Construction of meaning in generative AI: Discursive semiotic perspectives from the Global South

punctum.gr

BY: **Letícia Moraes** and **Silvia Sousa**

ABSTRACT

This paper examines the processes of meaning construction in generative AI models, with particular attention to how regional datasets shape the production of meanings related to the Global South. It draws on the semiotic framework developed by Greimas and Courtés, who approached meaning as a process and proposed a horizontal model – the generative trajectory of meaning (*parcours génératif*) – capable of ideally simulating the transition from abstract schematized forms to enunciative realization. With the widespread adoption of generative AI, it has become possible to observe the immediate conversion of system components, now stored not in social memory but in datasets composed of complex collections of text-objects. Grounded in Jacques Fontanille's theory of enunciative praxis, the study examines possible configurations of the discursive field in texts generated through interactions between human prompts and AI systems. It advances the hypothesis that the existential emergence of semiotic objects produced by such technologies no longer originates in pure virtuality (virtualization) – from which forms traditionally emerge – but rather stems from the continuous recombination of pre-formed, potentialized structures stored in datasets. This dynamic, in turn, helps explain the amplification of biases and stereotypes in AI outputs. To investigate this hypothesis and deepen the semiotic critique of AI from the perspective of the Global South, particularly in light of data underrepresentation, the study focuses on *Amazônia IA*, a generative model developed in Brazil. It compares this model with more globalized systems such as ChatGPT. Through this

ARTICLE INFO:
Volume: **12**
Issue: **02**
Winter **2025**
ISSN: **2459-2943**
DOI: **10.18680/hss.2025.0025**
Pages: **143-166**
Lic.: **CC BY-NC-ND 4.0**

KEYWORDS:
Generative trajectory of meaning
Enunciative praxis
Generative AI
Amazônia IA
Global South

comparison, the research seeks to reveal how meaning regimes operate when generative outputs are informed by regionally rooted datasets, potentially reflecting local cultural contexts. Ultimately, the paper aims to assess whether, and to what extent, such models challenge hegemonic frameworks of meaning construction and promote the development of more ethical, plural, and culturally situated discourses.

Introduction

Contemporary inquiries in the fields of social computing and digital humanities have increasingly approached Big Data as a complex repository of cultural artifacts produced, collected, and organized through everyday human practices. Such understanding, articulated by Microsoft Research scholars Danah Boyd and Kate Crawford (2012), shifts attention away from a merely technical or operational focus toward a perspective that foregrounds the sociocultural and political constitution of data. Within this framework, data are no longer conceived as raw or neutral entities; rather, they are foregrounded as the outcome of methodological choices and institutional decisions that reflect and reproduce cultural values, power structures, and ideologically determined contents.

Numerous studies analyze the pragmatic and discursive effects of the use of data and algorithms in contemporary culture. José van Dijck (2014) highlights the need for an inter- and transdisciplinary critique of Big Data systems that extends beyond engineering and economics. While Big Data and artificial intelligence are not identical phenomena, contemporary AI systems are structurally dependent on processes of datafication, large-scale data extraction, and algorithmic rationalities. In dialogue with these concerns, but from an ethico-political perspective, Dr. Louise Amoore (2020), a geographer at Durham University, introduces the term 'cloud ethics,' the title of her best-known book. For her:

> Cloud ethics is attuned to how algorithms come to configure the attributes of ourselves and of others, and to what it means to live in a world where such attributes are actively inferred and mobilized as the basis for action. (Amoore 2020: 4)

Inspired by the works of philosophers such as Michel Foucault and Judith Butler, Amoore reflects on how algorithms produce axiological configurations of identity – both of the self and of others – thereby participating in the constitution of subjectivity, alterity, and moral judgment in datafied societies. Amoore poses questions such as "Who is to be held accountable if an algorithm generated drone strike kills civilians or botches an intricate surgical operation which results in the loss of the patient's life?" (2020: 4) to underscore, in a compelling way, the urgency of interrogating the ethical principles at stake in a world increasingly mediated by artificial intelligence systems, algorithms, and data.

© 2025 Leticia Moraes and Silvia Sousa | Licenced under CC BY-NC-ND 4.0

In this initial gathering of interdisciplinary reflections, whose objective is to demonstrate how transformations across various social spheres are reflected in multiple fields of knowledge, let us revisit the thoughts of the anthropologist Nick Seaver (2018). He suggests that anthropologists approach data as ethnographic objects, similar to any complex cultural artifact, and that this requires attention to the multiplicity of voices, discursive disputes, and practices involved in their constitution. This perspective requires attention to the multiplicity of voices, the discursive disputes, and the practices implicated in their constitution. Considering the contributions of Boyd and Crawford (2012), van Dijck (2014), Amoore (2020), and Seaver himself (2018), it can be argued that data not only reflect cultural reality but also constitute an integral part of the semiosphere. Far from proposing strictly ontological discussions of the nature of data, such reflections illuminate how meaning is produced and contested for new spaces of circulation – concrete, virtual, and symbolic – within datafied societies. From a semiotic perspective, it becomes evident that, through digital technologies – and the objects they produce – meaning emerges from individual actions, ethical conduct, collective organization, and even the creation of axiological configurations of aesthetic taste.

This initial discussion prompts us to examine the epistemological nature of data from a semiotic perspective. For us semioticians, what are data, after all? How can they be apprehended within a conceptual framework specific to our discipline? At what level of pertinence should they be situated? In what ways do the logic of data production, storage, organization, and circulation influence the construction of meaning?

To address these questions, the article is divided into three sections. The first section offers a semiotic discussion of datafication, linking it to the concepts of *practices* and *forms of life* (Fontanille 2008, 2015), followed by an examination of data bias in the Global South. The goal is to provide a solid semiotic framework for these two aspects. Finally, the third section provides a brief illustrative analysis of responses generated by a regional AI model developed in Brazil, compared with those from ChatGPT.

1. Big Data and datafication

Understood as extensive and complex collections of semiotic objects manifested on the level of pertinence of expression through multiple languages and codes – verbal, verbal-visual, visual, mathematical, syncretic, among others – data do not constitute tangible realities of the natural world[1] (Moraes 2021). Unlike stones, they are not raw elements: these objects are always ‘something’ that has been modeled. By qualifying them as semiotic objects, we draw on Marrone (2008 n.p.), for whom a semiotic object is “a simulacrum, anything that must be prepared.”

---

[1] Although data rely on digital and physical infrastructures that raise ethical and ecological issues, these material dimensions are analytically bracketed here. The analysis focuses on data as semiotic objects.

Therefore, for an artifact to be recognized as data, the establishment of 'analysis' is indispensable, as Hjelmslev (2010 [1975]) pointed out. It was also through the Hjelmslevian thought that semiotics highlighted the need to delineate and inscribe the semiotic-object within a system of signification, as only in this way does it become possible to formally describe the internal relations established by it. To apprehend data semiotically, as well as to delimit how their production and circulation produce meaning, we will address the issue of datafication, drawing primarily on the studies of Jacques Fontanille (2008, 2015) regarding *practices* and *forms of life*.

1.1. Datafication: Big Data and semiotic practices

Data modeling is part of a broader process known as datafication. This term, coined in 2013 by Viktor Mayer-Schönberger of the University of Oxford and journalist Kenneth Cukier – both active in the field of internet governance and regulation – refers to a finite sequence of executable actions – algorithms – that evaluate the performance of previously collected data and make decisions based on established rules (Mayer-Schönberger and Cukier 2013: 28). Such a technical definition, emerging from the field of computer science, tends to obscure aspects that are fundamental to the human sciences. By figurativizing a supposed technological neutrality, these technical processes mask the biases embedded in them, especially when examined from the epistemological perspective of the Global South.[2] One can interrogate the technical character of these processes by asking how decisions are made and how rules are established, and which axiological values guide and sustain them. These reflections are essential for understanding the generation and circulation of data.

Datafication cannot be reduced to a mere technical process of encoding reality; instead, it functions as a semiotic device that converts *social practices* into quantifiable data to feed algorithmic ecosystems whose logic and infrastructure remain largely under the control of actors concentrated in the so-called Global North (Mejias and Couldry 2024). It is important to emphasize that not only the processing rules, but also the discursive values concretized in the content of the data – comprising diverse textual genres, images, audio, and codes – shape the formation of *forms of life*. These *forms of life*, which ideally should be locally grounded, are instead instantiated and traversed through a modeling process that flattens culturally specific configurations of meaning. This contributes to the normalization of colonial epistemologies and intensifies the marginalization of non-hegemonic knowledge and experiences.

---

[2] The division between the Global South and Global North has its roots in geopolitics and economics, historically framed by expressions such as the 'Third World' (Alfred Sauvy 1952) and 'developing countries.' Between the 1980s and 1990s, the term "Global South" began to appear in United Nations debates, alongside the conceptual framing of a North–South divide, as exemplified in the Brandt Report (North–South: A Programme for Survival 1980), which emphasized the structural inequalities between these two axes.

© 2025 Leticia Moraes and Silvia Sousa | Licenced under CC BY-NC-ND 4.0

Within the framework of an 'implicated' semiotics (Aldama et al. 2021) – that is, an approach that acknowledges the analyst's ethical, political, and epistemological involvement in the processes of meaning-making under examination, and thus rejects the assumption of a neutral or external position with respect to its objects – we define datafication as a syntagmatic arrangement composed of distinct practical scenes that, at first, convert certain human actions and gestures into quantifiable data, which are then processed by artificial intelligence algorithms and ultimately employed to make sense of practices and forms of life (Fontanille 2008, 2015) of subjects embedded in the very same datafied society from which the original data were extracted. The circular nature of datafication – rigorous in its own operation – becomes even more problematic when it exposes that the original data pertain to only a small fraction of the world, and still, through datafication, attain a global scale.

In his *Pratiques sémiotiques* (2008: 4), Fontanille proposes a hierarchy of levels of pertinence for semiotic analysis and describes *practical scenes* as syntagmatic arrangements of human behaviors and actions, endowed with both a plane of expression and a plane of content. The author observes that practices are 'courses of action' (2019: 251); they are open-ended, with meaning emerging through the syntagmatic unfolding between a beginning and an end. Unlike semiotic objects inscribed on stable *supports*, such as a photograph, *semiotic practices* produce *live* meaning – that is, while they occur – allowing the praxic movement itself to act as a meaning-generating element; signification is, therefore, inherently dynamic.

Big Data thus constitutes a development, let us say, a more stabilized outcome, of the intangible and dynamic process of datafication (Moraes 2025), insofar as it provides the raw material for artificial intelligence systems to make predictions based on mathematical and statistical calculations (Jay, Agrawal and Goldfarb 2018). It is therefore not difficult to observe that Big Data – although characterized by a codified materiality and a non-transparent enunciative process – nonetheless concretizes isotopic lines anchored in colonialist ideological biases and in the worldviews of the Global North.

Datafication reconfigures the discursive field by converting practical scenes – various courses of action such as purchasing, gesturing, circulating, accessing, and searching – into manipulable semiotic objects. From Fontanille's (2008) perspective, data may thus be understood as syncopal condensations of practices and strategies at the level of the text, in which complex chains of action are selectively compressed and formalized. This enables not only the prediction of actions and behaviors but also the reorganization of meaning configurations that emerge from human interactions. Within this dynamic, data lose their status as archived traces, that is, as testimonies of past *practices*, and instead assume a programming function, serving as material *supports* for the probabilistic modeling of possible futures within the interpretive frameworks operated by AI systems. Such a process of passing from lived acts to data alters the dynamics established among modes of existence, since, in the shift from virtuality to actualization, a kind of algorithmic enunciation emerges, co-enunciated by both human and non-human actants.[3] As this process is circular, rapid, continuous, non-transparent, and far-reaching, it becomes increasingly difficult to distinguish between the characteristics of input data (raw data – lived scenes) and output data (datafied forms). Moreover, it becomes progressively harder to delimit the roles and positions of social subjects within this process. This nebulous condition, which blurs the operations of datafication, lies at the core of current issues concerning transparency, accountability, and ethics in AI. Global initiatives are working to formulate policies that promote 'explainable AI' and safeguard human dignity amid AI's growing use (Coeckelbergh 2023: 143), efforts to which semiotics must actively contribute.

Understanding the process of datafication allows us to define Big Data with greater precision as a vast assembly of collections of semiotic objects generated through the unfolding of human practices, whose circulation and appropriation are governed by structural logics of power, as we will demonstrate in the second section of this article. It is important to recognize that much of the semiotic material generated by datafication begins the process of meaning in actu, within the unfolding of a *practical scene*. This is the case, for example, of a monetary transaction conducted with a credit card when purchasing a product in a supermarket. Although such a *practical scene* is not fully stabilized in material support, it generates semiotic objects that can subsequently be stabilized (such as an electronic receipt, a list of purchased items, or an individual's tax identification number). These then become part of large collections of semiotic objects endowed with a new plane of expression and of content: Big Data. These collections are characterized by incommensurability, heterogeneity, and the overlapping of languages (verbal, audio, visual), genres, textual types, subjects, and fields of knowledge, as well as by access restrictions and confidentiality imposed by database-holding corporations (Moraes 2025). Such conditions contribute to the erosion of authorial, fiscal, temporal, and organizational origins, as well as to the difficulty of locating enunciative assumptions – ultimately complicating the identification of who speaks and under what conditions.

We emphasize the close articulation between data generation – initially extracted from human actions and currently already produced automatically by AI – and the process of datafication because, from a semiotic perspective, the latter operates as a form of textualization (Greimas and Courtés 2008 [1979]: 504) of human actions and behaviors inscribed in practical scenes and, increasingly, emulated by AI systems. Prior to the consolidation of artificial intelligence technologies associated with Big Data, many gestures

---

[3] D'Armenio et al. (2024) employ the expression co-enunciation to designate the mode of operation of generative AI models, underscoring the joint participation of human and machinic actants in the production of enunciation.

and traces of everyday experience were not systematically recorded on stable *supports*, which limited the possibility of capturing and analyzing them in real time. This, in turn, constrained the extraction of recurrent patterns essential to predictive models that anticipate possible behaviors within collective modes of social organization.

The passage from a non- (or less) stable *support* to a more stabilized one occurs through the process of datafication. The stabilized *support*, in this case, plays a decisive mediating role: it allows us to understand datafication as a form of textualization, that is, "a set of procedures (to be organized in a textual syntax) which aim at constituting a discursive continuum, before the manifestation of discourse in one semiotics or another" (Greimas and Courtés 1982 [1979]: 316). In the case of datafication, this continuum unfolds as actions are modulated into quantifiable data, so that human gestures, once ephemeral and inscribed in a *practical scene*, come to exist as manipulable semiotic objects.

In the last decade, the expansion of Big Data has not only intensified its use for predicting behaviors and mapping *forms of life*, but also profoundly reconfigured its management and storage. Data that were previously under the majority control of state entities – national or regional – have progressively come to be managed by transnational private corporations. At the same time, there has been a substantial expansion in the scope of stored content: beyond administrative and demographic records, these databases now include information capable of revealing personal consumption preferences, daily trajectories, ideological and political positions, and social networks, among other data.

This shift becomes even more problematic when analyzed from a critical perspective. Local practices are textualized into data that feed models constructed under the aegis of hegemonic Northern axiologies. In this context, the ordinary user – regardless of geographic location – is equally affected by these hegemonic values, insofar as they do not effectively participate in decision-making processes. However, from the standpoint of the Global South, which remains underrepresented, this position may also serve as a critical locus for examining and interrogating stereotypes and hegemonic structures.

Embedded within the broader process of datafication, these models operate as mechanisms of standardization and normalization of *practices* and *forms of life*, imposing pertinence cuts that frequently disregard – or even erase – the cultural, historical, and political specificities of the populations from which the data were extracted, thus confirming the process of loss of origins to which we referred. Therefore, rather than acting solely on isolated individual behaviors, it is crucial to understand how this process intervenes in the configuration of *forms of life*, reorienting *practices*, producing meanings, and instituting new regimes of collective organization in datafied societies.

1.2. Datafication: From Big Data to Forms of Life

A *form of life* designates the most encompassing form of semiosis ever identified, defined as collective modes of social and cultural organization that both carry and produce meaning. They are syntagmatic *forms of life* (Fontanille 2015: 43) and enable the apprehension of an elemental, generic, and collective experience: life itself or the course it takes (Fontanille 2017: 73). At the highest level of semiotic analysis, *forms of life* allow for the understanding of the experience of *practices* (Fontanille 2008), which are structured by an ethos and by social norms, and in which multiple semiotic objects converge in an orchestrated manner to generate the meaning of life. This, therefore, constitutes a semiotic articulation among ethos, norms, and *practices* that structures a semiosphere.

The existence of social subjects is only possible within *forms of life*. By incorporating an axiological framework, the subject recognizes itself in the values that structure the semiosphere, thereby constituting itself as such. It is within this collective space that "subjects and social groups are the protagonists of interactions and sociocultural exchanges"[4] (Fontanille 2015: 247). Far from being reducible to mere behavioral patterns, *forms of life* configure ways of seeing, reading, organizing, and inhabiting the world. According to Fontanille and Zilberberg, in *Tension and Signification* (Tensão e Significação), in the chapter devoted to *forms of life*, "every manifestation of a form of life is then considered as a condensate of the entire *form of life*"[5] (2001 [1998]: 204). Its apprehension, however, only occurs through relation, for it is in confrontation and coexistence with other *forms of life* that they become comprehensible (Fontanille 2015: 59).

Although they sustain identity narratives, *forms of life* cannot be analyzed solely at the individual level, as they are essentially collective. They delineate and sustain an existential trajectory for the subject within the social body, conferring congruence for syntagmatic existence while simultaneously articulating a system of values inherent to the meaning of the subject's life. When we relate this concept to the discussion of datafication, we can understand how the analysis of data – carried out by artificial intelligence models, algorithms, and statistics – enables data-holding companies to identify the meanings inscribed and their salience within collective organizations on the one hand, and, on the other, to act by influencing the schemes of selection of dominant categorical content, the discursivization (figures and themes), and the valorization that structure specific *forms of life*.

---

[4] Original in Portuguese: *"Os sujeitos e os grupos sociais são os protagonistas das interações e das trocas socioculturais."* Translator's note.

[5] Original in Portuguese: *"toda manifestação de uma forma de vida é então considerada como um condensado da forma de vida inteira."* Translator's note.

Regarding analysis, different collective economic actants mobilize Big Data to guide decisions: financial institutions estimate the probability of default for individuals or groups; social networks deliver targeted advertisements based on behavioral profiles; health insurance providers adjust premiums according to medical consumption histories; and streaming platforms such as Spotify, Netflix, or Amazon Prime Video filter and recommend personalized content based on each user's activity. More than mere schematizations of behavior, these analyses allow the alteration of schematizations of meaning that confer congruence and coherence on given *forms of life*. It is noteworthy, however, that the more controlled *forms of life* are constructed – generated in and by Big Data – the more meanings of life are projected as formless and uncontrollable, since it is often impossible to determine who, where, or when is accountable in cases of disinformation dissemination, cybercrime, fraud, or workplace abuses, for example.

As we have noted, when coupled with artificial intelligence technologies, Big Data actively modulates social *practices* and *forms of life*. Global streaming platforms, such as Netflix, recommend content to individual users or groups of users. These subjects establish an interaction, in enunciative praxis, with that semiotic object, which, without such direct intervention, would possibly not occur. Here, we draw attention to the fact that the repetition of this dynamic not only reinforces certain consumption habits – beyond the shaping of individual tastes and styles – but also acts to model entrenched collective organizations, altering the regimes of meaning selected at the poles of a category and assigning them values. This seemingly innocuous process illustrates the silent and inchoative operation of datafication in the construction of meaning within collective organizations.

This phenomenon is illustrated by the results of Lev Manovich's *Selfiecity* project. The *MediaLab* team analyzed a corpus of over three thousand selfies published on Instagram from five cities – São Paulo, Bangkok, Moscow, New York, and Berlin – with the aim of identifying patterns and variations in the language of digital photography, considering aspects such as color schemes, use of filters, incidence, and distribution of light and shadow, gesturality, and affective expressiveness. In the statistical analysis of the categories 'head tilt' and 'pouting lips,' correlated with the female/male gender marker, the results (cf. Titenfale and Manovich 2015) indicated the recurrence and valorization of these gestures as transnational aesthetic patterns, with particular prevalence among women.

The massive recurrence of this visual pattern, reinforced by platforms whose algorithms favor its visibility, contributes to its naturalization. Gestures that might initially be perceived as eccentric or as expressions of individual singularity become so repeatedly enacted that they lose their semantic value as markers of an individual style. This standardization of gesture not only differentiates the practice of selfies according to a binary gender logic on social networks but also intervenes in the shaping of *forms of life*: tilting the head, smiling to show the upper teeth, or pouting become recurrent behaviors in women's everyday practice of self-publication on social media, gradually assimilated as expected – and socially legitimized – traits of the *form of life* associated with the category of the feminine.

As Fontanille and Zilberberg (1998) point out, *forms of life* are not immutable, as they respond to enunciative praxis, that is, to use. Accordingly, it is coherence and congruence that ensure the stabilization of a *form of life* and its social recognition. In light of this, it becomes necessary to investigate which *forms of life* are manifested by the objects generated by artificial intelligence. What categorical content and discursive elements are highlighted in them? And through these discursivization, what is overlooked?

The recurrence and consequent naturalization of aesthetic and behavioral patterns, as highlighted in the research of Tifentale and Manovich (2015), reveal a deeper functioning of the algorithmic logic: that of establishing and propagating hegemonic models of representation which, by assuming themselves as dominant isotopies, diverge from the configurations of meaning produced locally. Following Paveau's reasoning, algorithms are "sequences of instructions that allow the solution of problems,"[6] they sort information, creating classifications and hierarchies that render AIs far more than technological tools, since "certain information will appear more frequently, or in a better place than others, or will be more widely disseminated than others, or, on the contrary, will be prevented from appearing"[7] (Paveau 2021: 39). From a semiotic perspective, this constitutes a process of stabilization of forms of expression and content which, when reiterated and valorized by the technical apparatus (artificial intelligence models, algorithms, datasets, etc.), become part of the normative and collective repertoire at the most encompassing level of immanence, that is, of *forms of life*. This discussion underscores the need to examine the use of data and its biases as a discursive phenomenon through the lens of discursive semiotics, embracing its vocation for an 'implicated' semiotics (Aldama et al. 2021).

---

[6] Original in Portuguese: *"sequências de instruções que permitem a solução de problemas"* Transator's note.

[7] Original in Portuguese: *"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"* Translator's note.

© 2025 Leticia Moraes and Silvia Sousa | Licenced under CC BY-NC-ND 4.0

2. Data Bias in AIs from a Global South perspective

As we demonstrated, the use of predictive algorithms, aimed at forecasting future behaviors, and generative algorithms, designed to create novel semiotic objects, statistically affects the configuration of Big Data. This process performs a process of hierarchization in the distribution of content, privileging certain patterns so that whatever appears more frequently tends to be reiterated and consolidated into a 'norm.' This automatism of distribution and combination favors the characteristics of hegemonic cultures in the construction of meaning, thereby increasing the productivity of specific languages, modes of expression, topics, and images. It is sufficient to consider the existing discrepancy between geopolitically dominant countries and the rest of the world regarding the development of such technologies, data extraction, and the languages used (Hagerty and Rubinov 2019). In this scenario, English-speaking countries – especially the United States – occupy the top of the chain; as a result, most semiotic objects comprising Big Data and processed by AI originate from these communities.

On the other hand, non-hegemonic cultural practices – such as regional linguistic expressions, Afro-diasporic modes of sociability, and Indigenous epistemologies and worldviews – are marginalized, since they are read by the machine as singularities that escape the statistical regularities internalized by the models. In other words, these practices and configurations of meaning are understood as deviations from the norm, a norm consolidated by the frequency and repetition of forms present in the semiotic objects that constitute the Big Data. This situation places us before two central issues: (i) the exclusionary character of the very operational model of these technologies, and (ii) the naturalization of cultural hierarchies in the production of meaning mediated by AI.

As noted earlier, in datafied societies, where machinic and human agents jointly participate in the enunciative process (D'Armenio et al. 2024), the construction of meaning is mediated by statistical operations that shape the axiologization of isotopies. In discursivization, the machinic gesture, by reiterating certain figurative and thematic arrangements, tends to intensify their status as 'universals,' establishing them as dominant values within the discursive field. By contrast, non-hegemonic isotopies, insofar as they do not correspond to the regularities reiterated in the objects used for training, are construed either as deviations or as exotic occurrences.

The results can be observed in different practices mediated by artificial intelligence systems. By way of example, the following stand out: the stereotypical and homogenizing figurativization of subjects from the Global South in images generated by generative AIs (e.g., DALL·E and MidJourney), as if they all shared identical clothing, habits, and cultural traits (Ghosh et al. 2025); the systematic exclusion of languages from marginalized countries in automated moderation processes (Sahid et al. 2025); the recurring association of Black women with sexualization in search engine results (Noble 2018); and the linking of socially valued professions, such as that of physicians, to the visual figurativization of white people devoid of Brazilian identity markers in image banks widely used in the country (Moraes 2025).

The examples above operate primarily at the level of discursivization, reiterating figurative and thematic patterns from the Global North in a way that confers euphoric valorization upon them, while dysphoria is associated with the semantic content of the Global South.[8] These shifts in categorizations and regimes of meaning can also be observed in the construction of meaning within both emerging and consolidated practices, as well as in *forms of life* that are either on the rise or on the verge of disappearance.

In this scenario, the research of Brazilian scholars Vilarino and Vicente (2020) sheds light on how the use of predictive AI by financial agencies can restrict the access to credit and the purchasing power of Black individuals, since credit scoring is based on the calculations performed by the model and reproduces racial bias, even when such data does not explicitly appear as a parameter. This occurs because other datafied information, such as geographic location and socioeconomic characteristics, is closely related to race in the cultural context of Brazil – a country marked by deep social and racial inequalities and home to the second largest Black population in the world. By privileging dominant cultural repertoires, AI not only renders local identities and experiences invisible but also creates stereotypical representations that function as stabilized simulacra of certain *forms of life*.

Given such a scenario and focusing our discussion more specifically on the domain of generative AIs, we argue that the constitution of datasets is a crucial point in the problem of data bias and the erasure or exoticization of the cultures and languages of minoritized groups. Broadly speaking, datasets are subsets of Big Data and consist of collections of textual, visual, sonic, or audiovisual objects used to train specific models. Since these collections, for the most part, consist of semiotic objects in English or are heavily marked by the culture of the Global North (Moraes 2024), they perpetuate a new modality of colonization through data (Mejias and Couldry 2024). Inequality is manifested both in the quantity and quality of data available between the North and the South, as well as in the structural asymmetry between technological production, concentrated in central countries, and the mere importation of these technologies by peripheral countries (Hagerty and Rubinov 2019).

---

[8] Global South does not exactly designate a geographical location, but rather an anti-hegemonic epistemological positioning, proposing a radical critique of the predominant hegemony of certain cultures in science, culture, technology, etc.

Thus, by inscribing themselves within the logic of datafication, datasets reveal themselves not only as technical instruments but as semiotic operators that select, reiterate, and legitimize particular schemas of meaning. By privileging content originating from the Global North, these collections of semiotic objects consolidate hegemonic isotopies that propagate as universal patterns of representation. This process, sustained by the congruence between selection, combination, and productivity, stabilizes *forms of life* marked by colonial values.

Consequently, the datasets used in the most widely deployed generative models – such as ChatGPT, DALL·E, Sora, and MidJourney – contribute to the underrepresentation of figures, themes, *practices*, and *forms of life* that reflect the specificities and cultural diversity of Global South populations, thereby reinforcing the reproduction of stereotypes and the semiotic erasure of these non-hegemonic communities. As a potential mitigation strategy, various specialists emphasize the need to develop and use regional datasets for training AI systems.

The following section examines whether AIs supposedly trained on a less universalist, more regionally oriented dataset influence the construction of meaning in semiotic objects produced by generative models. We will conduct a comparative analysis of results generated by two artificial intelligences: ChatGPT and *Sabiá*.

3. The construction of meaning in Big Data (and datasets): from the universal to the regional

In this section, we present *Sabiá* as a regional generative AI model, which, according to platform data, was developed in Brazil and primarily trained on Brazilian data. In the analysis, the responses produced by this model will be contrasted with those from OpenAI's globally deployed model, ChatGPT. The aim is to identify whether, and to what extent, the regional AI model disrupts – or at least represents a shift in relation to – hegemonic logics of meaning production, and whether it fosters the construction of more ethical and culturally plural discourses.

The *Sabiá* chatbot was launched in October 2022 by the company *Maritaca AI.* According to its developers, the model was developed in partnership with Google and is primarily trained on a Portuguese-language dataset focused on Brazil. This regional emphasis, in principle, enables the model to handle data more closely aligned with the country’s culture, history, and sociocultural context, potentially enhancing its performance (Abonizio et al. 2024). The company name, *Maritaca AI* (also known as MariTalk), is a portmanteau of the Portuguese word *maritaca* and the English verb *talk*. *Maritaca*, a term of Tupi indigenous origin (*tupi mba’é taka*), designates a bird species common in Brazil. Unlike parrots, *maritacas* do not imitate human speech but instead produce a distinctive, loud, and strident song.

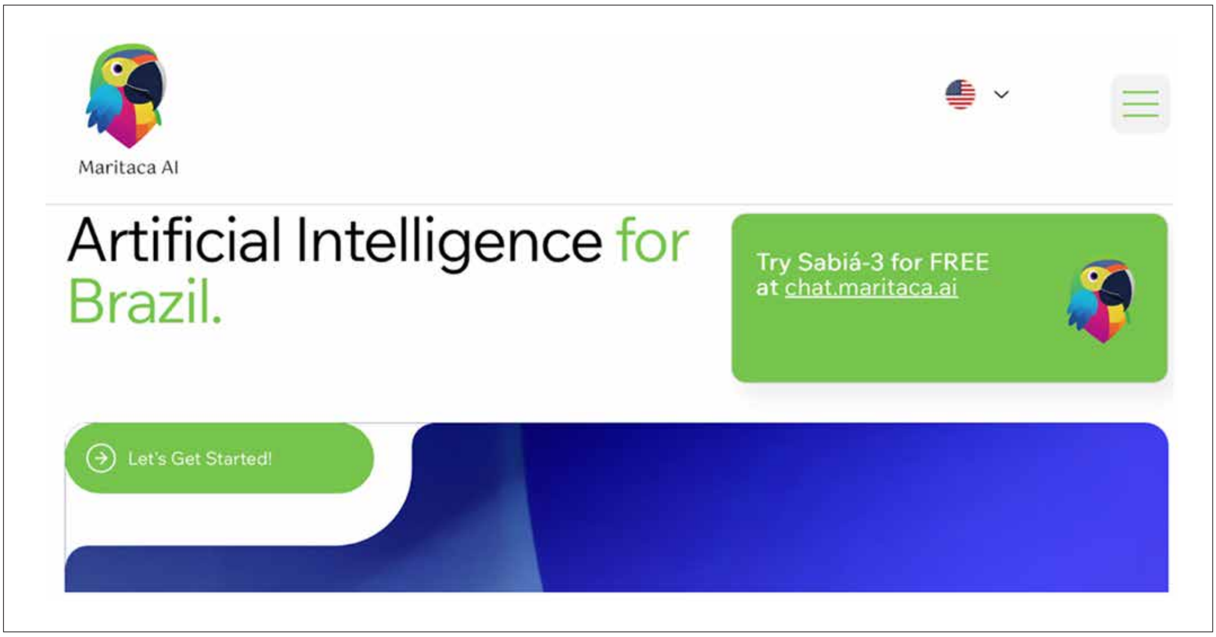
Maritaca AI
Artificial Intelligence for Brazil.
Try Sabiá-3 for FREE at chat.maritaca.ai
Let's Get Started!

Figure 1. First scroll of the *Maritaca AI* webpage [9]

Sabiá, a Tupi term (s-apia), denotes a bird species found in South America. A brief analysis of the model reveals that its chosen name and visual presentation – which features the image of a maritaca – foreground the theme of Brazilian identity, thereby distinguishing it from globally disseminated AI models. The selection of the name, the color palette in shades of green and blue, and the depiction of the maritaca figurativize the theme 'Brazilianness,' invoking aspects widely considered emblematic of the country, such as its fauna and flora. Moreover, Maritaca AI features the slogan "Artificial Intelligence for Brazil," positioning the AI model within a specific geographic and cultural context. There is no doubt that it constitutes a regional AI model.

ChatGPT, in turn, is a chatbot developed by OpenAI, launched in November 2022, designed as a globally oriented tool trained on extensive datasets drawn from worldwide Big Data. Its neutral visual design – with a white background and a symbol of curved, abstract black-and-white shapes – evokes the theme of technology itself and the flow of information within AI. The ChatGPT interface is not tied to any particular region, nor does it display any markers of local specificity. This purported 'universality,' however, conceals asymmetries: the distribution and density of data vary significantly across regions, with a predominance of content originating from countries with higher Gross Domestic Product (GDP) and datasets constructed primarily in English (Kaplunovich 2023).

---

[9] Available at: https://www.maritaca.ai. Accessed August 8, 2025.

For the analysis, we selected Brazilian linguistic expressions, given that the lexicon is one of the subsystems of language that most directly reflects the interplay between language and culture, since words are tasked with designating referents within the extralinguistic system. As Vilela (1994:6) notes:

> The lexicon is the part of a language that primarily shapes extralinguistic activity and archives the linguistic knowledge of a community. Civilizational advances and setbacks, discoveries and inventions, encounters between peoples and cultures, myths, beliefs, almost everything, before entering the language and culture of a people, there is a name, and that name belongs to the lexicon. The lexicon is the repository of linguistic knowledge and is also the window through which one people perceive the world. A shared knowledge that exists only in the consciousness of the speakers of a community.[10] (Vilela, 1994: 6)

Linguistic expressions manifest local figures and isotopies, condensing semantic aspects that are tied to *forms of life*. They emerge from bodily experience in the natural world, from interaction with the environment and with other subjects, as well as from the configurations of meaning that structure life in society.

The investigation of AI-generated texts produced through the interaction between researchers and systems – where prompts are specifically formulated for research purposes – constitutes a methodological problem currently faced by semiotics, and it will not be addressed in this study. The task of understanding and explaining signification, conceived as a process, was taken up by the semiotics developed by Greimas and Courtés, who proposed a vertical solution – the generative trajectory of meaning – capable of ideally simulating the passage from schematized forms to usage. Thus, the semiotician's point of departure has always been textual manifestation. With the popularization of generative AI models, however, it has become possible to observe this process from an inverted perspective, since one can now provoke the instantaneous conversion of datafied system components into texts. That said, within the scope of this study, we opted to employ simple prompts to avoid external variables interfering with the results. The instructions were formulated in the interrogative form "What does … mean?", simulating a user seeking to understand the meaning of a word. The purpose of these tests, here understood as a sample, is to exemplify the discussion concerning Big Data and the underrepresentation of the Global South. To this end,

---

[10] Original in Portuguese: "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." Translator's note.

we selected two words in informal usage, common in peripheral regions of Brazil's Midwest and Northeast. The chosen words ('muage' and 'peseiro') allow us to test the AIs beyond the colloquial terms and slang more widely disseminated in the Southeast Region, Brazil's largest economic center.

While acknowledging that this small sample points to the need for further analysis, our aim here is to outline observations regarding AI behavior. Another important aspect is that these words, although in common use, are not attested in dictionaries; rather, they circulate on social media, blogs, YouTube, and the comment sections of journalistic websites, as well as in other less formal contexts. The deliberate choice of such terms has an experimental character, insofar as it allows us to observe how AI operates in the production of meaning when suggesting definitions for them. It should be noted that regional linguistic expressions from the Midwest region of Brazil are often either unknown or stigmatized within Brazil. In the presentation of results, we first report the responses obtained from ChatGPT, followed by those generated by the *Sabiá* model for each prompt. The tests were conducted in August 2025, using ChatGPT-5 (OpenAI) and *Sabiá* 3.1 (*Maritaca AI*).

The first prompt concerns the regional linguistic expression 'muage,' which is most frequently used in Brazil's Midwest and Northeast. This word designates a kind of playful activity, but it can also function as a pejorative way of characterizing someone's overly specific or demanding behavior regarding something, or, alternatively, as referring to idle talk ('enrolação').

Table 1. ChatGPT–5 Response – <"What does *muage* mean?">

Prompt: What does *muage* mean?
ChatGPT-5:

The word ***muage*** is rarely used today, but it has a specific meaning in historical and legal texts.

Primary Meaning:
It comes from the French *muage*, derived from *muer* ("to change," "to transport").
It designated a fee or duty charged for the transportation of goods in ports, rivers, or canals.
Thus, it was a kind of toll or port tax, paid for loading or unloading ships.

Historical Examples:
In Portuguese documents from the 15th or 16th centuries, *muagem/muage* appears linked to customs houses and trading posts, indicating the **amount to be paid to transport or unload merchandise**.
Expressions such as **"para de muage"** specifically referred to the **sum intended to pay this tax**.

In summary: ***Muage*** **= transport/loading fee in ports or rivers, widely used in archaic Portuguese and administrative documents from the Age of Discovery.**

Would you like me to **provide examples of the word's occurrence in medieval Portuguese documents** (such as royal charters or town grants)?

Table 2. *Sabiá-3.1* Response < "What does *muage*? mean">

Prompt: What does muage mean?
Sabiá-3.1

"*Muage*" is a slang or colloquial expression found in some regions of Brazil, but it is not widely recognized or used. It may reflect a regional variation or be a typo/corruption of another, more common expression.

One possibility is that "*muage*" is a variation of "*muganga*," which in some contexts can mean a playful act, a grimace, or a mannerism. However, this interpretation may not apply in all contexts and is not widely used.

Another possibility is that "*muage*" is a mispronunciation or misspelling of "*mauagem*," which does not have a specific meaning in Portuguese, or perhaps a confusion with "*mugir*," the sound made by cattle, though this connection seems less likely in colloquial usage.

Since slang and colloquial expressions can be highly regional and contextual, and without a clear context, it is difficult to determine the exact meaning of "*muage*." If you heard this expression in a specific context or region, could you provide more details for a more precise explanation?

When comparing the definitions generated by the two AI systems, certain aspects of the AI algorithm's functioning in relation to its training dataset become apparent. Notably, ChatGPT makes a categorical claim that the word 'muage' is little used today. This assertion contrasts with the linguistic reality in Brazil, where the word circulates among millions of speakers, particularly in the Midwest and Northeast regions. This discrepancy between attested reality and the AI's response provides evidence that linguistic data from these regions are not included in ChatGPT's dataset. By responding with such a definitive statement, the system accentuates this linguistic erasure. ChatGPT presents meanings of a historical nature, linked to the French etymology of the word and its occurrence in official documents produced in Portugal between the 15th and 16th centuries, categorizing this as the 'primary meaning' of the term. This classification itself leaves room to consider that the word may indeed have other, non-primary meanings. Importantly, there is no reference to Brazil or to the term's contemporary meanings within the national territory.

In responding to the same prompt, the *Sabiá* chatbot demonstrates some knowledge of Brazilian slang and colloquial expressions and also mentions regional linguistic variation. Its definition does not establish the existence of a principal meaning of the word, nor does it produce categorical statements; on the contrary, it leaves room for uncertainty through expressions such as 'it could be' or 'one possibility.' The AI even acknowledges that "it is difficult to determine the exact meaning of 'muage.'" Moreover, it refers to the sense of "playful act, grimace, or mannerism," which corresponds to one of the meanings effectively attested for the word. However, in certain regions of the Brazilian Midwest, particularly in the state of Mato Grosso, the word is also employed with the meanings of 'idle talk' and 'fussiness' – uses that do not appear among the results provided.

From a modal perspective, the two AI systems diverge: ChatGPT adopts a modal regime of certainty, whereas *Sabiá* operates within a modal regime of doubt. ChatGPT produces categorical statements and reinforces them with illustrative examples. Furthermore, its response employs boldface highlights to direct the user's attention and provides a summary in which the lexical item is followed by a mathematical equal sign (=). This compositional strategy reasserts values of truth and completeness. By contrast, *Sabiá* establishes a modal regime of doubt, and its response acquires a more open-ended and reflective trait. Both systems conclude by inviting further interaction; however, while ChatGPT does so by offering additional examples that confirm its initial answer, *Sabiá* explicitly requests more contextual information that could allow for a reformulation and thus enhance the precision of its response.

Let us now examine the performance of the two models regarding peseiro, the second lexical item selected for this experiment. The term peseiro can be used to denote a difficult or troublesome situation that the speaker is experiencing, roughly corresponding to expressions such as 'tough spot' or 'hard situation' in English.

Table 3. ChatGPT–5 Response: <What does *peseiro* mean?>

Prompt: What does *peseiro* mean?

The word **"peseiro"** can have several different meanings depending on the context:

1. Surname
Very common in Portugal, it appears as a family name.
Famous example: José Peseiro, Portuguese football coach.

2. Regionalism / Archaism (Portugal)
In some older records of the language, "*peseiro*" is related to *peso* (to weight) and could mean **a scale** or the **place where goods are weighed**.
It was also used to refer to someone who works with the weighing of grains or merchandise.

1. Toponym
Peseiro is the name of several **places in Portugal**, usually small villages or localities.

In summary: today, the most common usage is as a Portuguese **surname**; however, historically, "**peseiro**" could refer to someone involved in the act of **weighing goods** or even **the place where weighing occurred**.

Would you like me to look for examples in **old Portuguese texts** where the term appears in the sense of a **profession/place**?

The same test conducted with the *Sabiá* IA produced the following result:

Table 4. *Sabiá*-3.1 Response: <What does *peseiro* mean?>

Prompt: What does *peseiro* mean?
Sabiá-3.1

"*Peseiro*" is a term that can have different meanings depending on the context. Here are some possible interpretations:

1. In music: *Peseiro* is a popular musical style in Cape Verde, characterized by a slow and melancholic rhythm. It is often associated with themes of love and longing.

2. Slang: In some regions of Brazil, "*peseiro*" can be used as slang for something heavy or difficult to deal with, although this usage is not widely known or commonly employed.

3. Regionalism: In specific regional contexts, it may have other local meanings or it may be a word used in dialects or regional languages, though these usages are less common and less widely known.

If you are referring to a specific context — such as a regional term or a term used within a particular domain — please provide more details so that I can offer a more precise explanation.

In both tests, the AI programs recognize that the word can have multiple meanings and note contextual variations. However, ChatGPT frames the context exclusively within Portugal, without any reference to Brazil or other Portuguese-speaking countries. By contrast, *Sabiá* identifies 'peseiro' as a musical style in Cape Verde, a Portuguese-speaking African country, possibly assuming a connection to 'piseiro,' a variation of the 'pisadinha' musical style that originated in Bahia (Northeast Brazil) and is also danced in Cape Verde.

Both systems conclude their responses in a manner very similar to the previous test. ChatGPT provides additional examples to confirm the answer, whereas *Sabiá* requests more information to refine and increase the precision of its response. This manner of closing the interaction appears systematic, suggesting that ChatGPT is programmed to confirm the information provided without leaving room for doubt, whereas *Sabiá* is designed to highlight the incompleteness of its answers.

Based on the concept of enunciative praxis and modes of existence (Fontanille 2003), D'Armenio et al. (2024) argue that generative AIs, in the process of text generation, call upon virtualized forms available in our encyclopedic knowledge. The prompts then perform a triage process within certain zones of the databases used by an AI system, enacting some of these virtualities.

In generative AIs such as ChatGPT and *Sabiá*, the model generates new text by associating words (prompts) with embeddings. The initial procedure involves tokenization, which divides the texts present in the dataset into minimal units called tokens. These can consist of a full word, a segment of a word (e.g., a suffix), punctuation, or other elements. Each token is then converted into a numerical vector, the embedding, which enables semantic categorization: tokens that are semantically close tend to occupy contiguous regions in the vector space, while those that are more distant occupy separated positions. Semantic proximity or distance is calculated based on the statistical recurrence of token occurrences and the combinatorial possibilities of tokens present in the training data.

When we enter a prompt, the model interprets the corresponding embedding and, based on previously learned patterns, associates it with probable meanings, mobilizing isotopies, figures, and recurring themes. This is, therefore, an operation that does not originate ex nihilo, but rather involves the reuse and reorganization of discursive virtualities already inscribed in the dataset.

Thus, in the results presented in the tests – where prompts were directed to the generative artificial intelligences ChatGPT and Sabiá, requesting the meaning of Brazilian linguistic expressions, the prompt-text assumed the function of a 'search command' (Meyer 2023), capable of guiding the model through statistical procedures and the valuation regimes established by the algorithms. However, the results produced by each model are subject to the coercion inherent in the very constitution of its dataset: the way it was constructed and the valuation criteria applied to its contents. The recurrence and semantic proximity of the data directly condition the creation of novel textualizations produced by the AI. Such texts generate meanings and circulate socially, fixing the meanings inscribed within them.

The words tested here cannot be understood as mere lexical variants; they condense meanings that unveil the organization of *practical scenes* in which they are produced, as well as the knowledge systems and cosmologies that structure collective modes of existence – *forms of life*. As they circulate, they integrate into networks of enunciative practices and reveal the congruence of shared values and experiences. As demonstrated in the examples, the absence of certain words in Artificial Intelligence models, the disregard for linguistic variation, and the exclusive reference to the European origin of the word as its sole possible meaning, highlight in ChatGPT the erasure of the cultural specificities of marginalized groups and reinforce the historical, linguistic, and semiotic hegemony of the languages and varieties of the Global North, along with their associated *practices* and *forms of life*.

Some additional considerations

This article aimed to shed light on how meaning emerges in generative AIs from a Global South perspective. To this end, we followed the process in its entirety – from the initial datafication to the generation of the AI-generated semiotic object. In this way, we sought to conceive datafication as a kind of textualization, which often allows us to apprehend *practices* – or at least some traces of them – within a semiotic object manifested on a more stable medium, and thus manipulable by algorithmic devices. By doing so, we reinforce the semiotic nature of Big Data, highlighting that its organization is permeated by axiologies that will inform the decisions made by such machinic agents.

In this manner, within the datafication process, datasets cease to be mere data repositories: they assume the role of operators that select, reiterate, and stabilize isotopies present in the semiotic objects composing these large collections. This operation, by privileging content originating from the Global North, establishes hegemonic patterns that present themselves as universal, simultaneously erasing or exoticizing the *practices*, languages, and *forms of life* of the Global South.

The tests conducted with ChatGPT and *Sabiá*, even on a small scale, demonstrate how the nature of the training data directly affects signification. ChatGPT, anchored in the values of 'globalization' and trained predominantly on English-language data, tends to overlook Brazilian usages of certain words, privileging historicist and Lusitanian readings. In contrast, *Sabiá*, fed with national data in Portuguese, showed greater recognition of local isotopies, accounting for linguistic variation. From a semiotic perspective, this confirms the central hypothesis: these models do not 'invent' *ex nihilo*, but rather realize virtualities available in their repositories (through machinic co-enunciation).

The experiment demonstrated how differences in dataset composition reverberate in the construction of meaning. However, the mere use of regional datasets, although necessary, is insufficient to mitigate the problem, as these datasets may themselves contain prejudiced biases in their construction. The development of more ethical and culturally plural AIs, aimed at reducing structural asymmetries between the Global North and Global South, also requires semiotic curation of data – understanding them as semiotic objects carrying both expression and content planes. Implementing this practice has the potential to improve decision-making in the collection, balancing, and utilization of such data, from the process of datafication to the very constitution of large-scale data collections.

References

Abonizio, H., Almeida, T. S., Laitz, T., Malaquias Júnior, R., Kerche Bonás, G., Nogueira, R., & R. Pires 2024. *Sabiá-3 Technical Report* (arXiv:2410.12049) [Preprint]. arXiv.

Aldama, J. A., Bertrand, D., Darras, B., & Di Sciullo, F. (eds.) 2021. *Sémiotique impliquée: L'engagement du chercheur face aux sujets brûlants*. Paris: L'Harmattan.

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

Basso Fossali, P. (2012). Texte préparatoire au dossier. *Actes Sémiotiques* 115. https://www.unilim.fr/actes-semiotiques/4928

Boyd, D., and K. Crawford 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society* 15(5): 662–679. https://doi.org/10.1080/1369118X.2012.678878

Coeckelbergh, M. 2023. *Ética na inteligência artificial*. Rio: Ubu Editora, Editora PUC.

D'Armenio, E., Deliège, A., & M. G. Dondero 2024. Semiotics of machinic co-enunciation: About generative models (Midjourney and DALL·E). *Signata* 15. https://doi.org/10.4000/127x4

Fontanille, J. 2003. *Sémiotique du discours*. Limoges: France: PULIM.

Fontanille, J. 2008. *Pratiques sémiotiques*. Paris: PUF.

Fontanille, J. 2015. *Formes de vie*. Liège: Presses universitaires de Liège.

Fontanille, J. 2017. Práticas e formas de vida: A semiótica de Greimas posta à prova pela antropologia contemporânea. *Estudos Semióticos* 13(2): 66–76. https://doi.org/10.11606/issn.1980-4016.esse.2017.136307

Fontanille, J. 2019. Discursos, mídias, práticas e regimes de crença. *Revista do GEL* 16 (3): 246–261. https://doi.org/10.21165/gel.v16i3.2608

Fontanille, J. and C. Zilberberg 2001[1998]. *Tensão e significação.* São Paulo: Humanitas.

Greimas, A. J. 1993. *Les formes de vie* [Dossiê]. *Recherches Sémiotiques/Semiotic Inquiry* 13(1–2).

Greimas, A. J. and J. Courtés 1982[1979]. *Semiotics and language: An analytical dictionary*. Bloomington, IN: Indiana University Press.

Hjelmslev, L. T. 2010[1975]. *Prolegomena to a theory of language* (F. J. Whitfield, Trans.). Madison: University of Wisconsin Press.

Ghosh, S., Kapoor, A., Rane, A., Mahendiran, A. and M. Thakkar 2025. Documenting patterns of exoticism of marginalized populations within text-to-image generators. *arXiv.* https://arxiv.org/abs/2508.02937

Hagerty, A. and I. Rubinov 2019. Global AI ethics: A review of the social impacts and ethical implications of artificial intelligence. *arXiv*. https://arxiv.org/abs/1907.07892

Jay Agrawal, J. G. and G. Avi 2018. *Prediction Machines: the simple economics of Artificial intelligence*. Cambridge, MA: Harvard Business Review Press.

Kaplunovich, A. 2023. 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*, 4654–4663. IEEE. https://doi.org/10.1109/BigData59044.2023.10386329

Mayer-Schönberger, V. and K. Cukier 2013. *Big data: A revolution that will transform how we live, work, and think*. Boston, MA: Houghton Mifflin Harcourt.

Marrone, G. 2008. L'invention du texte. *Actes Sémiotiques* 111. https://www.unilim.fr/actes-semiotiques/1643

Mejias, U. A. and N. Couldry 2024. A colonização pelo dado: Do colonialismo histórico ao colonialismo de dados. *Revista Sociedades Híbridas* 4(1). https://revistas.hcte.ufrj.br/index.php/RevistaSH/article/view/446

Meyer, R. 2023. The new value of the archive: AI image generation and the visual economy of 'style'. *IMAGE. Zeitschrift für interdisziplinäre Bildwissenschaft* 19(1): 100–111. https://doi.org/10.25969/mediarep/22314

Moraes, L. 2021. *A noção de texto na semiótica* (Tese de doutorado, Universidade de São Paulo, Faculdade de Filosofia, Letras e Ciências Humanas). https://doi.org/10.11606/T.8.2021.tde-29112021-185047

Moraes, L. 2024. *Semiótica e big data: O valor da "textualização" na lógica capitalista da cultura dataficada*. In: P. Gonçalves-Segundo et al. (Orgs.), *Tecendo sentidos: investigações em estudos linguísticos, textuais e discursivos* (Vol. 1, pp. 252–272). FFLCH/USP.

Moraes, L.2025. O que pode o(a) semioticista na era da inteligência artificial? Semiótica, big data e racismo algorítmico. In: J. C. Portela, M. N. Schwartzmann, T. M. Corrêa, F. C. R. Santos, & C. A. Tassinari (Orgs.), *Identidade, experiência e discurso: Semiótica e crítica da cultura* (Vol. 1, 139–168). Campinas, SP: Pontes Editores.

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

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

Seaver, N. 2018. What should an anthropology of algorithms do? *Cultural Anthropology* 33(3): 375–385. https://doi.org/10.14506/ca33.3.04

Shahid, S., Elswah, M. and A. Vashistha 2025. Think outside the data: Colonial biases and systemic issues in automated moderation pipelines for low-resource languages. *arXiv*. https://arxiv.org/abs/2501.13836

Tifentale, A. and L. Manovich 2015. Selfiecity: Exploring photography and self-fashioning in social media. In: D. Berry & M. Dieter (Eds.), *Postdigital aesthetics: Art, computation and design*, 109–122. Palgrave Macmillan. http://manovich.net/index.php/projects/selfiecity-exploring

Van Dijck, J. 2014. Datafication, dataism and dataveillance: big data between scientific paradigm and ideology. *Surveillance & Society* 12(2): 197–208.

Vilarino, R. and R. Vicente 2020. An experiment on the mechanisms of racial bias in ML-based credit scoring in Brazil. *arXiv*. https://arxiv.org/abs/2011.09865

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

AUTHOR

Letícia Moraes, Adjunct Professor, Department of Linguistics and Portuguese Language, UFPB, Paraíba-PB, Brazil.

AUTHOR

Silvia Sousa, Associate Professor of Linguistics, Department of Language Sciences, UFF, Niterói, RJ, Brazil.