Book Chapter for the forthcoming *Artificial Knowledge of Language: A Linguist's Perspective on its Nature, Origins and Use*, Edited by José-Luis Mendívil-Giró

The Creative Aspect of Human Language Use

How Modern Language Models Fare with an Old Idea

By Vincent J. Carchidi

Abstract

Ordinary human language use is *stimulus-free*, *unbounded*, yet *appropriate and coherent* to the circumstances of its use, a characterization typically associated with Noam Chomsky and other generative linguists. This tripartite depiction of linguistic behavior is distinctive in that it is unique within the animal world and central to any scientific account of human language. The successes of Large Language Models (LLMs) in Natural Language Processing—chiefly in their ability to produce human-like text—altogether fail to shed any light on this "creative aspect of language use" (CALU). Such models fail to reproduce this ability without exception. Moreover, they offer no account of how humans acquire the ability to use their language in this tripartite fashion and what role this should play in a theory of human language. Against this background, I argue that Steven Piantadosi's (2023) critique of Chomsky's approach to linguistics (and related arguments) is misguided. More specifically, I argue that the failure to shed any scientific light on CALU amounts to a deficiency of a methodology that considers LLMs accurate models of human language and a vindication of the "Galilean" style of inquiry employed by Chomsky.

> "It is not a novel insight that human speech is distinguished by these qualities, though it is an insight that must be recaptured time and time again. With each advance in our understanding of the mechanisms of language, thought, and behavior, comes a tendency to believe that we have found the key to understanding man's apparently unique qualities of mind."
>
> - Noam Chomsky, *Language and Mind.*

Introduction

In March 2023, Steven Piantadosi (2023) published a wide-ranging critique of Noam Chomsky's approach to linguistics. "After decades of privilege and prominence in linguistics," he writes, "Noam Chomsky's approach to the science of language is experiencing a remarkable downfall" (Piantadosi, 2023, 1). The source: computational advances embodied in contemporary Large Language Models (LLMs). LLMs, he argues, are "bona fide linguistic *theories*" (Piantadosi, 2023, 7). Their ability to produce human-like grammatical utterances and coherent discourses is rightly considered an unprecedented simulation of human language's "high-level properties" (Piantadosi, 2023, 27). These models succeed, moreover, without methodological steps invoked by Chomsky in the study of human language, including requirements that LLMs "properly consider competence vs. performance, respect "minimality" or "perfection," and avoid relying on the statistical patterns of unanalyzed data" (Piantadosi, 2023, 15).

Responses to Piantadosi's critique and related arguments have already appeared (e.g., this volume; Katzir, 2023; Kodner, Heinz, and Payne, 2023; Leivada, Dentella, and Murphy, 2023; Milway, 2023; Moro, Greco, and Cappa, 2023; Rawski and Baumont, 2023).[1] These responses tackle matters including the poverty of the stimulus, the learnability and unlearnability of "possible" and "impossible" languages, and elements of linguistic theorizing such as explanatory adequacy and the deeper "why" questions about the nature of human language.

This chapter goes to the heart of this debate by focusing on a peculiarly absent notion associated with Chomsky's work: the "creative aspect of language use," or CALU. CALU refers to the notion that human language use is ordinarily *stimulus-free*, *unbounded*, yet *coherent and appropriate* to the circumstances of its use (Chomsky, 1968, 10-11; McGilvray, 2001, 6-13; Asoulin, 2013, 228-232).

Chomsky has maintained for decades that observations of this human ability can be traced back to Descartes' (1637) *Discourse on Methods* and subsequent Cartesian work (see, Chomsky, 2009a). The history of this concept is tied to the distinction between humans and machines, with the latter permanently lacking the ability to reproduce CALU. Just as Descartes distinguished between animals and machines, on the one hand, and human beings on the other, Chomsky has characterized the Turing Test as a 'resurrection' of the "Cartesian tests for the existence of other

[1] Note that, of these responses, Leivada, Dentella, and Murphy (2023), Milway (2023), and Rawski and Baumont (2023) will appear as chapters in the forthcoming volume.

minds" (Chomsky, 1988, 141). If a machine exhibits the use of language in a simultaneously stimulus-free, unbounded, yet appropriate fashion, it can reasonably be said to possess a "mind like ours" (Chomsky, 1988, 5).

Given that LLMs are simulations of human linguistic performance, it is striking not to find CALU cropping up more in discussions on their efficacy as models of human language. More specifically, given that Chomsky has long considered CALU to represent a set of facts about ordinary human language use with which "any science of language must contend" (McGilvray, 2001, 5), it is unusual not to find explicit discussion of this matter in Piantadosi's critique purporting to undermine the former's "approach" to linguistics.

The force of this chapter's argument is not exclusively in the claim that LLMs do not exhibit CALU. While I argue they do *not* use language in a manner that is stimulus-free nor appropriate—and can only be considered unbounded in a "weak" sense—this breakdown is illustrative and serves a deeper point: the attempt to explain human language via its wholesale simulation eschews the very conception of scientific inquiry that gives generative linguists an ability to both identify and grapple meaningfully with CALU and its role in a theory of language.

Generative linguistics is advantaged by its readiness to create an *abstraction* of the human language capacity. Concurrent with this abstraction is the selection of *crucial* data—very specific linguistic data whose relationship to theory yields deeper insight into this phenomenon than the depth of insight yielded by broader coverage of data. From the abstraction of language and the selection of crucial data, a distinction between an individual's knowledge or *understanding* of language—competence—and the *use* of this linguistic knowledge in concrete circumstances—performance—is justifiably drawn.

When generative linguists abstract away from specific uses of human language, they identify certain data that they argue are crucial to understanding it—the human being's reflexive acquisition of the language of their community and mastery of a rich linguistic understanding relative to the data to which they are exposed represents a set of observations that must be dealt with by a theory of language. Less commonly articulated, though no less central, is that human language use is *stimulus-free*, *unbounded*, yet *appropriate and coherent* to circumstance.

As this chapter demonstrates, this chain of reasoning is embodied in the "Galilean" method employed by Chomsky (1980) and explicitly rejected by Piantadosi (2023, 26-28), who argues that LLMs refute this method. An extensive appraisal of this method is thus provided alongside a breakdown of how LLMs fail to meet the challenges posed by CALU.

Specifically, it is shown that—in addition to not exhibiting CALU—LLMs taken to be theories of human language in their own right, or significant insights into human language, have three major shortcomings: (1) Their broad coverage of linguistic data comes at the expense of selecting particularly crucial data related to human language use that yield comparatively deeper insight; (2) LLMs-as-theories have unrealistically high expectations placed on their explanatory power and, as a result, gloss over the theoretical challenges posed by the *appropriateness* of stimulus-free and unbounded human language use; and (3) Sorely neglect the inextricable relationship between human language and thought, specifically in the human ability not only to construct new concepts but distinctively choose to engage this combinatorial cognition owing to CALU.

The chapter is structured as follows: we begin with a review of the literature surrounding Piantadosi's critique of Chomsky's approach to linguistics. Then, with a background established on the debate over the poverty of the stimulus, possible and impossible languages, and the notion of explanatory adequacy, we delve into the Galilean method and how Chomsky articulates it. Following this, we explore the implications of this method for CALU, its role in a science of language, and how LLMs fail to reproduce this ability or offer explanatory insight.

Large Language Models and Generative Linguistics

Piantadosi's (2023) critique represents the latest episode in a long-running drama about the utility of generative linguistics, particularly as articulated by Chomsky. The piece drew from the momentum generated by the success of OpenAI's (2022) ChatGPT-3.5. Piantadosi's central claim—that the Chomskyan approach to linguistics is undermined by computational advances—is not especially surprising given Chomsky's intra-field status as a "piñata" (Horgan, 2016).

Nevertheless, the grammatical output of LLMs, combined with an ability to maintain some level of internal consistency within their responses to prompts, rightfully triggers a rethink of fundamental claims pertaining to themes in generative linguistics such as innate linguistic knowledge, the poverty or richness of linguistic data, and the competence-performance distinction. Indeed, Piantadosi is far from alone in this respect—many scholars see direct implications for the study of human language from the rise of contemporary language models (e.g., Beguš, Dąbkowski, and Rhodes, 2023; Kallens, Kristensen-McLachlan, and Christiansen, 2023; O'Grady and Lee, 2023; Wilcox, Futrell, and Levy, 2023).

On March 8, 2023, Noam Chomsky, Ian Roberts, and Jeffrey Watumull wrote in *The New York Times* on the "False Promise of ChatGPT." These authors contrast "ChatGPT and its ilk, a lumbering statistical engine for pattern matching" with the human mind which they deem "a surprisingly efficient and even elegant system that operates with small amounts of information…" (Chomsky, Roberts, and Watumull, 2023). Their central claim is that there is a sharp distinction to be drawn between how humans acquire language and engage in linguistic cognition and how LLMs generate linguistic outputs.

Piantadosi's critique goes to the heart of this alleged distinction. An illustration of his central argument is given below.

Piantadosi's Critique of Chomsky's Approach

It is worth noting that while principally framed as a direct critique of Chomsky's "approach," Piantadosi's critique frequently veers into the broader territory of generative linguistics. This may be because, as Piantadosi (2023, 35) himself acknowledges in the critique's postscript, the nature of *scientific inquiry* seems to be lurking behind arguments between himself and his critics.

This observation is crucially relevant as this chapter seeks to highlight how Piantadosi's depiction of modern language models as linguistic theories—and more specifically, theories undermining Chomsky's approach—neglects some of the core methodological stances held by Chomsky and how this leads to divergent characterizations of the study of language, exemplified by CALU's role in generative linguistics.

Nevertheless, the central claims with respect to language models' relationship to linguistics made by Piantadosi are (1) They are linguistic theories in their own right; and (2) These theories undermine the Chomskyan approach to linguistics. Piantadosi argues that modern language models, because they "develop representations *of* key structures and dependencies," therefore "should be treated as bona fide linguistic *theories*" (Piantadosi, 2023, 7). He is adopting the position of Marco Baroni here, who argues that it is "appropriate…to look at deep nets as *linguistic theories*, encoding non-trivial structure priors facilitating language acquisition and processing…a general theory defining a space of possible grammars" prior to language-specific training (Baroni, 2022, 7).[2]

The claimed refutation of Chomsky's approach can be found in the idea that language models capture human-like representations of syntactic and semantic structure—they reproduce, in other words, the very aspects of natural language that linguists seek to explain. More specifically, Piantadosi's (2023, 5-8) argument is that—in the course of training on vast amounts of linguistic data—language models navigate a *space of possible theories*. Eventually, by *testing certain hypotheses* through said data and determining which hypotheses lead to the *desired output*, they settle on a computational procedure that gleans a human-like grasp on language *directly from* the data themselves.

Acknowledging that "any model will necessarily have certain tendencies and biases," Piantadosi thus conceives of "each model or set of modeling assumptions as a possible hypothesis about how the mind may work. Testing how well a model matches humanlike behavior then provides a scientific test of that model's assumptions" (Piantadosi, 2023, 11).

Piantadosi further argues that LLMs "embody several core desiderata of good scientific theories" (Piantadosi, 2023, 12): they are sufficiently *precise and formal* accounts to be implemented in computational systems, they make relevant *predictions*, such models can be *integrated* with knowledge originating in fields outside of linguistics, and they are *empirically testable* (Piantadosi, 2023, 12-14). The result of this reasoning is that LLMs serve as theories of *human* language by testing the space of possible linguistic theories—and they develop relevant linguistic representations "in a way which is unfamiliar to linguistics" (Piantadosi, 2023, 7).

In the postscript, Piantadosi (2023, 33) refers to this as a form of *inductive* reasoning, one that is common to scientific theorizing that (rightly, in his view) *predicts experimental data*, rather than being deduced *from* experimental data. In this way, the argument is that modern language models

[2] Baroni (2022, 6-7) takes a significantly firmer stance than Piantadosi on the matter of innate priors and language models. This difference in emphasis is no problem for Piantadosi (2023, 28)—who is decidedly *not* a proponent of blank slate arguments, arguing in favor of the reality of cognitive structures—but the strength of deep neural networks' innate priors relative to humans' innate biases is worth consideration.

can and do shed light on human language acquisition because they accurately predict the structure of human language via its simulation. This invocation of inductive reasoning hints at deeper issues of methodology and the formation of explanatory theory that will recur throughout this chapter.

Responses to Piantadosi

Naturally, not everyone agrees. Among the earliest responses was by Katzir (2023) who argues that "what makes Piantadosi's paper so surprising is his suggestion that LLMs are good theories of (actual) human cognition" (Katzir, 2023, 1-2). He suggests that Piantadosi's central claim is "akin to discovering that a newly designed drone accidentally solves an open problem in avian flight" (Katzir, 2023, 2).

Katzir runs a series of brief tests with ChatGPT-4 including (but not limited to): an assessment of its grammaticality judgments (Katzir, 2023, 2-4); probing its linguistic competence by presenting it with a center-embedded construction (Katzir, 2023, 5-7); and noting the system's failure to distinguish between *likelihood* and *grammaticality* in predicting the next word of a given sentence (Katzir, 2023, 7-8). Finding that ChatGPT-4 fails to produce responses comparable to a human's, he concludes that "Piantadosi's excitement is premature" (Katzir, 2023, 9). In other words, Katzir claims his findings contradict the claim that language models like GPT-4 exhibit a human-like grasp of linguistic structures once one peeks below the surface.

In another response, Rawski and Baumont (2023) argue within a single page that Piantadosi "fallaciously affirms the consequent" by moving from the premise that LLMs *correlate* with human behavioral data to the conclusion that LLMs, therefore, serve as a *theory* of human behavior. In response to this claim, Piantadosi (2023, 33) reaffirms the *inductive* nature of his argument, noting that the prediction of experimental data is of prime importance in physics and biology, and so, too, should it be for linguistics.

Rawski and Baumont (2023) also liken Piantadosi's reasoning to the mistake of pre-modern geocentric models in physics which provided high predictive accuracy but lacked the *explanatory* power of later, heliocentric models. The implication is that Piantadosi has mistaken the *predictive* power of LLMs in predicting the next token[3] in a human-like fashion with *explanation*, the latter of which may sacrifice predictive accuracy in exchange for depth of understanding. In a similar response, Piantadosi (2023, 33-34) reaffirms his emphasis on data and quantitative testing which he believes marks the successes of physics.

Other responses addressed dimensions relevant to the Argument from the Poverty of the Stimulus (APS) and the learnability or unlearnability of "possible" and "impossible" languages. Kodner, Payne, and Heinz (2023) argue that Piantadosi's invocation of LLMs in explaining human language acquisition neglects the original purpose of the APS:

> Under this hypothesis, children generalize from their limited input in specific ways, navigating a constrained space of possible natural language grammars. Consequently,

[3] "Tokens" are units that represent words or parts of words for LLMs.

> they do not consider all logically possible generalizations that are consistent with their linguistic experience. Rather, the particular structure of the hypothesis space facilitates the rapid development of their linguistic capabilities (Kodner, Payne, and Heinz, 2023, 2).

The authors distinguish the argument above from the "relatively unconstrained" learning Piantadosi (2023, 18) attributes to LLMs.

Kodner, Payne, and Heinz further suggest that—contra conventional wisdom in some corners—LLMs may *not* be so constrained. They question the need for machine learning researchers to "constantly tinker with the layers, the gating mechanisms, the architectures, and the tuning of the hyperparameters" of artificial neural networks in attaining the desired results (Kodner, Payne, and Heinz, 2023, 2). "But what are these biases, principles, and limitations," they ask, "if not some form of the Universal Grammar that Piantadosi (2023, 19) claims LLMs prove "to be wrong" (Kodner, Payne, and Heinz, 2023, 5)?

Piantadosi believes this is "deeply mistaken. There are no doubt *some* principles required for language. The question is whether they are language-specific (or syntax-specific), innate, and whether they have the form that Chomsky and similar theorists have said, as opposed to lower-level principles that work through emergence" (Piantadosi, 2023, 38).

Piantadosi (2023, 39) also argues that whatever priors "relatively constrained" language models do possess, they do not need to be as strong as generativists suggest. He hints that some generativists are contradicting themselves by simultaneously saying language models are *not* equipped with the appropriate innate biases because they learn impossible languages and that language models *are* so innately constrained as to be considered a form of Universal Grammar.

There are two overlapping matters here: the APS and possible and impossible languages. The common thread between them is an underlying dispute over whether modern language models do, in fact, capture language-specific, or syntax-specific, features of human language. We explore this and then tease out its implications for the APS and possible and impossible languages.

There is research outside of direct responses to Piantadosi on such matters. For example, Wilcox, Futrell, and Levy (2023) test the outputs of two Long Short-Term Memory (LSTM) models and two Transformer models (GPT-2 and GPT-3, notably). They argue that the results demonstrate, empirically, that the input is *not* as impoverished as proponents of the APS suggest in the domain of filler-gap dependencies and their island constraints. In direct response, Lan, Chemla, and Katzir (2023) test the outputs of two LSTM models and four Transformer models (including GPT-2 and GPT-3). The authors explicitly establish relatively lenient conditions for success in judging the grammaticality of examples of parasitic gaps and across-the-board movements but find that these models often fail to approximate wh-movement, and thus do not challenge the APS in this domain.

Furthermore, in systematic testing of three language models—GPT-3/text-davinci-002, GPT-3/test-davinci-003, and ChatGPT-3.5—Dentella, Günther, and Leivada find that all three models exhibit

> marginal overall above-chance accuracy and absence of response stability…the [language model] answers to questions tapping into the (un)grammaticality of prompts that pertain to different language phenomena are largely inaccurate (RQ1), ever-changing (RQ2), and not playing out in favor of a strategy that culminates in either more stable or more accurate answers (RQ3) (Dentella, Günther, and Leivada, 2023, 6).

Language models such as these thus exhibit an unusual contrast with humans: they simultaneously *appear* to master the form of language but do not produce the accurate and comparatively stable grammaticality judgments that *should* result from their apparently human-like linguistic competence.

As the authors explain, language models' "[i]nsensitivity to (un)grammaticality amounts to a qualitative mismatch problem…LMs which cannot contribute to the description of the language they have been trained on, by recognizing and ruling out its ungrammatical instances, fail to be "observationally adequate" (Dentella, Günther, and Leivada, 2023, 8). More than this, LMs receive systematic *negative feedback* for ungrammatical constructions; human children do not. This leads to a "double mismatch" in which "i) LMs receive rich information about strings of words that correspond to grammatically wrong sentences, while humans do not, but still, ii) they are not able to accurately judge grammatically wrong sentences as such, while neurotypical humans can" (Dentella, Günther, and Leivada, 2023, 8).

Leivada, Dentella, and Murphy (2023) expound on these results and further assessments of ChatGPT, observing that claims regarding the human-like linguistic competence of LLMs do not hold up to scrutiny. In particular, the inconsistent application of grammatical rules by ChatGPT (e.g., detecting attraction errors) is sharply out of line with human consistency in such domains. LLMs may instead be engaging in a probabilistic endeavor that does not rely directly on learned rules of grammar (Leivada, Dentella, and Murphy, 2023, 4-5).

While research techniques for probing the linguistic competencies of language models will doubtless evolve, it does not appear that state-of-the-art models possess the decisive grasp on elements of syntax that Piantadosi suggests—at least not in a manner sufficiently comparable to the accuracy and consistency of human linguistic competence. This is echoed in a general context by Shalom Lappin who, while explicitly *unsympathetic* to Chomsky's perspective on LLMs, states that Piantadosi's claim that LLMs serve as models of human language acquisition "go well beyond the evidence" (Lappin, 2023, 6).

With this in mind, we turn to the matter of possible and impossible languages and how this relates to the APS.

For background, some generative linguists (e.g., Chomsky and Moro, 2022; Moro, Greco, and Cappa, 2023) have argued that the ability of LLMs to acquire "impossible" languages just as readily as "possible" languages—an ability they allege is not shared by humans, with the former treated akin to puzzle-solving (Milway, 2023, 3)—indicates that human linguistic cognition is constrained in ways that LLMs are not, marking a sharp disanalogy between the two. Chomsky argues that there is a sense in which LLMs are *too strong*—they make no distinction between possible and impossible languages, capable of learning and mastering either with sufficient data. LLMs are, then, not models of human language because human linguistic cognition is not amenable to learning that is unconstrained (Milway, 2023, 3).

Piantadosi rejects this view. In the critique's postscript, he laments the lack of attention paid to the *amount of data* required for LLMs to acquire possible rather than impossible languages:

> There is historical irony to this move because, without missing a beat, the field switched from saying that innateness was true because learning was impossible ("poverty of stimulus"), to saying that learning can't be right because it works too well. But also, simultaneously somehow, we should doubt the models since they don't work on small amounts of data?…Of course, nobody says what timing for learning would be evidence *against* innateness (Piantadosi, 2023, 36).

Piantadosi appears to be making two claims here: (1) That the alleged purpose of the APS was to pinpoint the *un*learnability of mature linguistic competence in humans; (2) Proponents of the APS have contradicted their position by claiming LLMs do not model human linguistic competence because they learn *too well*, thereby undermining their claim that such learning is impossible.

Are generative linguists contradicting themselves? Milway suggests that these claims have been misinterpreted:

> Far from claiming that [Modern Language Models] could never do the things that ChatGPT can do, Chomsky has repeatedly claimed the opposite—that with enough data and computing power, a statistical model would almost certainly outperform any scientific theory in terms of empirical predictions (Milway, 2023, 2).

The APS does not imply that there is no amount of data for which a language could be used to train a model to provide the kind of predictive power of ChatGPT. It simply implies that the data to which humans are exposed during development—comparatively less in *amount* and different in *kind* than that of LLMs—is not sufficient to account for their mastery of natural languages.

In fact, in one of the few sources from which Piantadosi (2023, 26) directly quotes Chomsky on matters related to methodology and statistical approaches to language—a 2012 interview with Yarden Katz—Chomsky repeatedly notes that one *can* attain an "approximation" of unanalyzed data—what is in doubt is the scientific value of the approximation (Katz, 2012). There does not appear to be any contradiction between the persuasive approximation of human text-based language use via statistical modeling, on the one hand, and the development of a biologically endowed linguistic competence, on the other. Tension only arises when one tries to have the former *explain* the latter (or vice versa).

That said, there is a sense in which the APS is not a typical "argument." When Chomsky compares his granddaughter with a kitten and suggests that only one—with all the relevant data—acquires language while the other never does, he is articulating the parameters for a research program (Collins, 2008, 102). This is why criticisms of this idea, sometimes amusingly called the "rocks-and-kittens" argument (e.g., Behme, 2014, 22-24; Postal, 2005, ix-x) amount to a criticism that Chomsky pays no attention to new empirical studies.

The fundamental disagreement over "rocks-and-kittens" boils down to what Pullum (2011, 2)[4] identifies as the distinction between *general* nativism and *linguistic* nativism—the difference between having innate structure as opposed to having innate structure specific to language. The APS is poorly framed as an argument *strictly* about data: "Properly understood, the [APS] is no more than an invitation to look upon language acquisition as a process of biological development" (Collins, 2008, 102). The underdetermination of external stimuli on the development of cognitive capacities is an assumption about biological organisms: "…it is not so much that the stimulus is not rich enough, as if it *could* be rich enough; rather, we simply understand that biological development is characterized by internal processes of generation and organization free of stimulus determination" (Collins, 2008, 103).

Once we arrive at questions of general or linguistic nativism, the reasoning used to justify the latter is much the same: any account of this phenomenon of language acquisition has to recognize that if disparate capacities of the mind "form a "constellation" in humans, dissociated from their other functions," this in no way has "direct bearing on the richness of universal grammar (UG)" (Chomsky, 2013, 34). Language is a phenomenon that is *distinct* from anything else in the animal world and any way to account for distinctly linguistic behavior that is reflexively acquired over predictable intervals of time should start from the premise that its structure is biologically imposed, *relatively irrespective of the quantity or quality of data*[5] *to which an individual happens to be exposed.*[6]

There are three points of relevance here for those interested in leveraging LLMs for the study of human language. First, Piantadosi references the "BabyLM" challenge that aims to develop language models trained on a "developmentally plausible amount of data" (Piantadosi, 2023, 14). On this interpretation of the APS, though, this perhaps misses the point: human language acquisition is not strictly dependent on just the right amount of data.

Second, even if one *were* to focus only on the data, the aim must be to pinpoint the *minimal* amount of data needed for a human to acquire a typical human linguistic competence[7] and *then* compare them to language models on comparably low amounts.

Finally, the APS is not strictly about data and the emphasis in this interpretation of the APS is on an *abstraction* of human language—the quantity or quality of data comes later.

[4] This niche source was found via Behme (2014, 23).

[5] To be clear, *some* data are necessary to act as a "trigger" for the acquisition of language—but this is not the same as saying the data is the *focus* of such acquisition.

[6] On this interpretation, Chomsky appears to be saying that the burden of proof is on those denying linguistic nativism given the preponderance of evidence that most biological systems impose their own peculiar pattern of development onto external stimuli.

[7] An ethical landmine, to be sure.

Linguistic Intuitions and Explanatory Adequacy

This type of abstraction is a methodological sticking point between some generative and computational linguists, exemplified by debates over the Galilean method, as we will see. It is useful to clarify, however, that when generative linguists claim that LLMs fail to offer *descriptive* and *explanatory* adequacy in theorizing about human language, they are making a claim about the *internal representations* that humans *do* and *could* possess in contrast to LLMs.[8]

Recall that generative linguists make two critical observations: first, an individual has the ability to construct an infinite number of sentences from finite (and often deficient) experience. Second, individuals *also* produce "intuitive judgments about the properties and relations of sound, form, and meaning of novel expressions in her language…" (Mikhail, 2011, 44). Both observations help comprise the foundations of theory-formation in generative linguistics. The latter linguistic intuitions, for their part, are seen as a form of "negative data"—not only data about what an individual believes *could* or *will* be said, but what could *never* be said while remaining grammatical (see, Allot, Lohndal, and Rey, 2021, 517, 521).

Because the focus of this research program is on the speaker's *knowledge of language* (Chomsky, 1986), a *descriptively* adequate grammar "correctly describes its object, namely the linguistic intuition—the tacit competence—of the native speaker (Chomsky, 1965, 27). But characterizing this "mass of feelings and understandings that we might call "intuitions about linguistic form" (Chomsky, 1975, 62) via descriptive adequacy is *not* an explanation.

Explanatory adequacy advances beyond these criteria, accounting not only for the *tacit competence* of native speakers, but also to offer "general principled reasons for choosing it over imaginable alternatives" (Rizzi, 2017, 100). Rizzi notes that this effort to not merely describe but explain *why* human language learners appear to select and converge on highly particular hierarchical structures—including cases of subject-auxiliary inversion, also reviewed by Piantadosi (2023, 21-23)—*despite* the availability of simpler linear orderings in the data. Universal Grammar is formulated in a somewhat "restrictive" fashion to account for *both* situations in which adult language learning could have many conceivable routes but dispenses with the readily apparent ones based on the available data *and* those in which there *is* variability across languages (Rizzi, 2017, 100-108).

LLMs possess too many dissimilarities in the way they use and understand language—more specifically, their ability to match the accuracy and reliability of grammaticality judgments—to reach observational and descriptive adequacy, despite surface-level appearances to the contrary.[9]

[8] Observational adequacy and descriptive adequacy are distinguished between the weak and strong generative capacity and in their reference to a corpus in the external world (E-languages) and an internal representational system (I-language), respectively (Rizzi, 2017, 98-100).

[9] There are (at least) three kinds of serious dissimilarities between LLMs and humans. First, they differ in how they *acquire* language (i.e., the *curation* of text-based linguistic data and the *amount* therein). Second, they differ in how they *understand* language when their grammaticality judgments are probed (and using the term "understand" is for ease of reading only). Third, they differ in how they *use* language—as we will see, they do not replicate CALU.

We thus cannot ask, building from their descriptions, why is *human* language one way, and *not* another way (Chomsky, 2022, 347-348)?

Still, there is cross-talk between Piantadosi and his critics, and the use of *abstraction* by the latter may be responsible for different ways of theorizing about language. Indeed, Piantadosi reaffirms an emphasis on inductive reasoning in response to his critics: while researchers "do not, at present, know how the models achieve this…(i) these models do it *somehow* and (ii) how they do it is almost certainly different from Chomsky's approach, and (iii) their approach works *really*, *really* well" (Piantadosi, 2023, 35).

The relationship between scientific theory and data—perhaps more appropriately, the relationship between scientific theory and reality—is at stake here. The Galilean method—in which the abstraction of human language is a critical methodological move—rests at the heart of this disjuncture, itself leading to the competence-performance distinction and an emphasis on "crucial" facts about human language.

Except for a brief allusion by Moro, Greco, and Cappa (2023, 83), neither Piantadosi nor his critics highlight the role of CALU in the characterization of human language. For generative linguists, this set of crucial data which yields deep insight into the nature of human language may often be implicit, focusing explicitly on the novelty that is subsumed under the "unboundedness" criterion (Collins, 2008, 153). For Piantadosi, CALU is either unrecognized or considered unimportant.

While the identification of CALU alone is a *theory-neutral* insight (McGilvray, 2001, 2), generativists' abstraction of the human language capacity and the subsequent distinction drawn between competence and performance gives them a distinctive advantage in this identification. The failure of LLMs to draw such a distinction by virtue of this abstraction, as well as the use of LLMs to draw insights about human language via inductive reasoning, leads to a deficient characterization of human language and subsequent deficiency in explanatory theory.

To see how this is the case, we turn to the Galilean method.

The Galilean Method

The "Galilean" method rests at the heart of this controversy. This method, or "style," refers to an interpretation of Galileo's conception of science. The term was coined by Edmund Husserl who, critical of this conception, associated it with the "idealizing and mathematizing" of natural phenomena, concerned not "with the free fall of *this* body" but rather the "indirect mathematization of the world…" (Husserl, 1970/1936, 41).

Chomsky's invocation of the Galilean method—with approval, contra Husserl—follows this emphasis on the *abstraction* of natural phenomena (without the concomitant emphasis on mathematics). Chomsky's critics, however, frequently confuse the Galilean method as a disposition towards *data*, when in reality it is an attitude about abstraction that sets the parameters for his selection and de-selection of relevant data (see, Collins, 2023, 3). Behme

(2014),[10] for example, takes Chomsky's invocation of the Galilean method to be part of a broader pattern in his work in which, when confronted with uncomfortable data that threatens his theories, he clings to the core tenets of Universal Grammar *in the face of mounting contradictory data*.

This confusion is critically relevant to the identification of CALU and its categorization into "performance" rather than "competence." It is, furthermore, just as relevant to understanding Chomsky's position on "simplicity" in explanatory theory (a notion brought up several times by Piantadosi (2023)). Chomsky's positions on each of these draws significantly from the Galilean disposition on abstraction and an accompanying understanding of the development of physics over time.

First, let us see exactly how Piantadosi—who is critical of the Galilean method (Piantadosi, 2023, 26-28)—understands the relationship between experimentation and principles in explanatory theories. His extended remarks are provided for the sake of clarity:

> Often, the only way to study such complex systems is through simulation. We often can't intuit the outcome of an underlying set of rules, but computational tools allow us to simulate and just *see* what happens. Critically, simulations *test the underlying assumptions and principles* in the model: if we simulate traders and don't see high-level statistics of the stock market, we are sure to have missed some key principles; if we model individual decision making for honeybees but don't see emergent hive decisions about where to forage or when to swarm, we are sure to have missed principles. We don't get a direct test of principles because the systems are too complex. We only get to principles by seeing if the simulations recapitulate the same high-level properties of the system we're interested in. And in fact the *surprisingness* of large language models' behavior illustrates how we don't have good intuitions about language learning systems (Piantadosi, 2023, 27) (emphases in original).

Systems like human language are so complex that the optimal way to test the principles embedded in them is by *simulating their use*. LLMs perform exactly this kind of simulation, finding a computational procedure based on next-token prediction that gives rise to emergent phenomena—among them, an apparent grasp of the structure of human language. From this, Piantadosi believes the Galilean method is refuted—the success of LLMs in reproducing human-like language use is powerful evidence against the idea that innate, high-level linguistic principles are necessary to give rise to the properties of language in which linguists are interested.

This conception of science and the testing of explanatory theories and principles within them is a powerful, even intuitive notion: theories rise and fall, it says, with the support—or lack thereof—of data. The more data that one can test against the performance of the target of explanation—in

[10] Behme incorrectly attributes the term "Galilean" style to Chomsky, claiming he "coined a term for this unorthodox methodology" (Behme, 2014, 687).

this case via the simulation of human languages in LLMs—the stronger the resultant theory. What justification could Chomsky have to dismiss this data-driven approach?

Abstraction and a Lowering of Scientific Expectations

Chomsky began discussing the ‘Galilean style’ in detail in his 1980 *Rules and Representations*.[11] There, Chomsky articulates both the *abstraction* of ordinary phenomena and a *lowering of expectations* for what a science of mind can realistically accomplish. More specifically, he expresses a certain admiration for the success thus far of the natural sciences while simultaneously *downgrading* expectations for the effectiveness of transferring methods of the natural sciences to the study of the human mind (Chomsky, 1980, 9-11).[12]

In these remarks, Chomsky alludes to the Cartesian understanding of CALU in an indication that this aspect of linguistic behavior is directly related to his conception of science. The Cartesian view that he summarizes is one in which the “contingencies that guide action, drives and instinct” and the like *can* be productively studied, but no assessment of these factors will breach the foreseeable inexplicability of the “freedom to choose” (Chomsky, 1980, 9). This remark implies a familiar position for Chomsky: that one can study the mechanisms that enable free human action, but stimulus-free, unbounded, yet appropriate linguistic behavior may exceed the scope of scientific inquiry (Chomsky, 1982, 2009b). Indeed, it is *after* these remarks—that lower one’s expectation for what a science of the human mind can accomplish—that Chomsky invokes “the Galilean style” in physics (Chomsky, 1980, 9).

In this passage, Chomsky cites theoretical physicist Steven Weinberg’s (1976) “The Forces of Nature” in which Weinberg explicitly invokes the Galilean style: “we have all been making abstract mathematical models of the universe to which at least the physicists give a higher degree of reality than they accord the ordinary world of sensation” (Weinberg, 1976, 28). We see an increasing emphasis on the abstraction of ordinary phenomena *away* from commonsense notions. Weinberg writes, “As our knowledge increases, the abstract mathematical world becomes farther removed from the world of sensation” (Weinberg, 1976, 28). A distance is placed between the ordinary human experience of natural phenomena and the abstractions embedded in explanatory theories of them.

Weinberg expresses personal awe, noting that “there is nothing so evocative of the Galilean style in physics as the idea of broken symmetry, the idea that on a true mathematical level there is a deep degree of symmetry between the forces of nature…it is remarkable that physics in this Galilean style should work” (Weinberg, 1976, 28). Chomsky, approving of these remarks, speculates that a physics “developed through inquiry in the Galilean style, is a remarkable historical accident resulting from chance convergence of biological properties of the human mind with some aspect of the real world” (Chomsky, 1980, 9).

[11] Earlier remarks can be found in Chomsky (1978).

[12] As Mikhail (2011, 26) observes, Chomsky’s ambition from his earliest works owes to an ability to ask questions that admit greater humility.

The success, thus far, of this Galilean style of physics is *lucky*: humans have a cognitive architecture that just so happens to approximate some—not all—aspects of reality. Transferring this style to the study of the mind, however, is no easy path to success: however effective this Galilean style is, there are aspects of reality that the human mind cannot penetrate due to the same biological constraints that enable its science-forming capacity—and what works for physics may work *less* effectively for the mind.

We see, then, a dual emphasis on both *abstraction* and a tempering of *scientific expectations* in the study of nature and mind in Chomsky's methodological remarks. The two go hand-in-hand yet are often delinked. Botha, in fact, argued that Chomsky's remarks in *Rules and Representations* adhere to a "typical stick-and-carrot pattern" (Botha, 1982, 1). He suggests that Chomsky presents the reader with "a mode of inquiry whose use may lead to the enviable kind of success achieved by the natural sciences" (Botha, 1982, 2) (i.e., the carrot) while simultaneously disparaging the idea that the Galilean method may be unsuitable for linguistics and proclaiming that "serious" linguists should accept its use (Botha, 1982, 1-4) (i.e., the stick).

A close reading of the remarks laid out above reveals the opposite: Chomsky is *almost* pessimistic about the possibility of achieving physics-like success in linguistics through the Galilean method. The idea is that it may be useful to *lower* one's expectations for a science of the mind *in order* to make greater progress.

Chomsky has repeatedly implored scholars to adopt a "willingness to be puzzled" about human phenomena, a willingness he argues enabled figures of the modern scientific revolution to move beyond commonsense beliefs about the natural world (e.g., Chomsky, 2013, 38).

It is from these positions on abstraction and scientific expectations that his attitude toward data is derived:

> From this point of view, substantial coverage of data is not a particularly significant result; it can be attained in many ways, and the result is not very informative as to the correctness of the principles employed. *It will be more significant if we show that certain fairly far-reaching principles interact to provide an explanation for crucial facts—the crucial nature of these facts deriving from their relation to proposed explanatory theories.* It is a mistake to argue, as many do, that by adopting this point of view one is disregarding data. Data that remain unexplained by some coherent theory will continue to be described in whatever descriptive scheme one chooses, *but will simply not be considered very important for the moment*" (Chomsky, 1980, 11-12) (emphases added).

For Chomsky, the accuracy of the principles used to explain a particular phenomenon is not best tested in reference to the *breadth* of data that a theory attempts to explain. Instead, what matters is accounting for data that represent "crucial facts" that link one's theory to deeper accounts of a phenomenon. This emphasis on crucial facts is absent, for example, in Piantadosi's (2023, 26-28) critique, *prematurely* leaping to the Galilean method's pursuit of underlying principles.

Georges Rey observes that this approach “could initially sound a little odd” (Rey, 2020, 16). The Galilean method does aim to produce “*structures and principles of reality that underlie the data*,” but, as Rey explains, “this might well not include *all* the data, some of which may be due to a complex interaction of causes, but just data that are *revelatory* of that reality” (Rey, 2020, 17). This harkens back to the methodological innovations made during the modern scientific revolution:

> Galileo and Newton did not do physics by taking a careful inventory of all the variety of motions objects exhibit, attempting to explain the complicated trajectories of people running up hills or leaves swirling in a storm. Rather, they turned to what they had a hunch were *specific data* that were simple and free from interaction effects…These often supplied *crucial data* that, relatively free of interfering factors, began to reveal the underlying principles of motion that were not predicted by rival theories, but which applied only in immensely complex ways to the trajectories of people, horses, clouds, and leaves in a storm (Rey, 2020, 17) (emphases in original).

This passage reveals three facts about the Galilean method: (1) The selection of data is made on the basis that very specific data—“crucial data”—offer deeper insight into the nature of phenomena than maximizing coverage of data; (2) To get at said crucial data, one must be willing to put aside a great number of other, less crucial data that distort the phenomena in question; and (3) The competition among scientific theories that results from this method is not about which theory has the most to say about the most data, but rather which theory can best offer an account of very specific (“crucial”) data (should a theory fail to address crucial data, it fails to offer explanatory depth about the phenomenon in question *no matter the breadth of coverage of other data*).

Weinberg, in his 1974 “Reflections of a Working Scientist,” explicitly argues that the notion of science as breadth of data coverage is misguided:

> One often reads in popular histories of science that “So and so’s data showed clearly that this and that were false, but no one at the time was willing to believe him.” Again, this impression that scientists wantonly reject uncomfortable data is based on a misapprehension as to the way scientific research is carried on. The fact is that a scientist in any active field of research is continually bombarded with new data, much of which eventually turns out to be either misleading or just plain wrong…When a new datum appears which contradicts our expectations, the likelihood of its being correct and relevant must be measured against the total mass of previously successful theory which might have to be abandoned if it were accepted (Weinberg, 1974, 40).

It makes sense, against this background, that Chomsky says such things as, “[Galileo] was willing to say “Look, if the data refute the theory, the data are probably wrong” (Chomsky, 2002, 98).

Competence, Performance, and Understanding Natural Phenomena

As Collins (2023) details, the competence-performance distinction is a *result* of these positions on abstraction, data selection, and theory-formation. The competence-performance distinction reflects Galileo's effort to 'decompose' natural phenomena to rid one's model of the enormous number of interactions that occur in the natural world. Galileo did not "[presuppose] the existence of a perfect material sphere that might be in contact with a surface at a single point." Instead, he "is insisting that science is only possible once we abstract to 'perfect' mathematical systems" (Collins, 2023, 5).

The Galilean method thus uses abstract models to present phenomena *as they would be* in the absence of external material data. When we observe human beings speaking, we see how language is *used* in real-world circumstances. But such use is mired in interactions with other phenomena, and proponents thus seek to understand language *as it would b*e in the absence of such performance effects. The relationship between data and theory follows what Allot, Lohndal, and Rey (2021, 520-521) characterize as consistent with the contemporary philosophy of science—not the broadest possible coverage of data that lead to generalizations and re-descriptions of phenomena, but deeper explanatory principles that depend on specific kinds of idealizations.

In the construction of abstractions, it is not unexpected that refinements will occur over time. Indeed, the role of *simplicity* in Chomsky's work is related to this effort (Collins, 2008, 77). The idea is that if explanatory principles are designed in proportion to the broadest coverage of the data, the principle is shallower; conversely, if they are designed in proportion to crucial data, they have greater depth. It is not surprising, then, that "[o]ver the years [Universal Grammar] has been pared down more and more" (Jackendoff, 2011, 588). This is to be expected in a Galilean style of inquiry; the explanatory theories thereafter adhere better to simplicity.

Empiricist-leaning approaches to language tend to confuse the concepts involved with the Galilean method or mischaracterize Chomsky's invocation of Galileo (or both). Chater et al. (2015), for example, argue that Chomsky misses two critical principles of Galileo's theorizing: "first, that we must look not to books but to Nature, the real phenomena, if we are to understand the world, and second, the language in which Nature is written is mathematical, which is to say, quantitative in character" (Chater et al., 2015, 96).

As we have seen, it is the emphasis on the abstraction of ordinary phenomena *away* from commonsense notions that underlie what Weinberg (1976) and Chomsky (1980) referred to as the "Galilean" style. Galileo looked to natural phenomena only as a first step before, as Collins (2023) shows, 'decomposing' them. The use of mathematics in this endeavor flows *from*, rather than towards, this use of abstraction.

Moreover, Galileo, consistent with Chomsky's remarks, *lowered* his expectations for what human explanations of natural phenomena can achieve in the same vein that he *abstracted away* from specific cases. Consider his comments on sunspots:

> For in our speculating we either seek to penetrate the true and internal essence of natural substances, or content ourselves with a knowledge of some of their properties. *The former I hold to be as impossible an undertaking with regard to the closest elemental*

> *substances as with more remote celestial things*…all the things among which men wander remain equally unknown, and we pass by things both near and far with very little or no real acquisition of knowledge…But this final information about water *is no more intimate than what I knew about clouds in the first place* (Galilei, 1957/1613, 123-124) (emphases added).

Galileo expresses that he is ignorant about the "true and internal essence" of all natural phenomena, lacking the 'intimacy' that would be required of this knowledge. He continues to explain that

> If what we wish to fix in our minds is the apprehension of some properties of things, then it seems to me that we need not despair of our ability to acquire this respecting distant bodies just as well as those close at hand…Hence I should infer that although it may be vain to seek to determine the true substance of the sunspots, still it does not follow that we cannot know some properties of them, such as their location, motion, shape, size, opacity, mutability, generation, and dissolution (Galilei, 1957/1613, 124).

As Garber notes, these remarks hint at a "Galilean science [that] places strict limits on the knowability of the nature of the bodies it studies" (Garber, 2004, 149).

Galileo's interest in "real phenomena" (Chater et al., 2015, 96), then, is better stated as an interest borne of humility: a stance in which we do not, and cannot, have 'intimate' access to natural phenomena. The use of mathematics in understanding them does not capture their essence, but rather provides a means by which the "properties of things" (Galilei, 1957/1613, 124) may be grasped.

These can be radical remarks to some but note how they reveal the deficiencies of the conception of scientific inquiry espoused by Piantadosi: LLMs represent the broadest possible coverage of linguistic data via the simulation of human-like language use to-date. Yet, importantly, this amounts to a sort of *re-description* of linguistic data—and not, as we have seen, a wholly accurate one beneath the surface.

This owes to the reality that LLMs offer no identification of "crucial facts" about human language. This, in turn, owes to the fact that LLMs provide *no abstraction* of the phenomenon of human language, instead training on vast amounts of text-based linguistic data—that is, training and then, at best, describing linguistic *performance*. This is radically different than offering an account of human linguistic competence, inclusive of crucial facts pertaining not only to the APS and the biological presuppositions about language acquisition, but also to competence-borne grammaticality judgments (see, Allot, Lohndal, and Rey, 2021, 524).

All data in the LLM's training process are leveraged for the singular objective of next-token prediction. Whatever representations of the data the LLM derives from this will not be a reliable indicator of the crucial facts for *human* language. This point is vital, if counter-intuitive: a simulation of a natural phenomenon can only address the crucial facts about that phenomenon if it is *precisely constrained to the theoretical priors* that researchers have already employed in its study. LLMs are simulations of human language, but—if they are taken as theories or strongly theory-informative—the representations of language they construct during training will be relevant only to those theories that consider next-token (or next-word) prediction to be crucially relevant.

It is not optimal—*for purposes of explanation*—to *reproduce* a phenomenon via simulation over the broadest coverage of data for that phenomenon (imperfectly, in LLMs' case) without paying attention to abstraction and the crucial facts of the phenomenon. This presupposes that the end-result—the "high-level" properties that Piantadosi (2023, 27) claims LLMs "recapitulate"—are already filtered through an adequate theoretical framework that takes both abstraction and crucial data seriously. Piantadosi's argument turns too heavily, in this way, on the idea that language's crucial properties are perhaps obvious, and we just have to find a way to model the phenomenon in a manner that already makes sense.[13]

Let us turn directly to CALU and LLMs and flesh out this reasoning.

The Creative Aspect of Language Use and Large Language Models

CALU represents crucial data of its own. The stimulus-freedom, unboundedness, and appropriateness of linguistic productions are, to use Rey's (2020, 17) term, *revelatory* of the nature of human language. Indeed, Chomsky's "creative aspect observations, along with the poverty of stimulus observations, offer a set of facts with which his and—he holds—any science of language must contend" (McGilvray, 2001, 5).

If LLMs are to be considered theories of human language, then they should, at least, identify CALU or shed light on this phenomenon. Let us illustrate, first, how CALU manifests in humans.

Human Language Use:

Stimulus-Freedom: Humans use their language in a manner that is not determined by their immediate environments, or even their internal physiological states. Any attempt to provide a serious causal explanation for an individual's linguistic production will merely be an *interpretation*—and not a scientific one—of many factors about the given context—"[t]his is not the well-defined causality of serious theory…" (McGilvray, 2001, 7; see also, Asoulin, 2013, 230). Individuals may, in the face of a stimulus (e.g., being asked a question) say nothing at all, further illustrating their linguistic behavior's detachment from local contexts. Human language use is stimulus-*free*.

Unboundedness: Humans use their language in a way that is novel or innovative. That is, they construct new words, phrases, and sentences without any apparent upper limit on this ability.

[13] Although Piantadosi emphasizes that LLMs' representations "are parameterized in a way which is unfamiliar to linguistics" (2023, 7)—and even criticizes some generative linguists for allegedly believing that scientific theory "must be *intuitively comprehensible* to us" (Piantadosi, 2023, 35)—he is perhaps underestimating the intuitiveness of his own approach that leverage computational methods in this way—at least, so far as he understands the nature of explanation.

Individuals are not limited to a pre-sorted list of words, phrases, and sentences in the production of a thought or an utterance.

Appropriateness and Coherence to Circumstances: Finally, although an individual's use of language is both stimulus-free and unbounded, it is also appropriate to the circumstances of its use and coherent to others who hear one's remarks. More specifically, their remarks are judged to be relevant by others in the situation who may have thought or uttered similar remarks. Language use "is recognized as appropriate by other participants in the discourse situation who might have reacted in similar ways and whose thoughts, evoked by this discourse, correspond to those of the speaker" (Chomsky, 1988, 5).

Taken together, these criteria constitute CALU (Chomsky, 1968, 10-11). This entirely *ordinary* use of language "is not a series of random utterances but fits the situation that evokes it but does not cause it, a crucial if obscure difference" (Chomsky, 1988, 5). This ability to use one's cognitive capacities in a manner that is free of local circumstances, robustly innovative across contexts, yet appropriate and suitable for any given context is a "species-specific capacity, a unique type of intellectual organization which cannot be attributed to peripheral organs or related to general intelligence" (Chomsky, 2009a, 60). Note that, for human language use to be "creative," in this sense, all three factors must be present *simultaneously* (Baker, 2008, 237).

This capacity "for the free expression of thought or for appropriate response in any new context" (Chomsky, 2009a, 60) represents a set of crucial facts about human language. The uniquely human ability to deploy one's cognitive resources at will, most expressible through language, to any problem to which they see fit and have their productions make sense to the thoughts of others is central to a depiction of the human species.

Any theory of the human capacity for language, then, must take CALU seriously. Creative language *use*, of course, is intermingled with an enormous variety of interaction effects in the real world, and thus—consistent with the Galilean style—the target of explanation is not any *specific instance* of creative language use but an abstraction of its underlying mechanisms. In this way, when generative linguists attempt to characterize competence—I-language—they are *not* explaining creative language use, but merely the competence that enables it.

LLMs' Language Use

LLMs do not exhibit a creative use of language. Let us see why.

Stimulus-Controlled: An identifiable stimulus constrains an LLM's use of language to the *context* of its use. Usually, an external stimulus takes the form of *prompting* an LLM. The relevant prompt can be input by either a human end-user or a separate language model with which it is interacting (though, most often the former). In either case, the input value controls the output value—not only the *content* of the output value, that is, but the very *fact* that the LLM will respond.[14] Given a prompt or an input, the LLM exhibits no independent ability to respond

[14] More accurately, LLM-powered chatbots do not "respond" but instead continue the input value by predicting its most likely continuation.

or not respond. Nor does the LLM exhibit an ability to detach itself from the local contexts of its use and generate linguistic productions that correspond with the thoughts of a human end-user.

While an LLM could be manipulated by human programmers in such a way as to make its stimulus reactions variable (e.g., to not respond to every third input), the LLM is ultimately bound to either the internal state imposed by human programmers or, in normal conditions, the prompt given to the model (or a combination of both).

This is not how human beings use their language in response to stimuli. Chomsky will, drawing from the Cartesian tradition, speak of language use that may be "predictable" but only in the sense "that they will tend to do what they are incited and inclined to do, but they are nonetheless free, and uniquely so, in that they need not do what they are incited and inclined to do" (Chomsky, 1988, 6). Chomsky thus distinguishes between matters of *etiquette or coercion* and *choice*. "If, for example, I were to take out a machine gun, point it menacingly at you, and command you to shout "Heil Hitler," you might do it if you had reason to believe I was a homicidal maniac…" (Chomsky, 1988, 6).

Is this the same as stimulus-control? Not quite: "[Y]ou would have a choice in the matter, even if that choice is not exercised. The situation is not unknown in the real world; under Nazi occupation, for example, many people…became active or passive collaborators, but some resisted" (Chomsky, 1988, 6).

Chomsky's example is extreme but illustrative: human beings use their language in a manner that is detached from *local* circumstances, and they do so at will. It is thus stimulus-*free* whereas an LLM is stimulus-*controlled*.

Weak Unboundedness: There are two ways to interpret unboundedness in CALU, one "weak" and one "strong."[15]

The former refers to the production of an undefined number and kind of linguistic thoughts or utterances—this is typically the version entertained in the literature, particularly in Descartes' (1910/1637, 60-61) remarks. The latter goes a bit further, linking unlimited linguistic productions with semantic content or formal linguistic competence—an *understanding* of one's novel linguistic productions.[16]

I adopt the "weak" version of unboundedness here. Expanding the definition of unboundedness to cover semantics in a test of CALU bleeds into separate matters that concern both formal competence—which CALU does not test for—and the "appropriateness" criterion. We should be clear, furthermore, that the test of CALU is a test of a "mind like ours" (Chomsky, 1988, 5), not a test for the existence of generative grammar. These are overlapping but distinct concepts.

LLMs like GPT-3 or GPT-4 appear to achieve the "weak" version of unboundedness with ease, generating limitless numbers and kinds of sentences. The reality, though, is a bit more tempered

[15] This distinction is my own, hopefully useful insertion.

[16] For extended arguments on CALU's relationship to lexical semantics and the science of language see McGilvray (2001, 2005) and Asoulin (2013).

than it appears. McCoy et al. (2023) do find evidence that when decoupling syntax from semantics (i.e., decoupling well-formedness from meaning and coherence)—consistent with our "weak" interpretation—GPT-2 produces *varying* degrees of novel content that does not appear in its training dataset.[17] In compositional generalization, models including GPT-2 demonstrate an ability to 'combine familiar parts in novel ways,' though perhaps "limited to particular subcases" owing to scores that "are lower than the baseline" (McCoy et al., 2023, 664).[18]

Dentella, Murphy, Marcus, and Leviada (2023) find that testing LLMs on "less frequent" and exotic linguistic constructions reveals an inconsistency and unreliability in their grammaticality judgments. The sentence—which can trip up humans at first—"*More people have been to Russia than I have*" is judged to be grammatically correct and meaningful by GPT-3, eventually contradicting itself on the meaning of the sentence when "more" is replaced with "fewer" (Dentella, Murphy, Marcus, and Leivada, 2023, 4-5). Humans tend to correct their mistakes in judging this sentence's grammaticality upon reflection; LLMs are not so reliable.

These results decidedly cast doubt on the "strong" version of unboundedness.[19] However, they do not refute state-of-the-art LLMs' ability to generate words and sentences that are not strictly pre-determined. Whether the ability of an LLM like GPT-3 to generate novel linguistic productions is unlimited—not restricted to phenomena found in its training dataset—is a worthwhile question, but we take a charitable interpretation here given the difficulty of assessing this problem.

LLMs thus use language in a "weak" unbounded fashion.

Appropriateness and Coherence to Situations: The appropriateness of language use is the most difficult to evaluate, as this determination typically depends on the intuitive judgments of humans. Chomsky notes an association of thought and language here, remarking that the "normal use of language…is recognized as appropriate by other participants in the discourse situation who might have reacted in similar ways and whose thoughts, evoked by this discourse, correspond to those of the speaker" (Chomsky, 1988, 5). Does this mean that anyone who, for example, enjoys interacting with an LLM recreationally can claim the model achieves appropriateness by virtue of their own experience? Not exactly.

Let us identify what does *not* count as appropriateness. McGilvray marks out a distinction here: creative linguistic productions are stimulus-free, so "being appropriate" is not "being caused by environmental conditions." Nor is it being regularly correlated with the environment" (McGilvray, 2001, 9). He explains that this "is not to say that regular correlation, with or without causation, yields no conception of appropriateness at all" (McGilvray, 2001, 9). Creative appropriateness is not "functional," goal-oriented appropriateness and one cannot capture this criterion merely through "articulate scheme(s) or formula(e)" (McGilvray, 2001, 9). One need

[17] Note that this is *not* a test of the models' grammaticality judgments.

[18] McCoy et al. (2023, 652-653) note that while language models sometimes copy text from their training data, this is not as frequent as expected (though, one cannot simply *assume* the novelty of their outputs).

[19] One could plausibly argue that a machine could not possess a "mind like ours" *without* a productive and consistent linkage of syntax and semantics. I leave this to future research but note that *testing* for a mind like ours via CALU is not quite the same as *characterizing* such a mind.

not satisfy a standard or speak in reference to an established goal in one's environment for their remarks to be appropriate.

LLMs clearly violate this condition. Base language models, disconnected from a more complex conversational system,[20] generate linguistic productions that *are* regularly correlated with and, in fact, *caused* by their local environment—the inputs they receive. This is not appropriate in the creative sense not because their use of language is *sometimes* correlated with the environment, but because it *always* is—whether the initial stimulus was given in the form of a prompt and then left the model to its own devices, so to speak, or is engaged in a continuous input-output exchange with a human or separate agent, its use of language is oriented towards an established environmental stimulus. The "environment" that an end-user or programmer establishes for the LLM is the *only one* in which the LLM uses its language, even if separated by time.

Contrary to what one might expect, this situation is only worsened in the case of fine-tuning base language models via Reinforcement Learning from Human Feedback (RLHF). This technique was notably used before the release of ChatGPT in the design of InstructGPT to aid in the reduction of falsehoods and toxicity (aligning an LLM-powered chatbot with the end-user's intents) (Ouyang et al., 2022). We see, here, that the LLM's use of language is becoming *more* restricted to a given "environment"—the contextual needs of a specific class of human end-users.

RLHF is a form of fine-tuning that takes *human* preferences as input. It involves humans labeling data according to their judgments on the relevant content, training a model on these data via supervised learning, training a separate reward model to predict appropriate answers among different texts based on human-written rankings, and, finally, fine-tuning the base model to maximize the reward—the desired *human* outputs (Ouyang, 2022, 2).

The *conversational* capabilities of ChatGPT are the result of this technique. As Kocoń et al. note, "One of the latest iterations of InstructGPT is the ChatGPT model, which most likely exploited even more users' feedback on a greater variety of tasks" (Kocoń, 2023, 2). (The ChatGPT model at the time was ChatGPT-3.5.)

Any appropriateness learned by a system like ChatGPT, then, is inextricably tied to the judgments of others on the relevant content—that is, ChatGPT's "appropriateness" amounts to fitting a curve of data produced by humans *who already possess an ability to intuitively judge the appropriateness of linguistic productions*. ChatGPT, by virtue of its fine-tuning via RLHF, is more *tool-like* in its use of language than it might otherwise have been. This is distinct from the creative sense of appropriateness which, through stimulus-free and unbounded expressions, nonetheless corresponds with the thoughts of others. LLMs thus do not use their language appropriately in the creative sense.

Having achieved just one of the three criteria—weak unboundedness—LLMs do not exhibit CALU and thus cannot be said to reproduce or simulate the ordinary human use of language.

[20] As Shanahan explains, to create a *conversational* AI system, "the LLM will have to be embedded in a larger system to manage the turn-taking in the dialogue. But it will also need to be coaxed into producing conversation-like behaviour" (Shanahan, 2022, 4). The "base" language model—say, GPT-4—is *not* the entire conversational system.

A Galilean Lens for LLMs and CALU

Broad Coverage of Linguistic Data Tells Us Nothing About CALU

Neither base LLMs nor LLM-powered chatbots exhibit CALU. However, as our exploration of the Galilean method has shown, this shortcoming is only a major failure if one considers an explanatory theory of language to be one in which the "high-level properties" (Piantadosi, 2023, 27) of human language are predicted by one's model.

LLMs neglect crucial data relevant to human language in favor of the broad coverage of linguistic data. As a result, they are unable to provide an adequate description of target grammars and they fail entirely to account for—meaning to reproduce, identify, make sense of, or explain—human language's creative uses. The *ineffectiveness* of this statistical approach to explanatory theory reveals a *strength* of the Galilean method: the selection of some particularly crucial data whose relationship to theory yields deeper insight into the object of study. In this case, the careful identification of ordinary human language use's stimulus-free, unbounded, and appropriate dimensions and placing CALU in a relationship with the study of linguistic competence.

Piantadosi could object on the basis that LLMs are not *designed* to simulate CALU. As such, there is no reason to expect they will acquire this ability. However, this argument contradicts the basis of his central claim: that LLMs like ChatGPT, despite not being explicitly designed to acquire the abstract structure of human language according to the properties with which linguists are interested (instead designed to predict the next token and give the appearance of human-like language use), nonetheless accomplish this feat—with few, if any, characteristically Chomskyan methodological choices to boot.

So it is either that LLMs do not model human language—thereby undermining the critique they enable of Chomsky's approach to linguistics—or they do yet fail to provide an account of the crucial data that are bound up in CALU. This latter possibility would be a significant failure (again, only if one takes LLMs as theories or strongly theory-informative).

Piantadosi could counter this by emphasizing that LLMs are *simulations*—simulations only model text-based linguistic data, rather than the full cognitive architecture of the human mind, and thus should not be expected to take on the full character of *human* language use. This is perfectly true, which is why Allot, Lohndal, and Rey (2021, 524) emphasize that deep neural networks model only human linguistic *performance*—the behavior enabled by human linguistic competence.

This, then, merely highlights how LLMs are not capturing some of the most relevant dimensions of human language. This shortcoming is a result of not producing an abstraction of the human language capacity and therefore not invoking a competence-performance distinction. Because of these methodological shortcomings, they also fail to identify crucial facts—among them, that human linguistic performance has not been explained in any scientifically meaningful way through prediction, instead exhibiting a "creative" character.

LLM-based Theories Aim Too High

A further problem with LLMs-as-theories is that such theories are far too *ambitious*. Recall the emphasis in Weinberg (1976) on the surprising effectiveness of the Galilean style in physics thus far and Chomsky's (1980) warning that it is unlikely to be as effective in studying the mind as physical reality. The idea running throughout our overview of the Galilean method is that, in seeking an explanation for crucial data, one aims for *simplicity* of explanation relative to the scope of the inquiry, while remaining aware that one's explanation is premised upon an *abstraction* of the phenomenon.

This is nowhere more evident in generative linguistics than in discussions of CALU. Generativists hedge their bets on the explainability of CALU, arguing that the simultaneous presence of stimulus-freedom, unboundedness, and appropriateness is likely not amenable to scientific explanation.

Their reasoning goes like this: the unboundedness of language use can be productively studied and given a reasonable scientific account. The *novelty* of language use, indeed, falls directly under the purview of generative grammar (Collins, 2008, 153). Stimulus-freedom, furthermore, is open to less precision, but can at least be accounted for in reasonable scientific terms by appealing to the relative autonomy of cognitive systems (operating as sub-systems of the mind/brain) and the "flexible interrelationships" between them (McGilvray, 2005, 221-222).

It is the *appropriateness* of unbounded and stimulus-free language use that throws a curveball. It is simply not clear why there is an alignment of sorts between human beings' use of language. Consider that human language use is neither *determined* by context nor is it *random* or in some sense *probabilistic*—if it were determined by context, it would not be stimulus-free; if it were random or probabilistic, it may be stimulus-free and unbounded yet frequently inappropriate and incoherent.

Taken together, these three characteristics of language use do not readily admit explanation. Indeed, Chomsky follows Descartes in accepting not only that CALU "falls entirely beyond the conceptions of mechanics" (Chomsky, 1988, 139) but also that "the Cartesian questions of creative use…remains as much of a mystery now as it did centuries ago, and may turn out to be one of those ultimate secrets that ever will remain in obscurity, impenetrable to human intelligence" (Chomsky, 2009b, 200).[21] Science does not deal with phenomena exceeding the bounds of determinism or randomness. CALU, which does this, is unlikely to be explained.

Note that this does not mean that we do not have *rational* explanations for appropriateness. We routinely find *commonsense* explanations for creative language use. Consider a hypothetical example: you show up early to a Christmas party at a friend's house and deliver your gifts before the host arrives. The host, who came home later and is now unpacking some groceries, may casually observe how "Even Jim brought over some gifts earlier." The remark, which was made

[21] Chomsky (2006, 86) explicitly rejects Descartes' notion of a "second substance" in seeking an explanation for creative language use in favor of viewing the mind as a biological endowment of the human species (McGilvray, 2005, 222).

in the kitchen and not in response to any gift-specific stimulus, leads you to believe that the host thinks you did not bring over gifts, unlike Jim. The host's remark is appropriate, in this sense, produced without the stimulus of a packaged gift in the area and arranged in just one possible way.[22]

Yet, as McGilvray (2001, 13-18) aptly explains, this rational account could never serve as the basis for a *scientific* explanation of appropriateness. Trying to scientifically explain appropriateness via reasons of everyday life is akin to trying to scientifically explain chemical interactions involving water by dropping the pretension of "H_2O" and using, instead, our commonsense notion of "water." The latter move would be impermissible in modern chemistry; there is no reason to adopt it in a science of language use.

Recall that Piantadosi believes the Galilean method was refuted on the basis that simulations of language allow scholars to "*test the underlying assumptions and principles* in the model," in this way suggesting that "[w]e only get to principles by seeing if the simulations recapitulate the same high-level properties of the system we're interested in" (Piantadosi, 2023, 27). If one adopts this view or otherwise uses LLMs as direct insights into human language, then we must observe that such stances are too ambitious—they aim too high, trying to provide the broadest coverage of linguistic data while not distinguishing between competence and performance, incorrectly leading scholars to assume that ordinary examples of human language use can be explained by 'recapitulating' the latter.

A parallel problem alluded to above is that Piantadosi's (2023, 27) notion of "emergence" in LLMs—the reproduction of human linguistic qualities from low-level principles adopted by the model—violates Weinberg's (1976) notion that scientists move *away* from commonsense ideas about the phenomena they study. The reproduction of human linguistic qualities by LLMs simply re-asserts the picture of the phenomenon we started with: human linguistic performance. We are interested, however, in *explaining* the picture we started with, and the scientific explanation provided will take the form of an abstraction that moves *away* from the "ordinary world of sensation" (Weinberg, 1976, 28).[23] The appropriateness of stimulus-free and unbounded human language use—if it is to be explained—cannot have the "rational," everyday account that we provide for it, just as chemists think of water not as the liquid one drinks but as "H_2O."

Putting aside matters of competence and grammaticality judgments, note that even if statistical analysis of human language use led to reasonably accurate predictions about the next word an individual will actually speak (and whether they will speak at all), while this may be of some interest to a theory of performance, but it would plainly offer no explanation for *how* the individual accomplishes this feat of ordinary language. To do so, more than prediction would be needed, namely an account that offers some insight into a phenomenon that is neither determined nor probabilistic or random. Statistical methods do not provide this.

[22] Stimulus-freedom can go much further, speaking of not just circumstances removed from the nearby environment (e.g., the living room where the gifts are located) but entirely *fictional* circumstances.

[23] As Alexandre Koyre puts it, the "principle of inertial motion appears to us perfectly clear, plausible, and even, practically, selfevident…Yet it is nothing less than that…to the Greeks, as well as to the Middle Ages, they would appear as "evidently" false, and even absurd" (Koyre, 1943, 335).

The explanatory question for CALU itself (however ill-advised this is at our current level of understanding) would not be, “What is the next word?” but rather, “Why is the individual using this sequence of words in this context and *not* another?”

LLMs as theories thus aim too high. The risk to linguistics echoes the quote that begins this chapter: that scientists will once again believe they have found the key to the stimulus-free, unbounded, yet appropriate use of language by humans, only to realize at some later point that expectations were too high and the fact that humans possess this ability is an “insight that must be recaptured time and time again” (Chomsky, 1988, 88).

Combinatorial Cognition and the Role of Language in Thought

Finally, there is a sense in which the debate sparked by Piantadosi misses a larger point about studying language as one part of the more complex human mind/brain. This point concerns the more speculative relationship between *language* and *thought*.

According to Elizabeth Spelke (2010), a distinctive element of human thought is the ability to engage in combinatorial cognition; that is, while human beings possess “core knowledge systems” like non-human animals, each providing their own concepts, “the concepts human construct with our combinatorial system are more numerous and flexible.” Unlike non-human animals, “humans therefore can act by choosing among the options made available by the combinatorics of natural language.” Language, in this way, “may serve as a medium for constructing new concepts once words and expressions are linked to representations from multiple core systems” (Spelke, 2010, 208).

Let us parse out the nuances here. Human beings possess an *ability* to construct new concepts from their core knowledge systems through natural language. But this ability is formulated in the abstract and is not, as Spelke (2010, 208-209) acknowledges, an explanation for the specific *choices* or instances in which humans actually do this. These choices are precisely what CALU describes: the stimulus-free, unbounded, yet appropriate use of language.[24] While entirely ordinary, it illustrates the distinctive *way* in which human beings go about constructing new concepts and ideas.

Consider what this entails: human beings have an *ability* to construct new concepts through natural language’s combinatorics. If this ability were stimulus-controlled, however (to take one counterfactual), the intellectual contours of the species would be sharply limited. This is because the ability would be, in this instance, bound to local conditions.

But human language is not like this. Intellectual ruminations that can be constructed and/or expressed with language are stimulus-free, unbounded, yet appropriate to a particular subject, a characteristic of humans that is lost in depictions of LLMs as theories of human language. Language serving as a sub-system of the more complex mind/brain, used intentionally by humans to link representations from disparate domains together, means not that *all* thought is

[24] To be clear, CALU describes ordinary language use that is *either* constructing or not constructing new concepts.

linguistic, but that language and its creative uses are central to the intellectual lives of the species.

As noted, this is a somewhat speculative line of thought. It is, however, important to emphasize that it is unwise to—as Elliot Murphy (2023) explains in recounting his debate with Piantadosi—downplay either the modular architecture of the mind posited by generativists or to delink syntax from semantics based on a misinterpretation of the "autonomy of syntax."

Moreover, those who argue for the *dissociation* of language and thought in the study of humans' and LLMs' formal and functional competence, as Mahowald et al. (2023) do prominently, must—in addition to having a sound depiction of the former (see, Leivada, Dentella, and Murphy, 2023)— grapple first with the fact that human thought is not only uniquely combinatorial but uniquely executed. Only then can a comparison to LLMs start on the proper footing.

Conclusion

Perhaps the greatest misconception about Noam Chomsky is that his linguistic work is emblematic of a new scientific method, abruptly departing from the scientific status quo in favor of his own peculiar conception of Universal Grammar. It is more accurate to suggest, as John Collins did, that "the greatest service Chomsky has provided for philosophy is to do philosophy of science via the construction of a new science. In this light, the best way to proceed is to present the science and pay attention to the philosophy as we go" (Collins, 2008, 25).

The greatest value of Piantadosi's critique of Chomsky's approach to linguistics, in my view, is that it forces scholars working within the generative tradition to return to the subfield's foundations. Upon returning, we reassess the strengths and vulnerabilities of methodological moves that previously may have been taken for granted before the very public rise of LLMs and linguistic work premised on their successes. Piantadosi's piece compels us, in this way, to reformulate what it means to engage in a science of the human mind.

What this chapter finds in this reassessment is that crucial facts about human language go beyond the APS. Namely, the stimulus-free, unbounded, and appropriate and coherent use of language falls outside the explanatory scope of theories premised upon or within LLMs. Concomitantly, it finds that the Galilean method that Chomsky's critics deride has distinctive advantages in the identification of CALU in humans, its role in a theory of language, and how this relates to the simulation of human language via the broad coverage of data. Thus, Chomsky's (1988, 88) insight that CALU is an old idea that must be remembered time and again has enduring wisdom in the age of LLMs.

Bibliography

Allott, N., Lohndal, T., & Rey, G. (2021). "Chomsky's "Galilean" Explanatory Style." In: N. Allot, T. Lohndal, & G. Rey (Eds.), *A Companion to Chomsky* (pp. 515-528). Hoboken, NJ: John Wiley & Sons, Inc.

Asoulin, E. (2013). "The Creative Aspect of Language Use and the Implications for Linguistic Science." *Biolinguistics*, *7*, 228-248. https://doi.org/10.5964/bioling.8963.

Baker, M. C. (2008). "The Creative Aspect of Language Use and Nonbiological Nativism." In: P. Carruthers, S. Laurence, and S. Stich (Eds.), *The Innate Mind, Volume 3: Foundations and the Future* (pp. 233-253). New York, NY: Oxford University Press.

Baroni, M. (2022). "On the Proper Role of Linguistically Oriented Deep Net Analysis in Linguistic Theorizing." In: Shalon Lappin and Jean-Philippe Bernardy (Eds.), *Algebraic Structures in Natural Language* (pp. 1-16). Boca Raton, FL: CRC Press.

Beguš, G., Dąbkowski, M. and Rhodes, R. (2023). Large Linguisitc Models: Analyzing Theoretical Linguistic Abilities of LLMs. *ArXiv*. 1-28. https://arxiv.org/abs/2305.00948.

Behme, C. (2014). "A 'Galilean' Science of Language." *Journal of Linguistics*, *50*(3), 671-704. https://doi.org/10.1017/S0022226714000061.

Botha, R. P. (1982). "On the Galilean Style of Linguistic Inquiry." *Lingua*, *58*(1-2), 1-50. https://doi.org/10.1016/0024-3841(82)90056-0.

Chater, N., Clark, A., Goldsmith, J., Perfors, A. (2015). *Empiricism and Language Learnability*. New York, NY: Oxford University Press.

Chomsky, N. (1964). *Current Issues in Linguistic Theory*. The Hague: Mouton & Co.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. 50th Anniversary Edition. The MIT Press.

Chomsky, N. (1968). *Language and Mind*. San Diego, CA: Harcourt Brace.

Chomsky, N. (1975). *The Logical Structure of Linguistic Theory*. Chicago: The University of Chicago Press.

Chomsky, N. (1978). "A Theory of Core Grammar." *Glot*, *1*(1), 7-26.

Chomsky, N. (1980). *Rules and Representations*. New York, NY: Columbia University Press.

Chomsky, N. (1982). "A Note on the Creative Aspect of Language Use." *The Philosophical Review*, *91*(3): 423-434. https://doi.org/10.2307/2184692.

Chomsky, N. (1986). *Knowledge of Language: Its Nature, Origin, and Use*. Westport, CT: Praeger Publishers.

Chomsky, N. (1988). *Language and Problems of Knowledge: The Managua Lectures*, Third Printing. Cambridge, MA: The MIT Press.

Chomsky, N. (2002). *On Nature and Language*. Cambridge, UK: Cambridge University Press.

Chomsky, N. (2006). *Language and Mind*, Third Edition. New York, NY: Cambridge University Press.

Chomsky, N. (2009a). *Cartesian Linguistics: A Chapter in the History of Rationalist Thought*, Third Edition. Cambridge University Press.

Chomsky, N. (2009b). "The Mysteries of Nature: How Deeply Hidden?" *The Journal of Philosophy*, *106*(4), 167-200. https://doi.org/10.7312/bric14474-002

Chomsky, N. (2013). Problems of Projection. *Lingua*, *130*, 33-49. https://doi.org/10.1016/j.lingua.2012.12.003.

Chomsky, N. (2022). "Genuine Explanation and the Strong Minimalist Thesis." *Cognitive Semantics 8*, 347-365. https://doi.org/10.1163/23526416-bja10040.

Chomsky, N. and Moro, A. (2022). *The Secrets of Words*. Cambridge, MA: The MIT Press.

Chomsky, N., Roberts, I., and Watumull, J. (2023). "Noam Chomsky: The False Promise of ChatGPT." March 8, 2023. Accessed December 27, 2023. https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html.

Collins, J. (2008). *Chomsky: A Guide for the Perplexed*. London, UK: Continuum International Publishing Group.

Collins, J. (2023). "Generative Linguistics: 'Galilean Style.'" *Language Sciences*, *100*, 1-14. https://doi.org/10.1016/j.langsci.2023.101585.

Descartes, R. (1910/1637). *Discourse on the Method of Rightly Conducting the Reason, and Seeking Truth in the Sciences*. Chicago, IL: The Open Court Publishing Company.

Dentella, V., Murphy, E., Marcus, G., and Leivada, E. (2023). "Testing AI Performance on Less Frequent Aspects of Language Reveals Insensitivity to Underlying Meaning." *ArXiv*, 1-24. https://arxiv.org/abs/2302.12313.

Dentella, V., Günther, F., and Leivada, E. (2023). "Systematic Testing of Three Language Models Reveals Low Language Accuracy, Absence of Response Stability, and a Yes-Response Bias." *PNAS*, *120*(51), 1-10. https://doi.org/10.1073/pnas.2309583120.

Galilei, G. (1957/1613). *Discoveries and Opinions of Galileo: Translated with an Introduction and Notes by Stillman Drake*. New York, NY: Anchor Books.

Garber, D. (2004). "On the Frontlines of the Scientific Revolution: How Mersenne Learned to Love Galileo." *Perspectives on Science*, *12*(2): 135-163. https://doi.org/10.1162/106361404323119853.

Horgan, J. (2016). "Is Chomsky's Theory of Language Wrong? Pinker Weighs in on Debate." *Scientific American*. November 28, 2016. Accessed December 27, 2023. https://blogs.scientificamerican.com/cross-check/is-chomskys-theory-of-language-wrong-pinker-weighs-in-on-debate/.

Husserl, E. (1970/1936). *The Crisis of European Sciences and Transcendental Phenomenology: An Introduction to Phenomenological Philosophy*. Translated by David Carr. Evanston, Illinois: Northwestern University Press.

Jackendoff, R. (2011). "What Is the Human Language Faculty? Two Views." *Language*, *87*(3): 586-624. http://dx.doi.org/10.1353/lan.2011.0063.

Kallens, P. C., Kristensen-McLachlan, R. D., and Christiansen, M. H. (2023). "Large Language Models Demonstrate the Potential of Statistical Learning in Language." *Cognitive Science: A Multidisciplinary Journal*, *47*(3): 1-6. https://doi.org/10.1111/cogs.13256.

Katz, Y. (2012). "Noam Chomsky on Where Artificial Intelligence Went Wrong." *The Atlantic*. November 1, 2012. Accessed December 28, 2023. https://www.theatlantic.com/technology/archive/2012/11/noam-chomsky-on-where-artificial-intelligence-went-wrong/261637/.

Katzir, R. (2023). "Why Large Language Models Are Poor Theories of Human Linguistic Cognition: A Reply to Piantadosi." *Biolinguistics*, 17, 1-12. https://doi.org/10.5964/bioling.13153.

Kocoń, J., et al. (2023). "ChatGPT: Jack of All Trades, Master of None." *Information Fusion*, *99*, 1-37. https://doi.org/10.1016/j.inffus.2023.101861.

Kodner, J., Payne, S., & Heinz, J. (2023). "Why Linguistics Will Thrive in the 21st Century: A Reply to Piantadosi." (2023). *ArXiv*. 1-28. https://doi.org/10.48550/arXiv.2308.03228

Koyre, A. (1943). "Galileo and the Scientific Revolution of the Seventeenth Century." *The Philosophical Review*, *52*(4): 333-348. https://www.jstor.org/stable/2180668.

Lan, N., Chemla, E., and Katzir, R. (2023). "Large Language Models and the Argument From the Poverty of the Stimulus." *LingBuzz Preprint*. 1-48. lingbuzz/006829. https://lingbuzz.net/lingbuzz/006829.

Lappin, S. (2023). "Assessing the Strengths and Weaknesses of Large Language Models." *Journal of Logic, Language and Information*. 1-12. https://doi.org/10.1007/s10849-023-09409-x.

Leivada, E., Dentella, V., and Murphy, E. (2023). "The Quo Vadis of the Relationship Between Language and Large Language Models." *ArXiv*. 1-17. https://arxiv.org/abs/2310.11146.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., and Fedorenko, E. (2023). "Dissociating Language and Thought in Large Language Models." *ArXiv*, 1-41. https://arxiv.org/abs/2301.06627v2.

McGilvray, J. (2001). "Chomsky on the Creative Aspect of Language Use and Its Implications for Lexical Semantic Studies." In: F. Busa & P. Bouillon (Eds.), *The Language of Word and Meaning* (pp. 5-27). New York, NY: Cambridge University Press.

McGilvray, J. (2005). "Meaning and Creativity." In: J. McGilvray (Ed.), *The Cambridge Companion to Chomsky* (pp. 204-222). Cambridge, UK: Cambridge University Press.

McCoy, R. T., Smolensky, P., Linzen, T., Gao, J., and Celikyilmaz, A. (2023). "How Much Do Language Models Copy From Their Training Data? Evaluating Linguistic Novelty in Text Generation Using RAVEN." *Transactions of the Association for Computational Linguistics*, *11*, 652-670. https://doi.org/10.1162/tacl_a_00567.

Mikhail, J. (2011). *Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment*. New York, NY: Cambridge University Press.

Milway, D. "A Response to Piatandosi (2023)." *LingBuzz Preprint*, 1-4. lingbuzz/007264. https://lingbuzz.net/lingbuzz/007264.

Moro, A., Greco, M., and Cappa, S.F. (2023). "Large Languages, Impossible Languages, and Human Brains." *Cortex*, *167*, 82-85. https://doi.org/10.1016/j.cortex.2023.07.003.

Murphy, E. (2023). "Notes on Large Language Models and Linguistic Theory." *Elliot Murphy Blog*, April 26, 2023, Accessed January 13, 2024. https://elliot-murphy.com/2023/04/26/notes-on-large-language-models-and-linguistic-theory/.

O'Grady, W. and Lee, M. (2023). "Natural Syntax, Artificial Intelligence and Language Acquisition." *Information*, *14*(7), 1-10. https://doi.org/10.3390/info14070418.

OpenAI. (2022). "*OpenAI*. November 30, 2022." Accessed December 27, 2023. https://openai.com/blog/chatgpt.

Ouyang, L., et al. (2022). "Training Language Models to Follow Instructions with Human Feedback." ArXiv, 1-68. https://arxiv.org/abs/2203.02155v1.

Piantadosi, S. (2023). "Modern Language Models Refute Chomsky's Approach to Language." *Lingbuzz Preprint*, 1-57. lingbuzz/007180. https://lingbuzz.net/lingbuzz/007180.

Postal, P.M. (2005). "Foreword." In: Geoffrey Sampson (Ed.), *The 'Language Instinct' Debate: Revised Edition*, pp. vii-xi. London, UK: Continuum.

Pullum, Geoff. (2011). "Message 1: Remarks by Noam Chomsky in London." *LINGUIST List*, November 14, 2011, Accessed January 12, 2024. https://sites.socsci.uci.edu/~lpearl/courses/readings/Pullum2011_CommentsOnChomskyUG.pdf.

Rawski, J. and Baumont, L. (2023). "Modern Language Models Refute Nothing." *LingBuzz Preprint*, 1. lingbuzz/007203. https://lingbuzz.net/lingbuzz/007203.

Rey, Georges. (2020). *Representation of Language: Philosophical Issues in a Chomskyan Linguistics*. New York, NY: Oxford University Press.

Rizzi, L. (2017). "The Concept of Explanatory Adequacy." In: I. Roberts (Ed.), *The Oxford Handbook of Universal Grammar* (97-113). New York, NY: Oxford University Press.

Shanahan, M. (2022). "Talking About Large Language Models." *ArXiv*, 1-13. https://arxiv.org/abs/2212.03551v5.

Spelke, E. (2010). "Innateness, Choice, and Language." In: J. Franck and J. Bricmont (Eds.), *Chomsky Notebook* (pp. 203-210). New York, NY: Columbia University Press.

Weinberg, Steven. (1974). “Reflections of a Working Scientist.” *Daedalus*, *103*(3), 33-45. http://www.jstor.com/stable/20024218.

Weinberg, Steven. (1976). “The Forces of Nature.” *Bulletin of the American Academy of Arts and Sciences*, *29*(4), 13-29. https://doi.org/10.2307/3823787.

Wilcox, E. G., Futrell, R., and Levy, R. (2023). “Using Computational Models to Test Syntactic Learnability.” *Linguistic Inquiry*, 1-44. https://doi.org/10.1162/ling_a_00491.