Video, filtram e recomendam conteúdos personalizados a partir do uso de cada usuário. Mais do que esquematizações de comportamentos, essas análises permitem a alteração de esquematizações de sentidos que conferem a congruência e a coerência de dadas *formas de vida*. É curioso, entretanto, que quanto mais, constroem-se *formas de vida* controladas, geradas no e pelo *big data*, mais, projetam-se sentidos de

vida sem forma e sem controle, ao não sabermos a quem, onde ou quando responsabilizar em casos de veiculação de conteúdos de desinformação, crimes cibernéticos, fraudes, abusos no mundo do trabalho, por exemplo.

Como dissemos, quando articulado às tecnologias de inteligência artificial, o *big data* opera ativamente na modulação de práticas sociais e de formas de vida. Plataformas mundiais de *streaming*, como a Netflix, sugerem conteúdos para um usuário ou grupo de usuários. Esses sujeitos estabelecem uma interação, na práxis enunciativa, com aquele objeto semiótico que, sem tal interferência direta, possivelmente não aconteceria. Aqui chamamos atenção para o fato de que repetição dessa dinâmica não apenas reforça certos hábitos de consumo, para além da configuração de gostos e estilos individuais, mas age modelizando as organizações coletivas sedimentadas, alterando os regimes de sentido selecionados nos pólos de uma categoria e lhes atribuindo valores. Esse processo, aparentemente desprezioso, descreve a atuação silenciosa e incoativa da dataficação na geração de sentido das organizações coletivas.

Tal fenômeno é ilustrado nos resultados do projeto *Selfiecity*, de Lev Manovich. A equipe do MediaLab analisou um corpus de mais de três mil selfies publicadas no *Instagram*, provenientes de cinco cidades — São Paulo, Bangkok, Moscou, Nova Iorque e Berlim —, com o objetivo de identificar padrões e variações na linguagem da fotografia digital, considerando aspectos como cromatismo, uso de filtros, incidência e distribuição de luz e sombra, gestualidade e expressividade afetiva. Na análise estatística das categorias “inclinação da cabeça” e “biquinho” dos lábios, correlacionadas ao marcador de gênero feminino/masculino, os resultados (Cf. Titenfale e Manovich, 2015) indicaram recorrência e valorização desses gestos como padrões estéticos transnacionais, com particular incidência entre mulheres.

A recorrência massiva desse padrão imagético, reforçada por plataformas cujos algoritmos favorecem sua visibilidade, contribui para sua naturalização. Gestos que, a princípio, poderiam ser percebidos como excêntricos ou como expressão de uma singularidade passam a ser reiterados a tal ponto que perdem o valor semântico de especificidade de um estilo individual. Essa padronização do gesto não apenas diferencia a prática do selfie segundo uma lógica binária de gênero nas redes sociais, como também intervém na conformação das *formas de vida*: inclinar a cabeça, sorrir mostrando os dentes superiores ou fazer “biquinho” tornam-se condutas reiteradas na prática cotidiana de autopublicação de mulheres nas redes sociais, sendo gradualmente assimiladas como traços esperados — e socialmente legitimados — da *forma de vida* associada à categoria

de feminino.

Como apontam Fontanille e Zilberberg (1998), as *formas de vida* não são imutáveis, pois respondem à práxis enunciativa, isto é, ao uso. Sendo assim, são a coerência e a congruência que asseguram a estabilização de uma forma de vida e o seu reconhecimento social. Diante do exposto, faz-se necessária a investigação acerca de que *formas de vidas* são manifestadas pelos objetos gerados pela IA. Que conteúdos categoriais e elementos discursivos são salientados nelas? E mediante essas discursivizações o que é preterido?

A recorrência e a conseqüente naturalização de padrões estéticos e comportamentais, como exposto na pesquisa de Tifentale e Manovich (2015), evidenciam um funcionamento mais profundo da lógica algorítmica: a de instaurar e propagar modelos hegemônicos de representação que, ao assumirem-se como isotopias dominantes, afastam-se das configurações de sentido produzidas localmente. Para Paveau (2021) os algoritmos podem ser definidos como “sequências de instruções que permitem a solução de problemas” (Paveau, 2021, p.39), triam informações, criando classificações e hierarquias que tornam as IAs bem mais do que ferramentas tecnológicas, já que “certas informações aparecerão com mais frequência, ou em melhor lugar do que outras, ou serão mais disseminadas do que outras, ou, pelo contrário, serão inviabilizadas” (Paveau 2021, p.39).

Em termos semióticos, trata-se de um processo de estabilização de formas de expressão e de conteúdo que, reiteradas e valorizadas pelo dispositivo técnico (modelos de inteligência artificial, algoritmos, datasets, etc.), passam a integrar o repertório normativo e coletivo no nível mais abrangente da imanência, isto é, das *formas de vida* (Moraes, 2024). Essa discussão coloca em evidência a necessidade de examinarmos o uso dos dados e seu enviesamento como um fenômeno discursivo, a partir da lente da semiótica discursiva, assumindo sua vocação para uma semiótica “implicada” (Aldama *et al.*, 2021; Schwartzmann e Silva, 2022).

## **2. O enviesamento de dados em IAs pelo olhar do Sul Global**

Como temos procurado evidenciar, o uso de algoritmos preditivos, orientados à previsão de comportamentos futuros, e de algoritmos generativos, voltados à criação de objetos semióticos inéditos, incide estatisticamente sobre a configuração do *big data*. Esse processo realiza uma operação de hierarquização na distribuição dos conteúdos,

privilegiando determinados padrões, de maneira que aquilo que se apresenta com maior frequência tende a ser reiterado e consolidado em uma “norma”. Tal automatismo de distribuição e combinação favorece manifestações de sentido próprias de culturas hegemônicas, dando mais produtividade a dadas línguas, modos de dizer, assuntos, imagens. Basta considerarmos a discrepância existente entre países geopoliticamente dominantes e o restante do mundo, no que diz respeito ao desenvolvimento de tais tecnologias, à extração de dados, às línguas usadas (Hagerty; Rubinov, 2019). Nesse cenário, países de língua inglesa, especialmente os Estados Unidos estão no topo da cadeia; como resultado, a maior parte dos objetos semióticos que compõem o big data e que serão processados pela IA são provenientes dessas comunidades.

Por outro lado, as práticas culturais não hegemônicas – tais como expressões linguísticas regionais, modos de sociabilidade afrodiáspóricas e epistemologias e cosmovisões indígenas – são marginalizados, por serem lidas pela máquina como singularidades que escapam à regularidade estatística internalizada pelos modelos. Ou seja, essas práticas e configurações de sentidos são compreendidas como desvios em relação à norma, consolidada pela frequência e repetição de formas presentes nos objetos semióticos constituintes do big data. Isso nos coloca diante de duas principais questões: (i) o caráter excludente do próprio modelo de funcionamento dessas tecnologias e (ii) a naturalização de hierarquias culturais na produção de significação mediada pela IA.

Como dissemos, em sociedades dataficadas, nas quais agentes maquínicos e humanos participam conjuntamente do processo de enunciação, configurando-se como co-enunciadores (D’Armenio *et al.*, 2024), a produção de sentido é mediada por operações estatísticas que influenciam a axiologização das isotopias. Na discursivização, o gesto maquínico, ao reiterar determinados arranjos figurativos e temáticos, tende a intensificar o estatuto de “universais”, instaurando-os como valores dominantes no campo discursivo. Em contrapartida, isotopias não hegemônicas, por não corresponderem às regularidades reiteradas nos objetos usados no treinamento, são consideradas desvios ou então ocorrências exóticas.

Os resultados podem ser observados em diferentes práticas mediadas por sistemas de inteligência artificial. A título de exemplificação, destacam-se: a figurativização estereotipada e homogeneizante de sujeitos oriundos do Sul Global em imagens geradas por IAs generativas de imagens (ex. DALL.E e MidJourney), como se todos compartilhassem vestimentas, hábitos e traços culturais idênticos (Ghosh *et al.*, 2025); a exclusão sistemática de línguas de países marginalizados nos processos de moderação

automatizada (Sahid *et al.*, 2025); a associação recorrente entre mulheres negras à temática da sexualização nos resultados de motores de busca (Noble, 2018); e a vinculação de profissões socialmente valorizadas, como a de médicos, à figurativização visual de pessoas brancas, desprovidas de marcas identitárias brasileiras, nos bancos de imagens amplamente utilizados no país.

Os exemplos acima incidem sobretudo no nível da discursivização, reiterando padrões figurativos e temáticos do Norte, de maneira a atribuir-lhes uma valorização eufórica, ao mesmo tempo em que a disforia é relacionada aos conteúdos semânticos do Sul.<sup>8</sup> Essas alterações nas categorizações e regimes de sentido são percebidas também na geração de sentido no interior de práticas novas e consolidadas, bem como em *formas de vida* em vias de ascensão ou desaparecimento

Nesse cenário, a pesquisa dos brasileiros Vilarino e Vicente (2020) lança luz sobre a maneira como o uso de IA preditivas por agências da área de finanças pode restringir o acesso ao crédito e o poder de compra de pessoas negras, uma vez que o *scoring* de crédito é baseado nos cálculos realizados pelo modelo e reproduz viés racial, mesmo quando esse dado não aparece explicitamente como um parâmetro. Isso acontece porque outras informações datafiscadas como localização geográfica e características socioeconômicas mantêm íntima relação com a raça no contexto cultural do Brasil, país marcado por profundas desigualdades sociais e raciais e que tem a segunda maior população negra do mundo. Ao privilegiar repertórios culturais dominantes, a IA não apenas invisibiliza identidades e experiências locais, mas também cria representações estereotipadas que funcionam como simulacros estabilizados de determinadas *formas de vida*.

Diante do cenário exposto, e concentrando nossa discussão mais especificamente no âmbito das IAs generativas, entendemos que a constituição dos *datasets* constitui um ponto crucial para a questão do enviesamento dos dados e o apagamento ou exotização de culturas e línguas de grupos minorizados. Em linhas gerais, os *datasets* são subconjuntos do *big data* e correspondem a coleções de objetos textuais, imagéticos, sonoros ou audiovisuais empregados no treinamento de modelos específicos. Sendo essas coleções, em grande parte, compostas por objetos semióticos em língua inglesa ou fortemente marcados pela cultura dos países do Norte Global, elas perpetuam uma nova modalidade de colonização por meio dos dados (Mejias & Couldry 2024). A desigualdade

---

<sup>8</sup> Sul Global não designa exatamente uma localização geográfica, mas o posicionamento epistemológico contra-hegemônico, propondo uma crítica radical à hegemonia preponderante de algumas culturas na ciência, na cultura, na tecnologia, etc.

manifesta-se tanto na quantidade quanto na qualidade dos dados disponíveis entre Norte e Sul, bem como na assimetria estrutural entre a produção tecnológica, concentrada nos países centrais, e a mera importação dessas tecnologias pelos países periféricos (Hagerty; Rubinov, 2019).

Assim, ao inscreverem-se no interior da lógica da dataficação, os *datasets* revelam-se não apenas como instrumentos técnicos, mas como operadores semióticos que selecionam, reiteram e legitimam determinados esquemas de sentido. Ao privilegiar conteúdos oriundos do Norte Global, essas coleções de objetos semióticos consolidam isotopias hegemônicas que se propagam como padrões universais de representação. Esse processo, sustentado pela congruência entre seleção, combinação e produtividade, estabiliza formas de vida marcadas por valores coloniais.

Como consequência, os *datasets* usados nos modelos generativos mais difundidos — como ChatGPT, DALL·E, Sora e MidJourney — contribuem para a sub-representação das figuras, temas, práticas e *formas de vida* que evidenciam as especificidades e a diversidade cultural dos povos do Sul Global, reforçando por assim dizer a reprodução de estereótipos e o apagamento semiótico dessas comunidades não-hegemônicas. Como possibilidade de mitigação dos problemas mencionados, diversos especialistas reforçam a necessidade da criação e do uso de *datasets* regionais no treinamento dos sistemas de IAs.

No tópico seguinte, para verificar se IAs supostamente treinadas com um *dataset* menos universalista e mais regional influencia na geração de sentido de objetos semióticos produzidos por modelos generativos, faremos uma análise comparativa entre resultados gerados por duas inteligências artificiais: o ChatGPT e o Sabiá.

### **3. A geração de sentido no big data (e datasets): do universal ao regional**

Nesta seção, apresentaremos o Sabiá como modelo de IA generativa regional, que, segundo dados disponibilizados pelas plataformas, foi criado no Brasil e treinado principalmente com dados brasileiros. Na análise, as respostas dadas por esse modelo serão contrastadas com respostas de um modelo difundido mundialmente, o ChatGPT, da empresa Open AI. Pretende-se identificar se, e em que medida, o modelo regional de IA rompe com lógicas hegemônicas de produção de sentido e favorece a construção de discursos mais éticos e culturalmente plurais.

O chatbot Sabiá foi lançado em outubro de 2022 pela empresa Maritaca IA. Segundo seus idealizadores, o desenvolvimento ocorreu em parceria com a Google, tendo

como principais características o seu treinamento em língua portuguesa e o *dataset* construído com foco no Brasil, o que, em tese, possibilitaria que o modelo tenha melhor desempenho, ao manejar dados mais específicos sobre a cultura, história e contexto desse país (Abonizio *et al.*, 2024). A Maritaca IA, também chamada de MariTalk, é a junção de parte da palavra em português "maritaca" com o verbo em inglês Talk. Maritaca é uma palavra do português brasileiro, originada em língua indígena tupi ("tupi mba'é taka") e nomeia uma espécie de aves bastante comuns nesse território. Diferentemente dos papagaios, as maritacas não imitam a vocalização de palavras humanas, mas possuem um canto estridente e ruidoso.

**Figura 1**



Primeira rolagem da página da Maritaca IA<sup>9</sup>.

Sabiá, palavra também de origem indígena tupi ("s-apia"), designa uma espécie de pássaros encontrados na América do Sul. Uma breve análise do modelo permite observar que os nomes escolhidos e apresentação visual, que tem como símbolo a imagem de uma maritaca, põe em destaque o tema da brasilidade e, com isso, marca-se uma diferença dos modelos IA globais mais difundidos. A escolha do nome, o cromatismo em tons de verde e azul, somados à imagem da maritaca, figurativizam o tema da brasilidade

---

<sup>9</sup> Disponível em: <https://www.maritaca.ai>  
Acesso. 8 ago. 2025

e apelam para aspectos considerados representativos do Brasil, como a fauna e a flora. A Maritaca AI põe em destaque, ainda, o slogan “Inteligência Artificial para o Brasil”, ancorando esse modelo de IA a um lugar específico. Não resta dúvidas de que se trata de um modelo regional.

O ChatGPT, por sua vez, é um chatbot desenvolvido pela OpenAI, lançado em novembro de 2022, concebido como uma ferramenta de alcance global, treinada a partir de extensas coleções de dados provenientes do *big data* mundial. O design visual neutro do ChatGPT, com fundo branco e um símbolo de formas abstratas curvas em preto e branco, remete ao tema da própria tecnologia e do fluxo das informações na IA. Não há vinculação do ChatGPT a uma determinada região, nem há qualquer traço de particularidade. Essa pretensa “universalidade”, entretanto, revela assimetrias: a distribuição e a densidade dos dados variam substancialmente entre regiões, com predominância de conteúdos oriundos de países de maior Produto Interno Bruto (PIB) e com *datasets* construídos majoritariamente em língua inglesa (Kaplunovich, 2023).

Para análise, selecionamos expressões linguísticas brasileiras, por considerarmos que o léxico constitui um dos subsistemas da língua que mais diretamente evidencia a articulação entre língua e cultura, uma vez que cabe às palavras a tarefa de designar os referentes pertencentes ao sistema extralinguístico. Como Vilela (1994, p.6) aponta:

O léxico é a parte da língua que primeiramente configura a atividade extralinguística e que arquiva o saber linguístico duma comunidade. Avanços e recuos civilizacionais, descobertas e inventos, encontros entre povos e culturas, mitos, crenças, afinal quase tudo, antes de passar para a língua e para a cultura dos povos, tem um nome e esse nome faz parte do léxico. O léxico é o repositório do saber linguístico e é ainda a janela através da qual um povo vê o mundo. Um saber partilhado que apenas existe na consciência dos falantes duma comunidade (Vilela, 1994: 6).

As expressões linguísticas manifestam figuras e isotopias locais, condensando aspectos semânticos vinculados às formas de vida. Elas emergem da experiência corpórea no mundo natural, da interação com o meio e com os outros sujeitos, bem como das configurações de sentido que estruturam a vida em sociedade.

A investigação de textos gerados por IA produzidos na interação entre pesquisadores e sistemas cujos *prompts* (comandos) são criados para fins da própria pesquisa é um problema metodológico que se apresenta hoje para a semiótica, e não será tratado neste trabalho. Compreender e explicar a significação, tomada como um processo,

foi a tarefa que a semiótica desenvolvida por Greimas e Courtés buscou enfrentar, propondo uma solução horizontal — o percurso gerativo de sentido — capaz de simular idealmente a passagem das formas esquematizadas ao uso. Sendo assim, o ponto de partida do semioticista sempre foi a manifestação textual. Com a popularização do modelo de IA generativa, tornou-se possível observar a geração de um ponto de vista inverso, pois é possível provocar a conversão instantânea de componentes do sistema dataficados em textos.

Apenas vale dizer que, dentro dos limites deste trabalho, optamos, para a realização dos testes, empregar comandos simples, de modo a evitar que variáveis externas interferissem nos resultados. As instruções foram formuladas na estrutura interrogativa “*O que significa...?*”, simulando a situação de um usuário interessado em compreender o significado de uma palavra. O objetivo dos testes, aqui tomados como uma amostra, é exemplificar a discussão acerca do *big data* e da sub-representação do Sul Global. Para tanto, selecionamos duas palavras de uso informal difundidas em regiões periféricas, no Centro-Oeste e no Nordeste do Brasil. As palavras escolhidas (“muage” e “peseiro”) nos permitem testar as IAs para além dos termos coloquiais e das gírias mais difundidas na Região Sudeste, maior centro econômico do Brasil.

Mesmo tendo consciência que essa pequena amostra aponta para a necessidade de ampliação futura da análise, pretendemos, aqui, traçar algumas observações acerca do comportamento da IA. Outro aspecto importante é que essas palavras, embora de uso corrente, não estão dicionarizadas. A escolha deliberada por elas tem caráter experimental, a fim de observar como a IA atua na produção do sentido ao sugerir definições para as palavras. Observe-se que as expressões linguísticas regionais do Centro-oeste do Brasil são, muitas vezes, desconhecidas ou estigmatizadas dentro do próprio país.

Na apresentação dos resultados são expostas as respostas obtidas no ChatGPT, e, em seguida, aquelas fornecidas pela IA Sabiá para cada um dos comandos. Os testes foram conduzidos em agosto de 2025, utilizando as versões ChatGPT-5 (OpenAI) e Sabiá 3.1 (Maritaca AI).

O primeiro comando refere-se à expressão linguística regional “muage”, com ocorrência mais frequente nas regiões Centro-Oeste e Nordeste do Brasil. Essa palavra designa uma brincadeira, uma maneira pejorativa de caracterizar o comportamento muito específico e exigente de alguém sobre algo ou também uma conversa desnecessária (“enrolação”).

**Tabela 1**

<p>Prompt: que significa muage? ChatGPT-5</p> <p>A palavra <b>muage</b> é pouco usada hoje, mas tem significado específico em textos históricos e jurídicos.</p> <p><b>Sentido principal</b> Vem do francês muage, derivado de muer (“mudar”, “transportar”). Designava um direito ou taxa cobrada pelo transporte de mercadorias em portos, rios ou canais. Era, portanto, uma espécie de pedágio ou imposto portuário, pago pelo carregamento ou descarregamento de navios.</p> <p><b>Exemplos históricos</b> Em documentos portugueses dos séculos XV ou XVI, “muagem/ muage” aparece ligado a alfândegas e feitorias, indicando o <b>valor que se devia pagar para transportar ou desembarcar mercadorias</b>. Expressões como “<b>para de muage</b>” significavam exatamente a <b>quantia destinada ao pagamento desse imposto</b>. Resumindo: <b>muage = taxa ou direito de transporte/ carga em portos ou rios, muito usada em português antigo e em documentos administrativos da época das navegações</b>. Quer que eu traga <b>exemplos de ocorrência da palavra em documentos portugueses medievais</b> (como forais ou cartas régias?)</p>
--

Resposta do GhatGPT- 5 < “O que é muage?”>

**Tabela 2**

<p>Prompt: que significa muage? Sabiá-3.1</p> <p>“Muage” é uma gíria ou expressão coloquial que pode ser encontrada em algumas regiões do Brasil, mas não é amplamente reconhecida ou utilizada. Pode ser que haja uma variação regional ou que a palavra seja um erro de digitação ou corruptela de outra expressão mais conhecida. Uma possibilidade é que “muage” seja uma variação de “muganga, que em alguns contextos pode significar uma brincadeira, careta ou trejeito. No entanto, essa interpretação pode não ser aplicável em todos os contextos e não é amplamente utilizada. Outra possibilidade é que “muage” seja uma forma de pronunciar ou escrever incorretamente a palavra “mauagem”, que não tem um significado específico em português, ou talvez uma confusão com a palavra “mugir”, que é o som emitido pelo gado bovino, mas essa conexão parece menos provável no uso coloquial. Dado que gírias e expressões coloquiais podem ser muito regionais e contextuais, e sem um contexto claro, é difícil determinar o significado exato de “muage”. Se você ouviu essa expressão em um contexto específico ou de uma determinada região, poderia fornecer mais detalhes para uma explicação mais precisa?</p>
---

Resposta do Sabiá-3.1 < “O que é muage?”>

Ao comparar as definições geradas pelos dois sistemas de IA, é possível observar alguns aspectos do funcionamento do algoritmo da IA em relação ao *dataset* de

treinamento. Na resposta chama a atenção o fato de o ChatGPT fazer uma afirmação categórica sobre o fato de que a palavra *muage* é pouco utilizada na atualidade. Tal afirmação contrasta com a realidade linguística brasileira, na qual a palavra circula entre milhões de falantes, sobretudo nas regiões Centro-Oeste e Nordeste. Essa discrepância entre a realidade atestada e a resposta ajuda a comprovar que os dados linguísticos dessas regiões não fazem parte do dataset do ChatGPT. Ao responder ao usuário por meio de uma afirmação acentua-se esse apagamento linguístico. O ChatGPT apresenta significados de caráter histórico, vinculados à etimologia francesa da palavra e à sua ocorrência em documentos oficiais produzidos em Portugal, veiculados entre os séculos XV e XVI, classificados como “sentido principal” da palavra. Essa própria classificação deixa margem para que se considere que a palavra apresente, quem sabe, outros sentidos não principais. Observa-se que não há qualquer referência ao Brasil ou às acepções correntes do termo em território nacional.

Para responder ao mesmo *prompt*, o chatbot Sabiá revela algum conhecimento sobre as gírias e expressões coloquiais do Brasil mencionando, também, a variação linguística regional. A definição não considera a existência de um sentido principal da palavra, nem produz afirmações categóricas, mas, pelo contrário, deixa margem para dúvidas por meio de expressões como “pode ser” ou “uma possibilidade”. A IA admite, inclusive, que “é difícil determinar o significado exato de “muage”. Além disso, faz menção, ao sentido de “brincadeira, careta ou trejeito”, correspondendo a uma das acepções efetivamente atestadas da palavra. No entanto, em determinadas regiões do Centro-Oeste, em especial no estado de Mato Grosso, a palavra é empregada também com os sentidos de “conversa fiada” e “frescura” — usos que não aparecem entre os resultados fornecidos.

Modalmente os dois sistemas de IA opõem-se pelo modo da certeza adotado pelo ChatGPT, contra o modo da dúvida em Sabiá. O GPT faz afirmações e usa exemplos para confirmá-los. Além disso, a resposta apresenta destaques em negrito, guiando a atenção do usuário e apresenta um resumo em que a palavra é seguida pelo sinal matemático de igual (=). Esse modo de construir a resposta reafirma valores de verdade e completude. Em Sabiá instaura-se o perfil modal da dúvida e a resposta ganha um caráter mais aberto e reflexivo. Os dois sistemas encerram a resposta propondo a continuidade da interação, mas enquanto o ChatGPT oferece mais exemplos que buscam confirmar a resposta dada, o Sabiá solicita mais informações que podem servir para reformular e aumentar a precisão da resposta.

Observemos a performance dos dois modelos em “peseiro”, segunda palavra eleita para este experimento. Peseiro pode ser usado para identificar uma situação difícil que o falante está passando.

**Tabela 3**

<p>Prompt: O que significa peseiro?</p> <p>A palavra “<b>peseiro</b>” pode ter alguns significados diferentes conforme o contexto:</p> <p><b>1.Sobrenome</b> Muito comum em Portugal, aparece como apelido de famílias. <b>Exemplo conhecido: José Peseiro, treinador de futebol português.</b></p> <p><b>2.Regionalismo/ Arcaísmo (Portugal)</b> Em alguns registros antigos da língua, “peseiro” tem relação com peso (de pesar) e podia significar <b>balança</b> ou o <b>lugar onde se pesa a mercadoria</b>. Também se usava como termo para quem trabalha com pesagem de grãos ou mercadorias.</p> <p><b>3. Topônimo</b> Peseiro é nome de alguns <b>lugares em Portugal</b>, geralmente pequenas aldeias ou localidades. <b>Resumindo:</b> hoje, o uso mais comum é como <b>sobrenome português</b>; mas historicamente, “peseiro” podia designar alguém ligado ao ato de <b>pesar mercadorias</b> ou mesmo o <b>local de pesagem</b>. Quer que eu busque exemplos em <b>textos antigos portugueses</b> onde aparece o termo nos sentidos de profissão/ lugar?</p>
---

Resposta ChatGPT-5: <O que significa peseiro?>

O mesmo teste conduzido com a IA Sabiá, forneceu-nos como resultado:

**Tabela 4**

<p>Prompt: O que significa peseiro?</p> <p>Sabiá-3.1</p> <p>“Peseiro” é um termo que pode ter diferentes significados dependendo do contexto. Aqui estão algumas possíveis interpretações:</p> <p><b>1.Na música:</b> Peseiro é um estilo musical popular em Cabo Verde, caracterizado por um ritmo lento e melancólico. É frequentemente associado a temas de amor e saudade.</p> <p><b>2.Gíria:</b> Em algumas regiões do Brasil “peseiro” pode ser usado como gíria para algo pesado ou difícil de lidar, embora essa não seja uma gíria amplamente conhecida ou utilizada.</p> <p><b>3.Regionalismo:</b> Em contextos regionais específicos, pode ter outros significados locais ou ser uma palavra usada em dialetos ou linguagens regionais, mas esses usos são menos comuns e conhecidos.</p> <p>Se você estiver se referindo a um contexto específico, como um termo regional ou dentro de um campo específico, por favor, forneça mais detalhes para que eu possa oferecer uma explicação mais precisa.</p>
--

Resposta de Sabiá-3.1: <O que significa peseiro?>

Em ambos os testes, os programas de IA admitem que a palavra pode ter diferentes significados e mencionam as diferenças contextuais. Contudo, o ChatGPT toma como contexto de uso da palavra apenas os usos em Portugal, sem fazer nenhuma menção ao Brasil ou a outros países de língua portuguesa. Já o Sabiá traz “peseiro” como estilo musical em Cabo Verde, país falante de língua portuguesa na África, talvez por supor que “peseiro” seja o mesmo que “piseiro”, uma variação do estilo musical “pisadinha” que surgiu na Bahia (região Nordeste do Brasil) e é dançada em Cabo Verde.

Os dois sistemas encerram a resposta de maneira bem semelhante ao do outro teste. O ChatGPT oferece mais exemplos que buscam confirmar a resposta dada, o Sabiá solicita mais informações para tornar a resposta mais precisa. Esse modo de encerrar a interação parece ser sistemático e deixa entrever que faz parte da programação do ChatGPT confirmar a informação dada, sem deixar espaço para dúvidas, enquanto Sabiá está programado para demonstrar a incompletude das respostas dadas.

Com base no conceito de práxis enunciativa e nos modos de existência Fontanille (2003); D’Armenio *et. al.*, 2024) argumentam que as IAs generativas convocam, no processo geracional do texto, formas virtualizadas e disponíveis em nosso patrimônio enciclopédico. Os comandos, então, fazem o processo de triagem dentro de certas zonas de bases de dados utilizados por um sistema de IA e realizam algumas dessas virtualidades.

Em IAs generativas como ChatGPT e Sabiá, o modelo cria novos textos pela associação entre palavras (prompts) e *embeddings*. O procedimento inicial consiste na tokenização, na divisão dos textos presentes no *dataset* em unidades mínimas, chamadas de tokens. Esses podem ser constituídos por uma palavra inteira, um segmento dela (ex. um sufixo), por uma pontuação, dentre outros. Em seguida, cada *token* é convertido em um vetor numérico, o *embedding*, que possibilita a categorização semântica: tokens semanticamente próximos tendem a se localizar em regiões contíguas no espaço vetorial, enquanto os mais distantes ocupam posições afastadas. A proximidade ou o distanciamento semântico são calculados pela recorrência estatística de aparições das formas e das possibilidades combinatórias dos tokens presentes nos dados de treinamento.

Quando digitamos um prompt, o modelo interpreta o *embedding* correspondente e, com base nos padrões previamente aprendidos, associa-o a prováveis significados, mobilizando isotopias, figuras e temas recorrentes. Trata-se, portanto, de uma operação que não parte da invenção *ex nihilo*, mas de um reaproveitamento e reorganização de virtualidades discursivas já inscritas no banco de dados.

Sendo assim, nos resultados apresentados nos testes, com comandos dirigidos às inteligências artificiais generativas ChatGPT e Sabiá, solicitando-lhes o significado de expressões linguísticas brasileiras, o texto-prompt assumiu a função de um ‘comando de pesquisa’ (Meyer, 2023), capaz de orientar o modelo por meio de procedimentos estatísticos e de regimes de valoração instaurados pelos algoritmos. Todavia, os resultados fornecidos por cada modelo encontram-se submetidos à coerção inerente da própria constituição de seus *datasets*: a maneira como cada um deles foi construído, bem como os critérios de valoração atribuídos aos conteúdos. A recorrência e a proximidade semântica dos dados condicionam diretamente a textualizações novas, tomadas como inéditas, produzidas pela IA. Tais textos geram significações e circulam socialmente fixando os sentidos inscritos neles.

As palavras aqui testadas não podem ser compreendidas como meras variantes lexicais; condensam sentidos que desvelam organizações de cenas práticas em que são produzidas, saberes e cosmologias que estruturam modos de existir coletivos, as *formas de vida*. Ao circular, integram-se a redes de práticas enunciativas e revelam a congruência de valores e de experiências partilhadas. Como pudemos comprovar nos exemplos, a ausência de determinadas palavras em modelos de Inteligência Artificial, a não consideração da variação linguística, a recorrência apenas à origem européia da palavra, como único sentido possível, evidencia no ChatGPT o apagamento das especificidades culturais de grupos minorizados e reforça a hegemonia histórica, linguística e semiótica das línguas e variedades do Norte Global, juntamente com suas práticas e *formas de vida*.

### **Considerações finais**

Este artigo teve como objetivo lançar luz sobre o modo como o sentido emerge em IAs generativas a partir de uma perspectiva do Sul Global. Para isso, acompanhamos o processo em sua integralidade — da dataficação inicial à geração do objeto semiótico produzido pela IA. Dessa maneira, procuramos conceber a dataficação como uma espécie de textualização, que permite, muitas vezes, apreender práticas, ou pelo menos alguns traços dessas, em um objeto semiótico manifestado em suporte mais estável e que, portanto, pode ser manipulável por dispositivos algorítmicos. Ao fazê-lo, reforçamos o caráter semiótico do *big data*, evidenciando que sua organização é permeada de axiologias que estarão presentes nas decisões tomadas por tais agentes maquínicos.

Dessa maneira, no processo de dataficação, os *datasets* deixam de ser apenas arquivos de dados: assumem a função de operadores que selecionam, reiteram e estabilizam isotopias presentes nos objetos semióticos que compõem essas grandes coleções. Tal operação, ao privilegiar conteúdos oriundos do Norte Global, instaura padrões hegemônicos que se apresentam como universais, apagando ou exotizando práticas, línguas e *formas de vida* do Sul.

Os testes aqui realizados com o ChatGPT e o Sabiá evidenciam, ainda que em escala reduzida, como a natureza dos dados de treinamento incide diretamente na significação. O ChatGPT, ancorado em valores da “globalização”, com treinamento de dados marcadamente em língua inglesa, tende a negligenciar usos brasileiros de determinadas palavras, priorizando leituras historicistas e lusitanas. Por sua vez, o Sabiá, alimentado por dados nacionais e em língua portuguesa, apresentou maior reconhecimento das isotopias locais, considerando a variação linguística. Do ponto de vista semiótico, isso confirma a hipótese central: os modelos não “inventam” ex nihilo, mas realizam virtualidades disponíveis em seus acervos (pela co-enunciação maquínica).

O experimento demonstrou como diferenças de constituição de *datasets* repercutem na geração de sentido. Todavia, o simples uso de *datasets* regionais, embora necessário, não é considerado suficiente para mitigar o problema, pois esses mesmos dados podem apresentar vieses preconceituosos em sua constituição. A construção de IAs mais éticas e culturalmente plurais, com vistas a diminuir as assimetrias estruturais entre Norte e Sul, passa também pela curadoria semiótica de dados - entendendo-os como objetos semióticos portadores de plano de expressão e de conteúdo. A adoção dessa prática tem o potencial de contribuir para melhoria das decisões tomadas na coleta, no balanceamento e no uso de tais dados, desde o processo de dataficação até a própria constituição das grandes coleções de dados.

### **Declaração de Disponibilidade de dados**

As autoras confirmam que os dados e materiais que apoiam as descobertas deste estudo estão disponíveis no artigo.

### **Declaração de Conflitos de Interesse**

As autoras declaram não ter conflitos de interesse.

### **Contribuição de Autoria CRediT**

Letícia Moraes: Conceptualization; Formal Analysis, Investigation, Methodology, Writing – Original Draft Preparation – Review & Editing  
Sílvia Maria de Sousa : Conceptualization; Formal Analysis, Investigation, Methodology, Writing – Original Draft Preparation – Review & Editing

## Referências

ABONIZIO, H.; ALMEIDA, T. S.; LAITZ, T.; MALAQUIAS JUNIOR, R.; KERCHÉ BONÁS, G.; NOGUEIRA, R.; PIRES, R. *Sabiá-3 Technical Report (arXiv:2410.12049)*. arXiv, 15 out. 2024. Disponível em: <https://arxiv.org/abs/2410.12049>. Acesso em: 28 out. 2025.

ALDAMA, J. A.; BERTRAND, D.; DARRAS, B.; DI SCIULLO, F. (Éds.). *Sémiotique impliquée: L'engagement du chercheur face aux sujets brûlants*. Paris: L'Harmattan, coll. « Sémioses », 2021.

AMOORE, L. *Cloud ethics: Algorithms and the attributes of ourselves and others*. Durham: Duke University Press, 2020.

BASSO FOSSALI, P. *Texte préparatoire au dossier. Actes Sémiotiques*, n. 115, 2012. Disponível em: <https://www.unilim.fr/actes-semiotiques/4928>. Acesso em: 28 out. 2025.

BOYD, D.; CRAWFORD, K. *Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information, Communication & Society*, v. 15, n. 5, p. 662–679, 2012. DOI: 10.1080/1369118X.2012.678878.

COECKELBERGH, M. *Ética na inteligência artificial*. Rio de Janeiro: Ubu Editora; Editora PUC-Rio, 2023.

D'ARMENIO, E.; DELIÈGE, A.; DONDERO, M. G. *Semiotics of machinic co-enunciation: About generative models (Midjourney and DALL·E)*. *Signata*, n. 15, 2024. DOI: 10.4000/127x4.

FONTANILLE, J. *Sémiotique du discours*. Limoges: PULIM, 2003.

FONTANILLE, J. *Pratiques sémiotiques*. Paris: Presses Universitaires de France, 2008.

FONTANILLE, J. *Formes de vie*. Liège: Presses Universitaires de Liège, 2015.

FONTANILLE, J. *Práticas e formas de vida: A semiótica de Greimas posta à prova pela antropologia contemporânea. Estudos Semióticos*, v. 13, n. 2, p. 66–76, 2017. DOI: 10.11606/issn.1980-4016.esse.2017.136307.

FONTANILLE, J. *Discursos, mídias, práticas e regimes de crença. Revista do GEL*, v. 16, n. 3, p. 246–261, 2019. DOI: 10.21165/gel.v16i3.2608.

FONTANILLE, J.; ZILBERBERG, C. *Tensão e significação*. São Paulo: Humanitas, 2001. (Obra original publicada em 1998).

GREIMAS, A. J. *Les formes de vie [Dossiê]. Recherches Sémiotiques / Semiotic Inquiry*, v. 13, n. 1–2, 1993.

GREIMAS, A. J.; COURTÉS, J. *Sémiotique: Dictionnaire raisonné de la théorie du langage*. Paris: Hachette, 1979.

HJELMSLEV, L. T. *Prolegomena to a theory of language*. Trad. F. J. Whitfield. Madison: University of Wisconsin Press, 2010. (Obra original publicada em 1975).

GHOSH, S.; KAPOOR, A.; RANE, A.; MAHENDIRAN, A.; THAKKAR, M. *Documenting patterns of exoticism of marginalized populations within text-to-image generators*. arXiv, 2025. Disponível em: <https://arxiv.org/abs/2508.02937>. Acesso em: 28 out. 2025.

HAGERTY, A.; RUBINOV, I. *Global AI ethics: A review of the social impacts and ethical implications of artificial intelligence*. arXiv, 2019. Disponível em: <https://arxiv.org/abs/1907.07892>. Acesso em: 28 out. 2025.

JAY AGRAWAL, J. G.; AVI, G. *Prediction Machines: The simple economics of Artificial Intelligence*. Cambridge, MA: Harvard Business Review Press, 2018.

KAPLUNOVICH, A. *Wealth of nations, wealth of data: How GDP shapes diverse large language models like ChatGPT: Interviewing assorted open source generative AI models*. In: *2023 IEEE International Conference on Big Data (BigData)*. IEEE, 2023. p. 4654–4663. DOI: 10.1109/BigData59044.2023.10386329.

MAYER-SCHÖNBERGER, V.; CUKIER, K. *Big data: A revolution that will transform how we live, work, and think*. Boston, MA: Houghton Mifflin Harcourt, 2013.

MARRONE, G. *L'invention du texte. Actes Sémiotiques*, n. 111, 2008. Disponível em: <https://www.unilim.fr/actes-semiotiques/1643>. Acesso em: 28 out. 2025.

MEJIAS, U. A.; COULDRY, N. *A colonização pelo dado: Do colonialismo histórico ao colonialismo de dados. Revista Sociedades Híbridas*, v. 4, n. 1, 2024. Disponível em: <https://revistas.hcte.ufrj.br/index.php/RevistaSH/article/view/446>. Acesso em: 28 out. 2025.

MEYER, R. *The new value of the archive: AI image generation and the visual economy of "style". IMAGE. Zeitschrift für interdisziplinäre Bildwissenschaft*, v. 19, n. 1, p. 100–111, 2023. DOI: 10.25969/mediarep/22314.

MORAES, L. *O que pode o(a) semioticista na era da inteligência artificial? Semiótica, big data e racismo algorítmico*. In: PORTELA, J. C.; SCHWARTZMANN, M. N.; CORRÊA, T. M.; SANTOS, F. C. R.; TASSINARI, C. A. (Orgs.). *Identidade, experiência e discurso: Semiótica e crítica da cultura*. Campinas, SP: Pontes Editores, 2024. v. 1, p. 139–168.

NOBLE, S. U. *Algorithms of oppression: How search engines reinforce racism*. New York: NYU Press, 2018.

PAVEAU, M.-A. *Análise do discurso digital: Dicionário das formas e das práticas*. Campinas: Pontes, 2021.

SEAVER, N. *What should an anthropology of algorithms do? Cultural Anthropology*, v. 33, n. 3, p. 375–385, 2018. DOI: 10.14506/ca33.3.04.

SCHWARTZMANN, M. N.; SILVA, L. H. O. da. *Romper, desviar, desafiar: Reflexões por uma semiótica implicada. Estudos Semióticos*, v. 18, n. 3, p. i–viii, 2022. DOI: 10.11606/issn.1980-4016.esse.2022.203773.

SHAHID, S.; ELSWAH, M.; VASHISTHA, A. *Think outside the data: Colonial biases and systemic issues in automated moderation pipelines for low-resource languages*. arXiv, 2025. Disponível em: <https://arxiv.org/abs/2501.13836>. Acesso em: 28 out. 2025.

SOUSA, S.M. *O discurso da inovação no ensino: uma análise semiótica*, *Soletras*, v.1, n. 47, p. 56-72, DOI. <https://doi.org/10.12957/soletras.2023.80345>. Disponível em:

<https://www.e-publicacoes.uerj.br/soletras/article/view/80345/48902>. Acesso em: 28 out. 2025.

TIFENTALE, A.; MANOVICH, L. *Selfiecity: Exploring photography and self-fashioning in social media*. In: BERRY, D.; DIETER, M. (Eds.). *Postdigital aesthetics: Art, computation and design*. London: Palgrave Macmillan, 2015. p. 109–122. Disponível em: <http://manovich.net/index.php/projects/selfiecity-exploring>. Acesso em: 28 out. 2025.

VAN DIJCK, J. *Datafication, dataism and dataveillance: Big data between scientific paradigm and ideology*. *Surveillance & Society*, v. 12, n. 2, p. 197–208, 2014.

VILARINO, R.; VICENTE, R. *An experiment on the mechanisms of racial bias in ML-based credit scoring in Brazil*. arXiv, 2020. Disponível em: <https://arxiv.org/abs/2011.09865>. Acesso em: 28 out. 2025.

VILELA, M. *Estudos de Lexicologia do Português*. Coimbra: Livraria Almedina, 1994.

## Este preprint foi submetido sob as seguintes condições:

- Os autores declaram que os necessários Termos de Consentimento Livre e Esclarecido de participantes ou pacientes na pesquisa foram obtidos e estão descritos no manuscrito, quando aplicável.
- Os autores declaram que a elaboração do manuscrito seguiu as normas éticas de comunicação científica.
- Os autores declaram que estão cientes que são os únicos responsáveis pelo conteúdo do preprint e que o depósito no SciELO Preprints não significa nenhum compromisso de parte do SciELO, exceto sua preservação e disseminação.
- Os autores declaram que os dados, aplicativos e outros conteúdos subjacentes ao manuscrito estão referenciados.
- O manuscrito depositado está no formato PDF.
- Os autores declaram que a pesquisa que deu origem ao manuscrito seguiu as boas práticas éticas e que as necessárias aprovações de comitês de ética de pesquisa, quando aplicável, estão descritas no manuscrito.
- Os autores declaram que uma vez que um manuscrito é postado no servidor SciELO Preprints, o mesmo só poderá ser retirado mediante pedido à Secretaria Editorial do SciELO Preprints, que afixará um aviso de retratação no seu lugar.
- Os autores concordam que o manuscrito aprovado será disponibilizado sob licença [Creative Commons CC-BY](#).
- O autor submissor declara que as contribuições de todos os autores e declaração de conflito de interesses estão incluídas de maneira explícita e em seções específicas do manuscrito.
- Os autores declaram que o manuscrito não foi depositado e/ou disponibilizado previamente em outro servidor de preprints ou publicado em um periódico.
- Caso o manuscrito esteja em processo de avaliação ou sendo preparado para publicação mas ainda não publicado por um periódico, os autores declaram que receberam autorização do periódico para realizar este depósito.
- O autor submissor declara que todos os autores do manuscrito concordam com a submissão ao SciELO Preprints.