Chapter 1

Modern language models refute Chomsky's approach to language

Steven T. Piantadosi[a,b]
[a]UC Berkeley, Psychology [b]Helen Wills Neuroscience Institute

The rise and success of large language models undermines virtually every strong claim for the innateness of language that has been proposed by generative linguistics. Modern machine learning has subverted and bypassed the entire theoretical framework of Chomsky's approach, including its core claims to particular insights, principles, structures, and processes. I describe the sense in which modern language models implement genuine *theories* of language, including representations of syntactic and semantic structure. I highlight the relationship between contemporary models and prior approaches in linguistics, namely those based on gradient computations and memorized constructions. I also respond to several critiques of large language models, including claims that they can't answer "why" questions, and skepticism that they are informative about real life acquisition. Most notably, large language models have attained remarkable success at discovering grammar without using any of the methods that some in linguistics insisted were necessary for a science of language to progress.

Introduction

After decades of privilege and prominence in linguistics, Noam Chomsky's approach to the science of language is experiencing a remarkable downfall. The story is, in part, a cautionary tale about what happens when an academic field isolates itself from what should be complementary endeavours. Chomsky's approach and methods are often argued to be problematic (e.g. Harris 1993, Pullum 1989, Behme 2012, Postal 2012, Behme 2014), but it is yet to be widely recognized just how the underlying ideas have been undermined by recent computational advances.

The state of the art in virtually all computational language tasks makes use of deep neural networks (LeCun et al. 2015), often neural networks called *transformers* (Vaswani et al. 2017, Brown et al. 2020, Bommasani et al. 2021). These are models of text that are trained on huge datasets of internet-based text to predict upcoming linguistic material (often at a sub-word but supra-character encoding). Several advances have allowed these models to become wildly successful. First, we have become able to train them on massive data sets. This is in part due to advances in computing (e.g. computing derivatives in arbitrary models) and in part the availability of huge collections of text from the internet. A typical language model might be trained on hundreds of billions of tokens, estimated to cost millions of dollars in energy alone. Second, the model architecture flexibly deals with nonlocal dependencies and allows the prediction of a word to potentially depend on material that is far away. The key consequence is that leading models are able to generate not only grammatical sentences, but entire discourses, scripts, explanations, poems, etc. This pursuit has been led in recent years by AI companies (Ahmed et al. 2023).

A nice example can be provided by the following, produced by ChatGPT when prompted with only the text in bold. All examples in this paper show the first response generated by ChatGPT to these queries by the author in early 2023. Good queries often focus on weird or unusual requests in order to ensure that the output is well outside of the training set.

Write a short story explaining how an ant could sink an aircraft carrier.
Once upon a time, in a far-off land, there lived a colony of ants. These ants were not like any ordinary ants, they were highly intelligent and had the ability to communicate with one another telepathically. They had been observing the humans and their ways for a long time, and they had become fascinated by the massive aircraft carriers that the humans built and sailed across the seas.
One day, the leader of the ant colony, a wise old ant named Alex, had an idea. He called all the ants together and said, ”My friends, we have the ability to think and communicate like the humans do. We can also work together as a team to achieve great things. I propose that we build a ship of our own, a ship that can sink the human's aircraft carrier.”
The other ants were shocked and amazed by Alex's proposal, but they were also excited by the idea. They immediately set to work, gathering twigs and sticks from the forest floor, and using their own tiny bodies to weave them together.
Months passed, and the ants' ship grew bigger and bigger. It was not as big as

an aircraft carrier, but it was large enough to hold many ants and was heavily armed with miniature torpedoes and cannons.
Finally, the day arrived when the ant ship was ready. The ants climbed aboard and set sail towards the nearest aircraft carrier. As they approached the carrier, the humans were shocked to see such a small ship coming towards them. They laughed and jeered, thinking that the ants' ship was no threat to them.
But the ants were not to be underestimated. They had been studying the human's ship, and they knew its weak spots. As they neared the carrier, they launched their miniature torpedoes, which hit the carrier directly in its vulnerable areas. The carrier began to sink, and the humans panicked.
In the end, the ant ship emerged victorious, having successfully sunk the massive aircraft carrier. The ants returned home to their colony as heroes, and their story became a legend that was passed down through the generations.
And the human learned to never underestimate the power of collective intelligence and unity.
It is important to note that the story above is a fictional tale and it is impossible for an ant to sink an aircraft carrier, as in reality the size, strength and weapon capability of an aircraft carrier would far exceed the capability of a group of ants.

Ignore whatever theoretical persuasions you have for a moment and just recognize how *remarkable* that response is. A coherent and grammatical story unfolds from a simple prompt, even though the requested scenario strains plausibility and probability. The response tackles the core essence of the question to provide a compelling story about how an ant *might* sink an aircraft carrier. The answer is outside of the training set. In fact, after "Once upon a time, in a far-off land, there lived a colony of ants," a Google search returns no matching strings on the entire internet.

Anyone who has dealt in simpler language models like *n*-grams or context-free grammars will be impressed that this model is capable of incorporating not just longer dependencies within sentences, but across sentences. It also creates complex structures, with e.g. multiple embedded clauses, prepositional phrases, conjunctions, etc; it uses pronouns, determiners, quantifiers, adjectives, etc. in realistic ways; this and similar models have few apparent problems with agreement or pronoun reference (Gulordava et al. 2018, Goldberg 2019). Any one of these abilities has been the topic of countless dissertations in linguistics. Beyond that, such models succeed on tasks that formal linguistic theories haven't yet dreamed of—like interpreting the semantics of the words in question, maintaining not just coherent sentences but entire coherent discourses. Such models exhibit over 90% accuracy in Winograd-style questions (Kocijan et al. 2020, 2022), which require interpreting language in world contexts (e.g. "The trophy doesn't fit in the brown suitcase because it is too large. What is too large?"). What is clear is that these models *use language in a way that is remarkably human* (Mahowald & Ivanova et al. 2023).

This kind of example hearkens back to Turing 1950, who considered asking an artificial intelligence to write a sonnet on the topic of "The Forth Bridge"[1], and suggested that success in solving such tasks would demonstrate many key aspects of humanlike intelligence. Contemporary language models can do much more—they can write raps, explain things, imagine conversations, create a mash-up script of Seinfeld and Star Trek, etc. They are aware of even sub-word properties like rhymes:

Come up with a limerick about syntactic island constraints in linguistics.
On syntactic islands, we must be precise,
For constraints on movement can be quite nice.
But when the constraints are breached,
Our grammars are impeached,
And our linguistic analyses pay the price.

And they are aware of style:

Explain the fundamental theorem of arithmetic in the style of Donald Trump.
Well, let me tell you folks, the fundamental theorem of arithmetic is a really great theorem, it's tremendous, and it's all about prime numbers. Believe me, I know a lot about prime numbers. They're very important, very powerful, very beautiful numbers.
Now, the fundamental theorem of arithmetic tells us that every whole number greater than 1 can be expressed as a unique product of prime numbers. That's right, every single number can be broken down into its prime factors, and there's only one way to do it. It's like a code, a secret code that only primes can unlock.
This is a really amazing theorem, folks, and it's incredibly useful. It's used in cryptography, number theory, and all sorts of other fields. It's a cornerstone of mathematics, and it's something that we should all be very proud of.
So, in conclusion, the fundamental theorem of arithmetic is a fantastic theorem, it's tremendous, and it's all about prime numbers. I love prime numbers, they're the best. And this theorem tells us that every number can be broken down into primes in a unique way. It's a great thing, believe me

[1]According to Wikipedia, a cantilevered rail bridge from 1890 in Scotland that was voted in 2016 to be Scotland's greatest human-made wonder, apparently beating out Ewan McGregor.

Note that this specific example was not in the model's training set—there is no possibility that Trump understands prime numbers. However information about the fundamental theorem of arithmetic was, as was the speech style of Donald Trump. "Believe me, I know a lot about prime numbers" is probably not something Trump has actually said, but certainly something he *would* say.

Examples like these show why dismissals from cognitive scientists like Gary Marcus[2] that the models are just the same as "autocomplete" systems on your phone are far too shallow. The model is able to put together things in its training *in new ways* that maintain a considerable amount of linguistic and conceptual coherence. That requires more than merely guessing things it has seen before—it requires modeling the dynamics of language. If models only repeated what they had seen before, they would not be able to generate anything new, particularly complex sentence structures that are grammatical and coherent. It is somewhat difficult to convey how remarkable the models are currently. You just have to interact with them. They are imperfect, to be sure, but my qualitative experience interacting with them is like talking to a child, who happened to have memorized much of the internet.

How this magic happens

These tools are not just impressive, they are philosophically important. The reason they are important is that they succeed by following a very particular approach: they are trained *only on text prediction*.[3] This means that the models form probabilistic expectations about the next word in a text and they use the true next word as an error signal to update their latent parameters. This idea is one that dates back to at least Elman 1990, who showed how training a neural network on text prediction could lead it to discover key pieces of the underlying linguistic system.

Modern models are a resounding scientific victory for Elman's idea. But while modern models inherit his general setup, advances have added a few critical differences. Probably the most important is that modern models include an *attentional mechanism* that allows the next word in sequence to be predicted from some previous far in the past. For example, in the ant story above, when it says that "The other ants were shocked and amazed by Alex's...", it retrieves the name

[2]https://garymarcus.substack.com/p/nonsense-on-stilts

[3]The underlying neural network weights are typically optimized in order to predict text, but note that many applications of these models also use human feedback to fine-tune parameters and try to tamp down the horrible things text on the internet leads models to say.

"Alex" from dozens of words prior. This likely is the key property that distinguishes large language models from the most popular earlier models. An n-gram model, for example, would estimate and use a conditional probability that depends on just the preceding few words (e.g. $n = 2, 3, 4, 5$); context-free grammars make independence assumptions that keep lexical items from influencing those far away. Not only do large language models allow such long-distance influences, but they allow them to take a relatively unconstrained form and so are able to induce functions which, apparently, do a stellar job at in-context word prediction.

A second key feature of these models is that they *integrate semantics and syntax*. The internal representations of words in these models are stored in a vector space, and the locations of these words include not just some aspects of meaning, but properties that determine how words can occur in sequence (e.g. syntax). There is a fairly uniform interface for how context and word meaning predicts upcoming material—syntax and semantics are not separated out into distinguished components in the model, nor into separate predictive mechanisms. Because of this, the network parameters these models find blend syntactic and semantic properties together, and both interact with each other and the attentional mechanism in nontrivial ways. This doesn't mean that the model is incapable of distinguishing syntax and semantics, or e.g. mirroring syntactic structures regardless of semantics (see examples below), but it does mean that the two can be mutually informative. A related aspect of the models is that they have a huge memory capacity of billions to trillions of parameters. This allows them to memorize idiosyncrasies of language, and in this way they inherit from a tradition by linguists who have emphasized the importance of constructions (Goldberg 1995, Jackendoff 2013, Goldberg 2006, 2003, Tomasello 2000, McCauley & Christiansen 2019, Tomasello 2005, Edelman & Waterfall 2007) (see Weissweiler et al. 2023 for construction grammar analyses of large language models). Such models also inherit from the tradition of learning bottom-up, from data (e.g. Bod et al. 2003, Solan et al. 2005), and computational work which explicitly connects syntax and semantics (Steedman 2001, Siskind 1996, Ge & Mooney 2005, Kwiatkowski et al. 2012, Liang et al. 2009).

A good mental picture to have in mind for how massively over-parameterized models like these work is that they have a rich potential space for inferring *hidden variables and relationships*. Hidden (or latent) variables have been one of the key aspects of language that computational and informal theories alike try to capture (Pereira 2000, Linzen & Baroni 2021). In the middle of a sentence, there is a hidden variable for the latent structure of the sentence; in speaking an ambiguous word, we have in mind a hidden variable for which meaning we intend; throughout a discourse we have in mind a larger story arc that only unfolds across multiple sentences. The formalisms of linguistics attempt to characterize these hidden variables too. But what large language models do is *infer* likely hidden structure because that structure permits them to better predict upcoming material. This makes them conceptually similar to embedding theorems in mathematics (Packard et al. 1980, Takens 1981, Ye & Sugihara 2016), which show that sometimes the full geometry of a dynamical system can be recovered from a low-dimensional projection of its states evolving in time. Linguistic corpora are a low-dimensional projection of both syntax and thought, so it is not implausible that a smart learning system could recover at least some aspects of these cognitive systems from watching text alone (Piantadosi & Hill 2022) .

The structures present in large language models can be seen in detailed analyses, where as the model generates text after training, its internal states represent latent aspects of syntactic structure and semantic meaning (Manning et al. 2020, Futrell et al. 2019, Linzen & Baroni 2021, Pavlick 2022). The structure of the model's internal representation states and attentional patterns after training comes to capture tree structures with strong similarities to human-annotated parse trees (Manning et al. 2020), and the degree to which a model is tree-structured even predicts its generalization performance (Murty et al. 2022). The models seem to perform well on constructions involving tracking the right latent state, like function words (Kim et al. 2019) and filler-gap dependencies (Wilcox et al. 2018). In fact, the internal processing structure of some models seems to spontaneously develop an intuitive pipeline of representing parts of speech, followed by parsing, semantic analysis, etc. (Tenney, Das, et al. 2019, Liu et al. 2019).

All of this is possible because large language models develop representations *of* key structures and dependencies, it's just that these representations are parameterized in a way which is unfamiliar to linguistics. As argued by Baroni 2022, this means that language models should be treated as bona fide linguistic *theories*. Specifically, a space of possible theories is parameterized by the models and compared to data to find which theory is best in a formal sense. To make one version of the idea concrete, imagine a physicist who wasn't sure whether the law of gravitation force fell off with distance $1/r$ or distance squared $1/r^2$. To decide, the physicist might formulate a super-equation that captured both possibilities, for instance,

$$F(r, \alpha) = \alpha \cdot \frac{1}{r} + (1 - \alpha) \cdot \frac{1}{r^2}.$$

By seeing which parameter α best fits empirical data (e.g. measurements of forces F and distances r), they are *comparing* these two theories: $\alpha \approx 1$ means the former theory is right (since then $F = 1 \cdot \frac{1}{r} + 0$) and $\alpha \approx 0$ means the latter (since

$F = 0 + 1 \cdot \frac{1}{r^2}$). When the data are stochastic, a good way to measure how well any particular α does is to see what probability it assigns to the data. We can make a principled choice between parameters—and thus theories—by choosing the one that makes the data most likely (*maximum likelihood principle*), although often including some prior information about plausible parameter values or penalties on complexity (e.g. Bayesian estimation). Such a physicist might then find that the best parameter for capturing data has $\alpha \approx 0$, supporting the second theory.[4] In this case, inferring parameters *is* comparing theories: in computational modeling, there is no bright line between "just" fitting parameters and advancing theory.

Something very similar happens with many machine learning models, the main difference is that in these models, we don't explicitly or intentionally "build in" the theories under comparison ($1/r$ and $1/r^2$). There exist natural bases from which you can parameterize essentially *any* computational theory.[5] Parameter fitting in these models is effectively searching over a huge space of possible theories to see which one works best, in a well-defined, quantitative sense.

The bases required are actually pretty simple. Polynomials are one, but neural networks with sigmoid activations are another: fitting parameters in either of these can, in principle, realize countless possible relationships in the underlying domain. The challenge is that when these universal bases are used, it requires extra scientific work to see and understand what the parameters mean. Just to illustrate something roughly analogous, if we happened to write the above equation in a less transparent way,

$$F(r, \alpha) = \frac{\alpha \cdot (r - 1) + 1}{\log\left((e^r)^r\right)}$$

Then it might take some work to figure out which α values correspond to $1/r$ and which to $1/r^2$. Squint just a little and you can imagine that instead of algebra, we had a mess of billions of weighted connections between sigmoids to untangle and interpret. It becomes clear that it could be hard to determine what is going on, even though *the theory is certainly in there.*

In fact, we don't deeply understand *how* the representations these models create work (see Rogers et al. 2021). It is a nontrivial scientific program to discover how their internal states relate to each other and to successful prediction. Researchers have developed tools to "probe" internal states (e.g. Belinkov & Glass

[4] For decades, other fields have used statistical learning models that take empirical data and infer laws (Koza 1994, Langley et al. 1983, Schmidt & Lipson 2009, Udrescu & Tegmark 2020).

[5] Up to the capacity of the network.

2019, Tenney, Xia, et al. 2019, Kim et al. 2019, Linzen & Baroni 2021, Warstadt & Bowman 2022) and determined some of the causal properties of these models. At the same time, this does not mean we are ignorant of all of the *principles* by which they operate. We can tell from the engineering outcomes that certain structures work better than others: the right attentional mechanism is important (Vaswani et al. 2017), prediction is important, semantic representations are important, etc. The status of this field is somewhat akin to the history of medicine, where people often worked out what kinds of treatments worked well (e.g. lemons treat scurvy) without yet understanding the mechanism.

One thing that is interesting is how modern language models integrate varied computational approaches to language, not by directly encoding them, but by allowing them to emerge (Manning et al. 2020) from the architectural principles that *are* built-in (Elman et al. 1996). For example, the models appear to have representations of hierarchy (Manning et al. 2020) and recursion, in the sense that they know about e.g. embedded sentences and relative clauses. They also almost certainly have analogs of constraints, popular in approaches like harmonic (Smolensky & Legendre 2006, Prince & Smolensky 1997) and model-theoretic grammar (Pullum 2007, 2013). The models likely include both hard constraints (like word order) and violable, probabilistic ones (Rumelhart & McClelland 1986). They certainly memorize some constructions (Goldberg 1995, Jackendoff 2013, Goldberg 2006, 2003, Tomasello 2000, Edelman & Waterfall 2007). All of those become realized in the parameters in order to achieve the overarching goal of predicting text well.

The status of large language models as scientific theories

Many in language science see such models as at least relevant in some way to the future (Bommasani et al. 2021, Baroni 2022, Pater 2019). After all, they are the only models in existence that do a good job of capturing the basic dynamics of human language. However, in virtue of being neural networks, their—at least initial—state is are wholly unlike the rules and principles that have dominated generative approaches to language. As described above, their parameters come to embody a theory of language, including representations of latent state through a sentence and a discourse. The exact same logic of tuning parameters to formalize and then compare theories is found in other sciences, like modeling hurricanes or pandemics: any set of assumptions will generate a distribution of predictions and the assumptions are adjusted to make the best predictions possible. In this way, a learning mechanism configures the model itself in the space of theories in order to satisfy the desired objective function. For hurricanes or pandemics, this is as rigorous as science gets; for sequences of words, everyone seems to lose their mind.

In discussing GPT-3 with Gary Marcus,[6] for example, the most positive thing Chomsky could say was that it has an "ability to mimic some regularities in data", followed quickly by "In fact, its only achievement is to use up a lot of California's energy"[7]. In another interview, he summarized that the models "have achieved zero" in terms of our understanding of language. Chomsky et al. 2023 characterized the models as useful "in some narrow domains" but hampered by "ineradicable defects" that made them "differ profoundly from how humans reason and use language." As was quickly pointed out online, several of the examples they brought up—like reasoning with counterfactuals or understanding sentences "John is too stubborn to talk to"—current models actually get correct. Chomsky et al. 2023 tilt at an imagined version of these models, while ignoring the fact that the real ones so aptly capture syntax, a success Chomsky and others have persistently claimed was impossible.

Part of why some generative linguists dismiss these models is that they are seen as too unconstrained, and thus not explanatory. Writing to Marcus about the models, Chomsky explains,

> You can't go to a physics conference and say: I've got a great theory. It accounts for everything and is so simple it can be captured in two words: "Anything goes."
>
> All known and unknown laws of nature are accommodated, no failures. Of course, everything impossible is accommodated also.

This critique is a familiar rephrasing of his (and others') comments on language learning—essentially that one should not study an unconstrained system because it will not explain why languages have the particular form that they do.[8]

But it is too coarse a gloss to dismiss modern language models as "anything goes." The reason is that not all "anything goes" models are equivalent. A three-layer neural network is well-known to be capable of approximating any computable function (Siegelmann & Sontag 1995). That's also an "anything goes"

[6] https://garymarcus.substack.com/p/noam-chomsky-and-gpt-3

[7] One of the many limitations, concerns, and dangers (Bender et al. 2021, Bommasani et al. 2021) for these models is that they consume a lot of energy. It's estimated that these models take around 1000 MWh, compared to CA's daily generation of about 750, 000 MWh—so one model takes about 1/750'th of one day of CA's power.

[8] You have to wonder how a physics conference would react to someone saying, following Lasnik 2002, "I've got a great theory. It accounts for everything: physical laws 'might be a computationally perfect solution to the problem' of how objects move."

model. But the three-layer network will *not* work well on this kind of text prediction. Indeed, even some earlier neural network models, LSTMs, did not do as well (Futrell et al. 2019, Marvin & Linzen 2018, Hu et al. 2020); architectures generally vary in how well they capture computational classes of string patterns (e.g. Delétang et al. 2022).[9]

We are granted scientific leverage by the fact that models that are equally powerful in principle perform differentially. In particular, we may view 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. This is how, for example, the field has discovered that attentional mechanisms are important for performing well. Similarly, "ablation" experiments allow researchers to alter one part of a network and use differing performance to pinpoint what principles support a specific behavior (see Warstadt & Bowman 2022).

Even when—like all scientific theories—we discover how they fail to match people in terms of mechanism or representation, they still are informative. Heeding George Box's advice that "all models are wrong, some are useful," we can think about the scientific strengths, contributions, and weaknesses of these models without needing to accept or dismiss them entirely. In fact, these models have already made a substantial scientific contribution by helping to delineate what is *possible* through this kind of assumption-testing: Could it be possible to discover hierarchy without it being built in? Could word prediction provide enough of a learning signal to acquire most of grammar? Could a computational architecture achieve competence on WH-questions without movement, or use pronouns without innate binding principles? The answer to all of these questions is shown by recent language models to be "yes."

Beyond that, the models embody several core desiderata of good scientific theories. First, they are **precise and formal** enough accounts to be implemented

[9]Some also consider them not to be "scientific" theories because they are *engineered.* In an interview with Lex Friedman, Chomsky remarked, "Is [deep learning] engineering, or is it science? Engineering, in the sense of just trying to build something that's useful, or science, in the sense that it's trying to understand something about elements of the world … We can ask that question, is it useful? Yeah, it's pretty useful. I use Google Translator. So, on engineering grounds it's kinda worth having, like a bulldozer. Does it tell you anything about human language? Zero, nothing." In practice, there is often no clear line between engineering and science because scientists often need to invent new tools to even formulate theories: was Newton's calculus engineering instead or science? The machinery of transformational grammar? While the recent successes are due to engineering advances, researchers have been arguing for this form of model *as* cognitive theories for decades.

in actual computational systems, unlike most parts of generative linguistics. Implementation permits us to see that these theories are internally consistent and logically coherent. In virtue of being implemented, such models are able to **make predictions**. Just to list a few examples, the patterns of connectivity and activation within large language models appear to capture dependency structures in words via attention (Manning et al. 2020). Their predictability measures can be compared to psychological measures (Hoover et al. 2022, Shain et al. 2022). Transformer models "predict nearly 100% of explainable variance in neural responses to sentences" (Schrimpf et al. 2021).

Unlike generative linguistics, these models show promise in being **integrated with what we know about other fields**, specifically cognition and neuroscience. Many authors interested in human concepts have investigated the vector representations that the models form (Lake & Murphy 2021, Bhatia & Richie 2022). Surprisingly or not, the language model vectors appear to encode at least some aspects of semantics (Maas et al. 2011, Socher et al. 2013, Bowman et al. 2015, Grand et al. 2022, Bhatia & Richie 2022, Piantadosi & Hill 2022, Dasgupta et al. 2022, Petersen & Potts 2022, Pavlick 2022), building on earlier models that encoded semantics in neural networks (e.g. Rogers, McClelland, et al. 2004, Elman 2004, Mikolov et al. 2013). In fact, their semantic spaces can be aligned with the world with just a few labeled data points, at least in simple domains (Patel & Pavlick 2022). The representations that they learn can also be transferred to some degree across languages (Pires et al. 2019, Chi et al. 2020, Gonen et al. 2020, Papadimitriou & Jurafsky 2020, Papadimitriou et al. 2021, Hill et al. 2017), suggesting that they are inferring something deep about meaning. Following leading theories of concepts (Block 1986, 1998), the representations that language models learn may be meaningful in the sense of maintaining nontrivial *conceptual roles* (Piantadosi & Hill 2022), contrary to claims that meaning requires connections to the real world (Bender & Koller 2020). Building on the "parallel and distributed" tradition of cognitive modeling (McClelland et al. 1986), modern deep learning models are also likely able to be integrated with neuroscientific theories (Marblestone et al. 2016, Richards et al. 2019, Kanwisher et al. 2023, McClelland et al. 2020). In particular, they make predictions about neural data (e.g. Schrimpf et al. 2021, Caucheteux et al. 2022, Goldstein et al. 2022). Generative theories of syntax, by contrast, suffer from a "chronic lack of independent empirical support" and in particular have not been compellingly connected to neuroscience (Edelman 2019).[10]

[10]"Considering how central the existence of a brain basis for syntax is to Chomskian (bio)linguistics, the scarcity of behavioral and brain evidence for syntactic structures is strik-

Moreover, these models **are empirically tested**, especially as a theory of grammar. Modern language models are state of the art in most natural language processing domains (Bommasani et al. 2021). Approaches from generative syntax are not competitive in any domain and arguably have avoided empirical tests of their core assumptions (Edelman & Christiansen 2003). Several authors have sought to quantitatively evaluate large language models on the syntactic constructions that motivate much of linguistic theory. Early results found some successes across a variety of different architectures. For example, Warstadt, Singh, et al. 2019 evaluated LSTM neural network models on a corpus of more than ten thousand acceptability judgments published in the linguistics literature. They found that these models perform about in the mid 70% range, compared to human reliabilities in the upper 80% to 90% range. Warstadt, Cao, et al. 2019 look at BERT embeddings on negative polarity items and finds "significant knowledge" of these structures but that success depends heavily on how the structures are tested. More recently, GPT models show high performance on filler-gap structures, including various forms of islands (Wilcox et al. 2018, Wilcox et al. 2022).

Gauthier et al. 2020's *SyntaxGym* is a standardized environment for testing models against a number of standardized test suites that captures linguistic constructions and phenomena like clefts, center-embedding, cataphors, negative polarity items, filler-gap dependencies, subordination, agreement, etc. This project builds on similarly exciting benchmarking efforts in neuroscience (Schrimpf et al. 2020). As of the writing of this article, the state of the art for language was a variant of GPT-2 that achieved nearly 90% on these constructions. SyntaxGym is an ingenious resource that finally allows quantitative comparison of theories. A general trend is that the more recent language models perform better, though it appears that proponents of (e.g.) minimalist grammars have not compared theories on SyntaxGym. It is not clear there has ever been, in the history of science, an implemented computational system that achieves such accuracy, but which is dismissed by a field which has failed to develop even remotely comparable alternatives.

In the spirit of considering large language models as scientific theories, it's worth also highlighting their limitations. One is that while they succeed at language modeling, they are currently less successful in domains that require reasoning or thinking (Mahowald & Ivanova et al. 2023, Lake & Murphy 2021, Barrett et al. 2018, Collins et al. 2022). From an acquisition perspective, likely the most

ing. … In comparison to the basic phrase structure, evidence supporting the reality of more far-fetched theoretical constructs, such as movement, traces/copies, etc., remains elusive." (Edelman 2019) important limitation of current models is that are they trained on truly titanic datasets compared to children, by a factor of at least a few thousand (see Warstadt & Bowman 2022 for a comprehensive review of models in language acquisition). Moreover, these datasets are strings on the internet rather than child-directed speech. Work examining the scaling relationship between performance and data size shows that at least current versions of the models do achieve their spectacular performance only with very large network sizes and large amounts of data (Kaplan et al. 2020). However, Zhang et al. 2020 show that actually most of this learning is not about syntax. Models that are trained on $10-100$ million words "reliably encode most syntactic and semantic features" of language, and the remainder of training seems to target other skills (like knowledge of the world). This in fact matches in spirit analyses showing that syntactic knowledge requires a small number of bits of information, especially when compared to semantics (Mollica & Piantadosi 2019). Hosseini et al. 2022 present evidence that models trained on developmentally-plausible amounts of data already capture human neural responses to language in the brain.

Importantly, as Warstadt & Bowman 2022 outline, these models are in their early stages of development, so their successes are likely to be more informative about the path of children's language acquisition than the models' inevitable limitations. Current models provide a lower-bound on what is possible, but even the known state-of-the-art doesn't characterize how well future models may do. Our methods for training on very small datasets will inevitably improve. One improvement might be to build in certain other kinds of architectural biases and principles; or it might be as simple as finding better optimization or regularization schemes. Or, we might need to consider learning models that have some of the cognitive limitations of human learners, as in Newport 1990's "less is more" hypothesis. Such questions inspire the current "The BabyLM Challenge" (Warstadt et al. 2023), which aims to develop models capable of learning with a developmentally-plausible amount of data (see Geiping & Goldstein 2022 for training models with small amounts of compute resources). It is an interesting scientific question whether low-resource, low-data learning is possible—I'll preregister a prediction of yes, with small architectural tweaks.

The refutation of key principles

The success of large language models is a failure for generative theories because it goes against virtually all of the principles these theories have espoused. In fact, *none* of the principles and innate biases that Chomsky and those who work in his tradition have long claimed necessary needed to be built into these models (e.g. binding principles, binary branching, island constraints, empty category principle, etc.). Moreover, these models were created without incorporating any of Chomsky's key methodological claims, like ensuring the models properly consider competence vs. performance, respect "minimality" or "perfection," and avoid relying on the statistical patterns of unanalyzed data.

The next sections focus on a few examples.

Syntax is integrated with semantics

Chomsky and others have long emphasized the study of syntax as a separate entity, not only from the rest of cognition but from the rest of language (see, e.g. Chomsky 1957, Croft 1995, Newmeyer 1991, Adger 2018). Syntax in this approach is not supposed to be reducible to general statistics between words[11]—exactly the thing that large language models now provide.

Modern large language models integrate syntax and semantics in the underlying representations: encoding words as vectors in a high-dimensional space, without an effort to separate out e.g. part of speech categories from semantic representations, or even predict at any level of analysis other than the literal word. Part of making these models work well was in determining how to encode semantic properties into vectors, and in fact initializing word vectors via encodings of distribution semantics from e.g. Mikolov et al. 2013 (Radford et al. 2019). Thus, an assumption of the autonomy of syntax is not required to make models that predict syntactic material and may well hinder it.

We can have fun with the *colorless green ideas* sentence (see Gulordava et al. 2018) commonly offered as an example that syntax functions separately from semantics (and that transitional probability statistics are not enough.) The claim that this is out of reach of statistical models has been shown to be wrong (Pereira 2000, Norvig 2012, Gulordava et al. 2018), but it is an interesting example nonetheless, in part because ChatGPT can not only learn relevant statistics, but it can create—in this case surely mostly regurgitate—an explanation of why this sentence is interesting:

Why is the sentence "colorless green ideas sleep furiously" interesting?
The sentence "colorless green ideas sleep furiously" is similarly interesting because it is a grammatically correct sentence, but it is semantically meaningless. The words used in the sentence are arranged in a way that follows standard syntax and grammar rules, but they do not form a coherent or understandable

[11] "There is no current theory of linguistics that takes human syntactic capacities to be reducible to general probabilistic relations between elements of unanalysed data" (Adger 2018)

concept. This makes it an interesting linguistic oddity and a popular example of a nonsensical sentence.

We can test generalization with a follow-up:

Generate ten other sentences like this
Purple fluffy clouds dream wildly
Blue glittery unicorns jump excitedly
Orange sparkly penguins waddle energetically
Red shimmering elephants trumpet melodiously
Yellow dazzling lions roar majestically
Green glittery monkeys swing energetically
Pink sparkly giraffes walk gracefully
Black shiny kangaroos hop playfully
White glittery bears cuddle affectionately
Brown shimmering rabbits hop merrily.

This model successfully gets the referent of "this" (a sentence). The model resolves the ambiguity in "like this" to refer to the *structure* of the sentence—the exact thing that statistical models weren't even supposed to know! It generates some but not entirely low-frequency bigrams in the sentence. We can note a weakness in that it does not as readily generate wholly meaningless sentences ("black shiny kangaroos" is rare but not impossible), likely because meaningless language is rare in the training data. These results illustrate that even models which integrate syntax and semantics are able to make syntactic generalizations when appropriate. Syntax need not be autonomous in the underlying mechanism or model's analysis in order to show this behavior.

Probability and information are central

A second point is that for these models, probability and information theory are central. Chomsky has long been dismissive of probability, saying "it must be recognized that the notion of 'probability of a sentence' is an entirely useless one, under any known interpretation of this term" (Chomsky 1969), a position he has maintained for decades (Norvig 2012).[12] Often when those who work in Chomsky's tradition talk about probability models, they refer to simple things like Shannon 1948's *n*-gram models that count up sequential word co-occurrences

[12] Or in Chomsky 1957, "I think that we are forced to conclude that grammar is autonomous and independent of meaning, and that probabilistic models give no particular insight into some of the basic problems of syntactic structure."

and were long used in natural language processing tasks (Chen & Goodman 1999, Manning & Schutze 1999). But by now, such models are decades out of date.

Newer models use probability to infer entire generating processes and structures, a common cognitive task and modeling domain (e.g. Tenenbaum et al. 2011, Ullman et al. 2012, Lake et al. 2015, Goodman et al. 2011, Lake et al. 2017, Rule et al. 2020, Kemp & Tenenbaum 2008, Yang & Piantadosi 2022). Probability is central for models because a probabilistic prediction essentially provides an error signal that can be used to adjust parameters that themselves encode structure and generating processes. An analogy is that one might imagine watching a driver and inferring the relevant structures and dynamics from observation—rules of the road (which side you drive on), conventions (behavior of multiple cars at stop signs), hard and soft constraints (don't turn too hard), etc. Even a simple domain like this faces many of the problems of undetermination seen in language, but is is one where it is easy to imagine a skilled scientist or anthropologist discovering the key elements by analyzing a mass of data. Something similar goes on in machine learning, where a space of possible rules is implicitly encoded into the parameters of the model (see above).

It is worth noting that most models which deal in probabilities actually work with the log of probabilities, for reasons of numerical stability. Models that work on log probabilities are actually working in terms of description length (Shannon 1948, Cover 1999): finding parameters which make the data most likely (maximizing probability) is the same as finding parameters which give the data a short description (minimizing description length or complexity). Thus, the best parameters are equivalent to scientific theories that do a good job of *compressing* empirical data in the precise sense of description length. Far from "entirely useless," probability is the measure that permits one to actually quantify things like complexity and minimality.

Representations are continuous and gradient

The fact that predictions are probabilistic is useful because it means that the underlying representations are continuous and gradient. Unlike work formalizing discrete rules and processes, typical of generative linguistics (e.g. Chomsky 1956, 1995, Collins & Stabler 2016, Chomsky 1957, Pinker & Prince 1988), modern language models do not use (at least explicit) rules and principles—they are based in a continuous calculus that allows multiple influences to have a gradient effect on upcoming linguistic items. The foundation for this approach was laid by early modelers like Rumelhart & McClelland 1986, who argued for the key features of today's architectures decades ago, including that "cognitive processes are seen as graded, probabilistic, interactive, context-sensitive and domain-general." (McClelland & Patterson 2002).

Continuity is important because it permits the models to use gradient methods—essentially a trick of calculus—to compute what direction to change all the parameters in order to decrease error the fastest. Tools like TensorFlow and PyTorch that permit one to take derivatives of arbitrary models have been a critical methodological advance. This is not to say that these models end up with no discrete values—after all, they robustly generate subjects before verbs when trained on English. Similarly, the $F(r, \alpha)$ example might end up with a discrete answer like $\alpha \approx 0$. They key is that discreteness is a special case of continuous modeling, meaning that theories which work with continuous representations get the best of both worlds, fitting discrete patterns when appropriate and gradient ones otherwise. The success of gradient models over deterministic rules suggests that quite a lot of language is based on gradient computation. The success actually mirrors the prevalence of "relaxation" methods in numerical computing, where an optimization problem with hard constraints is often best solved via a nearby soft, continuous optimization problem. Thus, contrary to the intuition of many linguists, *even if* we wanted a hard, discrete grammar out at the end, the best way for a learner to get there might be via a continuous representation.

Learning succeeds in an unconstrained space

Perhaps most notably, modern language models succeed despite the fact that their underlying architecture for learning is relatively unconstrained. This is a clear victory for statistical learning theories of language (see (Contreras Kallens et al. 2023)). These models are capable of fitting a huge number of possible patterns, and while the principles of their architecture do constrain them to make some patterns easier than others, the resulting systems are incredibly flexible. Despite this lack of constraint, the model is able to figure out much of how language works. One should not lose sight of the role that "poverty of the stimulus" arguments have long played for generative linguists (e.g. Lasnik & Lidz 2016, Crain & Pietroski 2001, Legate & Yang 2002, Wexler & Culicover 1980, Laurence & Margolis 2001, Lasnik & Lidz 2016, Pearl 2022, Crain & Pietroski 2002). Poverty of the stimulus claims have been compellingly challenged both on empirical grounds about the nature of input, and through learning theories that acquire the relevant structures from input (e.g. Pullum & Scholz 2002, Clark & Lappin 2010, Perfors et al. 2011, Reali & Christiansen 2005, Solan et al. 2005). Large language models essentially lay this issue to rest because they come with none of the constraints that others have insisted are necessary, yet they capture almost all key phenomena (e.g. Wilcox et al. 2022). It will be important to see, however, how well they can do on human-sized datasets, but their ability to generalize to sentences outside of their training set is auspicious for empiricism.

Recall that many of the learnability arguments were *supposed* to be mathematical and precise, going back to Gold 1967 (though see Johnson 2004, Chater & Vitányi 2007) and exemplified by work like Wexler & Culicover 1980. It's not that we don't know the right learning mechanism; it's supposed to be that it can be proven none exists. Even my own generative syntax textbook from undergraduate syntax purports to show a "proof" that because infinite, productive systems cannot be learned, parts of syntax must be innate (Carnie 2021). Legate & Yang 2002 call the innateness of language "not really a hypothesis" but "an empirical conclusion" based on the strength of poverty of stimulus arguments. Proof of the impossibility of learning in an unrestricted space was supposed to be the power of this approach. It turned out to be wrong.

The notion that the core structures of language could be discovered without substantial constraints may sound impossible to anyone familiar with the generative syntax rhetoric. But learning without constraints is not only possible, it has been well-understood and even *predicted*. Formal analyses of learning and inference show that learners can infer the correct theory out of the space of possible computations (Solomonoff 1964, Hutter 2004, Legg & Hutter 2007). In language specifically, the correct generating system for grammars can similarly be discovered out of the space of all computations (the most unrestricted space possible), using only observations of positive evidence (Chater & Vitányi 2007).

In this view, large language models function somewhat like automated scientists or automated linguists, who also work over relatively unrestricted spaces, searching to find theories which do the best job of parsimoniously predicting observed data. It's worth thinking about the standard lines of questioning generative syntax has pursued—things like, why don't kids ever say "The dog is believed's owners to be hungry" or "The dog is believed is hungry" (see Lasnik & Lidz 2016). The answer provided by large language models is that these are not permitted under the best theory the model finds to explain what it does see. Innate constraints are not needed.

Representations are representationally complex, not minimal

Next, there is an important sense in which large language models are not minimal representationally, but maximal. What I mean is that there is not a single core nugget of representation or structure (like *merge*) that leads these models to succeed. Nor are any biases against derivational complexity likely to play a key role, since everything is a single big matrix calculation. This calculation moreover is not structurally minimal or "perfect" in the sense that minimalist linguistics means (e.g. Lasnik 2002). Instead, the attentional mechanisms of large language models condition on material that is arbitrarily far away, and perhaps not even structurally related since this is how they model discourses between sentences. A grammatical theory that relies on people's capacity for memorizing countless chunks of language changes the landscape of how we should think about derivation; as above, models based in probability give a formal footing to notions of complexity in grammar.

Deep learning has actually changed how people think about complexity in statistical learning too. It has long been observed that having too many parameters in a model would prevent the model from generalizing well: too many parameters allow a model to fit patterns in the noise, and this can lead it to extrapolate poorly. Deep learning turned this idea on its head by showing that some models will fit (memorize) random data sets (Zhang et al. 2021), meaning they can fit all the patterns in the data (including noise) *and* generalize well. The relationship between memorization and generalization is still not well-understood, but one of the core implications is that statistical learning models can work well, sometimes, even when they are over-parameterized.

While discussing statistical learning (before deep learning) with Peter Norvig, Chomsky noted that "we cannot seriously propose that a child learns the values of 10^9 parameters in a childhood lasting only 10^8 seconds." One has to wonder if a similar argument applies to biological neurons: humans have 80 billion neurons, each with thousands of synapses. If childhood is only 10^8 seconds, how do all the connections get set? Well, also note that the 3 billion base pairs of the human genome certainly can't specify the precise connections either. Something must be wrong with the argument.

Two missteps are easy to spot. First, even if a model has a billion parameters, they will not generally be independent. This means that a single data point could set or move thousands or millions or billions of parameters. For example, observing a single sentence with SVO order might increase (perhaps millions of) parameters that put S before V, and decrease (perhaps millions of) parameters that put S after V. Steps of backpropagation don't change one parameter—they change potentially *all* of them based on the locally best direction (the gradient).

Second, these models, or learners, often don't need to pinpoint exactly one answer. A conjecture called the *lottery ticket hypothesis* holds that the behavior of a deep learning model tends to be determined by a relatively small number of its neurons (Frankle & Carbin 2018). Thus, the massive number of parameters is not because they all need to be set exactly to some value. Instead, having many degrees of freedom probably helps these models learn well by giving the models directions they can move in to avoid getting stuck. It may be like how it is easier to solve a puzzle if you can pick the pieces up and put them down (move them in a third dimension) rather than just sliding them around the table. More degrees of freedom can *help* learning.

Hierarchical structure need not be innate

Finally, these models discover structure—including hierarchical structure–from their training (Hewitt & Manning 2019, Manning et al. 2020, Lin et al. 2019, Tenney, Xia, et al. 2019, Warstadt & Bowman 2020, Kim et al. 2020, Linzen & Baroni 2021, Mahowald & Ivanova et al. 2023). These models certainly could learn rules based on linear, rather than hierarchical, structure, but the data strongly leads them towards hierarchical, structural generalization. This capacity to discover hierarchical structure rather than building it in was long-emphasized by cognitive psychologists, for instance work inducing syntactic categories by clustering (Redington et al. 1998, Clark 2000), or Perfors et al. 2011 who showed that a Bayesian model comparison applied to child-directed speech could identify a context-free (hierarchical) grammar over alternatives. The most recent models build on earlier efforts to encode hierarchical processes into neural networks (e.g. Tabor et al. 1997, Christiansen & Chater 1999, Tabor 2000, Smolensky & Legendre 2006).

The familiar example of aux-inversion is meant to show that children must possess hierarchical syntax. As a simple experiment, we can also ask these models to form some questions. For example, the familiar aux-inversion:

Convert the following into a single question asking if the accordion is in the rain: "The accordion that is being repaired is out in the rain."
Is the accordion that is being repaired out in the rain?

More rigorous experiments demonstrating aux-inversion success even in earlier language models can be found in Warstadt & Bowman 2020. It is philosophically interesting that this model knows which "is" to "move", but I think it's staggering that unlike pretty much every other model in the whole of science, we can interrogate these models simply by *asking* them to perform a certain behavior. It's also interesting to ask it to generate a few more examples, without giving guidance on the nature of the question:

Form the following sentence into a question in ten different ways: "The accordion that is being repaired is out in the rain."

Is the accordion that is being repaired out in the rain?
Where is the accordion that is being repaired located in the rain?
In what condition is the accordion being repaired out in the rain?
Is the accordion being repaired, which is outside, getting wet in the rain?
Is the accordion being repaired, which is out in the rain, being protected from the water?
Is the accordion being repaired out in the rain, and if so, will it be damaged by the water?
How long has the accordion that is being repaired been out in the rain?
Is the accordion, which is being repaired, being kept in the rain outside?
What is happening to the accordion, which is being repaired, while it is out in the rain?
Is the accordion that is currently being repaired exposed to the rain outside?

It is remarkable too that the models maintain enough representation of its prior discourse to generate distinct questions. The responses also highlight integration of syntax and semantics, for instance knowing that such an accordion might be "protected from the water" or "damaged by the water."

The aux-inversion cases have been interesting only insofar as interrogatives are derived from declaratives: the question of how kids know which "is" to move is nonsensical outside of that assumption. These models, which are trained only on text prediction, provide an implemented account in which we don't need to think of interrogatives as derived from declaratives (it seems very unlikely this is what is happening inside the model). The models might thus lead us to consider connections between other constructions, such as those used in the above list, which are likely located nearby to the target question in the model's latent space of activations. For optimally predictive theories, constructions may be connected to each other but not in the way standard theories of syntax predict.

Language and thought dissociate

Human language, for Chomsky, is deeply interconnected to human thought Everaert et al. 2015. Chomsky 2002 describes language as "a system for expressing thought" in fact one which is used primarily for speaking to oneself. Interestingly, he does not draw on the literature on inner monologues, which show substantial variation between individuals, with some describing virtually no internal language use at all (e.g. Reed 1916, Heavey & Hurlburt 2008, Roebuck & Lupyan 2020). Chomsky's view, though, is made perhaps more plausible by arguments that thought itself shares many properties of language, namely a compositional, language-like structure (Fodor 1975, Fodor & Pylyshyn 1988, Goodman et al. 2014, Piantadosi & Jacobs 2016, Quilty-Dunn et al. 2022). Chomsky frequently contrasts his inner thought view of language with the idea that language primarily is structured to support communication (e.g. Hockett 1959, Bates & MacWhinney 1982, Gibson et al. 2019), although it's worth noting he sometimes draws the opposite predictions from what efficient communication would actually predict (e.g. Piantadosi et al. 2012). Mahowald & Ivanova et al. 2023 argue in a comprehensive review that large language models exhibit a compelling dissociation between linguistic ability and thinking. The models know so much syntax, and aspects of semantics, but it is not hard to trip them up with appropriate logical reasoning tasks. Thus, large language models provide a proof of principle that syntax can exist and likely be acquired separately from other more robust forms of thinking and reasoning. Virtually all of the structure we see in language can come from learning a good model of strings, not directly modeling the world.

Models therefore show a logically possible dissociation between language and thinking. But a considerable amount of neuropsychological evidence supports the idea that language and thought are actually separate in people as well. Fedorenko & Varley 2016 review a large patient literature showing that patients with aphasia are often able to succeed in tasks requiring reasoning, logic, theory of mind, mathematics, music, navigation, and more. Aphasic patient studies provide an in vivo dissociation between language and other rational thinking processes. They also review neuroimaging work by Fedorenko et al. 2011 and others showing that the brain regions involved in language tend to be *specific to language* when it is compared to other non-linguistic tasks. That is not what would be predicted under Chomsky's account.

This is not to say that there is no way language and thought are related—we are able to specify some kinds of reasoning problems, communicate solutions, and sometimes solve problems with language itself. A compelling proposal is that language may be a system for connecting other core domains of representation and reasoning (Spelke 2003).

Why this and not that

Chomsky maintains (in the same Marcus interview above) that large language models have achieved nothing because they fail to explain "Why this? Why not that?" The question of whether these models can explain why human language has the form that it does is an interesting one that likely depends on whether the language system evolved before language or concurrently with it. If language co-opted neural systems for general sequential prediction (e.g. Christiansen & Chater 2015), it's possible we had some architecture like these models before we had language, and therefore the form of language *is* explained by the pre-existing computational architecture. On the other hand, if the two co-evolved, language might not be explained by the processing mechanisms. Given this uncertainty, we may admit that there are some "why" questions that a large language model may not answer; this does not mean they have no scientific value. In the same way, Newton's laws don't answer why those are the laws as opposed to any other, and yet they still embody deep scientific insights. Anyone who has had a child ask "why" repeatedly will recognize that at some point, *everyone*'s answers ground out in assumption.

However, it is worth highlighting in this context that Chomsky's own theories don't permit particularly deep "why" questions either. In large part, he simply states that the answer is genetics or simplicity or "perfection", without providing any independent justification for these claims. For example, readers of Berwick & Chomsky 2016—a book titled *Why Only Us*—might have hoped to find a thorough and satisfying "why" explanation. Their answer boils down to people having *merge* (essentially chunking two elements into one, unordered). And when it comes down to explaining *why* merge, they fall down the stairs: they simply state that "merge" is the minimal computational operation, apparently because that's what they think and that's that. Forget the relativity of definitions of simplicity, articulated by Goodman 1965, where what is considered simple must ground out in some convention. Berwick & Chomsky do even attempt to explain why they believe "merge" is simpler than other simple computational bases, like cellular automata or combinatory logic or systems of colliding Newtonian particles—all of which are capable of universal computation (and thus encoding structures, including hierarchical ones). Or maybe more directly, what makes *merge* "simpler" or more "perfect" than, say, backpropagation? Or Elman et al. 1996's architectural biases? Berwick & Chomsky don't even consider these questions, even though the ability to scientifically go after such "why" questions is supposed to be the hallmark of the approach. One might equally just declare that a transformer architecture is the "minimal" computational system that can handle the dependencies and structures of natural language and be done with it.

We should not actually take it for granted that generative syntax has found any regularities across languages that need "why" explanations. Evans & Levinson 2009 has made a convincing empirical case that prior features hypothesized to be universal—and thus plausibly part of the innate endowment of language—actually are not actually found in *all* languages. Perhaps most damningly, not even all languages appear to be recursive (Everett 2005, Futrell et al. 2016), contradicting *the* key universality claim from Hauser et al. 2002. Dąbrowska 2015 highlights profound differences in adult grammars between languages and the lack of a coherent formulation from generative linguists. None of this is to say that we won't be able to find any universals; rather, the proposed ones aren't there, and the differences between languages may be more scientifically informative than their commonalities (Pullum & Scholz 2009). On the methodological side, statistical analyses show that in order to justifiably claim something as universal, one would need on the order of 500 *statistically independent* languages, which is likely beyond what is currently in existence (Piantadosi & Gibson 2014).

However, the question of why languages are the way they are *does* have plausible, testable hypotheses. These hypotheses are interdisciplinary (Reali & Christiansen 2009) and include variegated influences of communicative (Zipf 1965, Hockett 1959, Bates & MacWhinney 1982, Piantadosi et al. 2012, Gibson et al. 2013, 2019, Coupé et al. 2019, Hahn et al. 2020, Futrell & Hahn 2022), cultural (Everett 2005, Lupyan & Dale 2010, Dale & Lupyan 2012, Everett et al. 2015), ecological (Lupyan & Dale 2016), learning (Steinert-Threlkeld & Szymanik 2019, 2020), and cognitive factors (Gibson 2000, Futrell et al. 2015). Such pressures have led to efficient and useful properties, including minimization of dependency structures (Futrell et al. 2015), the presence of communicatively useful ambiguity (Piantadosi et al. 2012), and efficiency in lexical systems (Kemp & Regier 2012, Kemp et al. 2018, Zaslavsky et al. 2019, Steinert-Threlkeld 2020, Mollica et al. 2021, Mahowald et al. 2022, Denić et al. 2022). Recent advances in understanding cultural evolution also should shape our theories of linguistic nativism, highlighting that weak biases are often sufficient over the course of cultural transmission to lead to stable patterns across languages (Thompson et al. 2016, Kirby et al. 2014, Chater et al. 2009). All of these are factors that Chomsky and others have never really grappled with, much less successfully ruled out as alternative answers to the "whys" of language.

The refutation of method

Chomsky often describes his own approach as "Galilean" meaning that he seeks the underlying principles in phenomena rather than analysis of large amounts of data. The term is both a misnomer (Behme 2014) and a not-so-subtle insult to colleagues who choose to work from different assumptions. Of course, Galileo cared about quantitative measurement of the world in order to formulate theories, developing tools of his own and even trying to measure the speed of light.[13] Chomsky's view was clearly articulated in an interview with Yarden Katz in 2012 where, at the time, he was focused on explaining that Bayesian models were useless:

> … [S]uppose that somebody says he wants to eliminate the physics department and do it the right way. The "right" way is to take endless numbers of videotapes of what's happening outside the [window], and feed them into the biggest and fastest computer, gigabytes of data, and do complex statistical analysis—you know, Bayesian this and that—and you'll get some kind of prediction about what's gonna happen outside the window next. In fact, you get a much better prediction than the physics department will ever give. Well, if success is defined as getting a fair approximation to a mass of chaotic unanalyzed data, then it's way better to do it this way than to do it the way the physicists do, you know, no thought experiments about frictionless planes and so on and so forth. But you won't get the kind of understanding that the sciences have always been aimed at—what you'll get at is an approximation to what's happening.

It's worth pinpointing exactly where this kind of thinking has gone wrong because it is central to the field's confusion in thinking about large language models. Chomsky's view certainly does not address the above idea that parameter fitting in a statistical model often *is* theory building and comparison.

But another factor is missing, too. Over modern scientific history, many computational scientists have noticed phenomena of *emergence* (Goldstein 1999), where the behavior of a system seems somewhat different than might be expected from its underlying rules. This idea has been examined specifically in language models (Wei et al. 2022, Manning et al. 2020), but the classic examples are older. The stock market is unpredictable even when individual traders might follow simple rules ("maximize profits"). Market booms and busts are the emergent result of millions of aggregate decisions. The high-level phenomena would be hard to intuit, even with full knowledge of traders' strategies or local goals. The field of complex systems has documented emergent phenomena virtually everywhere, from social dynamics to neurons to quasicrystals to honeybee group decisions. The field to have most directly grappled with emergence is physics, where it is acknowledged that physical systems can be understood on multiple levels of organization, and that the same laws that apply one one level (like molecular

[13] "Measure what can be measured, and make measurable what cannot be measured."

chemistry) may have consequences that are difficult to foresee on another (like protein folding) (Anderson 1972, Crutchfield 1994b,a).

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.

We can contrast understanding emergence through simulation with Chomsky's attempt to state principles and reason informally (see Pullum 1989) to their consequences. The result is pages and pages of stipulations and principles (see, e.g., Collins & Stabler 2016 or Chomsky 1995) that nobody could look at and conclude were justified through rigorous comparison to alternatives. After all they *weren't*: the failure of the method to compare vastly different sets of starting assumptions, including neural networks, is part of why modern large language models have taken everyone by surprise. The fact that after half a century of grammatical theories, there can be a novel approach which so completely blows generative grammar out of the water on every dimension is, itself, a refutation of the "Galilean method."

An effective research program into language would have considered, perhaps even developed, these kinds of models, and sought to compare principles like those of minimalism to the principles that govern neural networks. This turn of events highlights how much the dogma of being "Galilean" has counterproductively narrowed and restricted the space of theories under consideration—a salient irony given Chomsky's (appropriate) panning of Skinner for doing just that.[14]

[14] "What is so surprising is the particular limitations [Skinner] has imposed on the way in which the observables of behavior are to be studied." (Chomsky 1959)

Enduring contributions of Chomsky's program

I have attempted to convey my own sense of excitement about large language models, as well as my own pessimism about several aspects of Chomsky's approach to linguistics. However, it is easy to see that beyond the critiques above, many of Chomsky's emphatic focuses will survive his specific theories. For example, one of Chomsky's lasting contributions to cognitive science will be his emphasis on the reality of cognitive structure, like Tolman, Newell & Simon, Miller, and others of the cognitive revolution (Nadel & Piattelli-Palmarini 2003, Boden 2008). The search for the properties of human cognition that permit successful language acquisition is clearly central to understanding not just the functioning of the mind, but understanding humanity. It is a deep and important idea to try to characterize what computations are required for language, and to view them as genuinely mental computations. Chomsky's focus on children as creators of language, and of understanding the way in which their biases shape learning is fundamental to any scientific theory of cognition. Linguistic work in Chomsky's tradition has done a considerable amount to document and support less widely spoken languages, a struggle for current machine learning (Blasi et al. 2021). The overall search for "why" questions is undoubtedly core to the field, even as we reject or refine armchair hypotheses.

Some of the ideas of Chomsky's approach are likely to be found even in language models. For example, the idea that many languages are hierarchical is likely to be correct, embodied in some way in the connections and links of neural networks that perform well at word prediction. There may be a real sense in which other principles linguistics have considered are present in some form in such models. If the models correctly perform on binding questions, they may have some computations similar to binding principles. But none of these principles needed to be innate. And in neural networks, they are realized in a form nobody to date has written—they are distributed through a large pattern of continuous, gradient connections. Moreover, the representation of something like binding is extraordinarily unlikely to have the form generative syntax predicts since the required, underlying representational assumptions of that approach (e.g. binary branching, particular derivational structures, etc) are not met.

Another key contribution of Chomsky's research program has been to encourage discovery of interesting classes of sentences, often through others like Ross 1967. Regardless of the field's divergent views on the reality of WH-movement, for example, the question of *what* determines grammaticality and ungrammaticality for WH-sentences is an important one. Similarly, phenomena like "islands" do not go away because of large language models—they are targets to be explained (and they do a pretty good job according to analyses by Wilcox et al. 2022). Such phenomena are often difficult to separate from theory, as in the example above about whether declaratives and interrogatives are actually connected in the real grammar. Regardless of theory, researchers working in Chomsky's tradition have illuminated many places where human linguistic behavior is more complicated or intricate than one might otherwise expect.

As articulated by Pater 2019, the field should seek ways to integrate linguistics with modern machine learning, including neural networks. I have highlighted some researchers whose approach to language clearly resonates with the insights of modern language models. The current upheaval indicates that we should foster a pluralistic linguistics that approaches the problem of language with as few preconceptions as possible—perhaps even a fundamental reconceptualization of what language is for and what it is like (Edelman 2019). Maybe many of the "syntactic" phenomena that Chomskyan theories have concerned themselves with are really about something else, like pragmatics or memorized constructions (Goldberg 2006, Liu et al. 2022). Maybe the universals of language—if there are any—come from aspects of use like communicative and cognitive pressures, or other cultural factors. Maybe linguistics could learn from the methods of cognitive science (Edelman 2007). Maybe theories of grammar should respect humans' unparalleled memory capacity for sequential material. Maybe we should have linguistics students learn information theory, probability, neural networks, machine learning, anthropology, numerical methods, model comparison, Kolmogorov complexity, cognitive psychology, language processing, multi-agent systems, etc. The most permanent legacy of Chomsky's approach could as an admonishment about what happens when the study of language is separated from the rest of science.

Conclusion

One must be frank about the state of the art for models that capture syntax. It's not that large language models offer slightly higher performance than other approaches in linguistics; its not that they better cover some corners of syntax. It's that there is *nothing comparable in all of linguistic theory* to the power of large language models in both syntax and semantics—much less discourse coherence, style, pragmatics, translation, meta-linguistic awareness, non-linguistic tasks, etc. They are game changers on all fronts. Optimists who believe them as at least a plausible direction for acquisition see them as a way to build in and test architectural principles and biases, long emphasized by connectionist modelers like McClelland et al. 1986, Elman et al. 1996, Smolensky & Legendre 2006, and others. Those who doubt they could function as models of acquisition should nonetheless see the success of gradient representations, architectural assumptions, and implicit or emergent principles as a theory of grammar. These models have opened the space of plausible linguistic theories, allowing us to test principles beyond the ones that have traditionally concerned linguists. They allow us to finally develop compelling theories of the interplay of structure and statistics. And they appear to solve many problems that generative syntacticians have worried about, but without using any of their theoretical tools and constructs. Large language models rewrite the philosophy of approaches to language.

It's worth ending with a reminder of how impressive these models are:

Explain how large language models undermine Chomsky's main claims about linguistics.

Noam Chomsky is a prominent linguist and philosopher who has made significant contributions to the field of linguistics and language theory. He has put forward the idea that language is an innate, biologically determined ability that is unique to humans, and that all humans possess a universal grammar, or a set of innate linguistic rules that allow us to understand and produce language.

However, the development of large language models like GPT-3 has challenged some of Chomsky's main claims about linguistics and the nature of language.

First, the fact that language models can be trained on large amounts of text data and can generate human-like language without any explicit instruction on grammar or syntax suggests that language may not be as biologically determined as Chomsky has claimed. Instead, it suggests that language may be learned and developed through exposure to language and interactions with others.

Second, the success of large language models in performing various language tasks such as translation, summarization, and question answering, has challenged Chomsky's idea that language is based on a set of innate rules. Instead, it suggests that language is a learned and adaptive system that can be modeled and improved through machine learning algorithms.

Finally, the ability of language models to generate coherent and coherent [sic] language on a wide range of topics, despite never having seen these topics before, suggests that language may not be as rule-based as Chomsky has claimed. Instead, it may be more probabilistic and context-dependent, relying on patterns and associations learned from the text data it was trained on.

In conclusion, while Chomsky's contributions to the field of linguistics have been significant, the development of large language models has challenged some of his main claims and has opened up new avenues for exploring the nature of language and its relationship with machine learning and artificial intelligence.

Models that can do this are a gift. They make language the most exciting arena in all of cognitive science and AI. Such models are also a tool that will help linguistics to refine theories and compare leading ideas to strong alternatives.

It's worth remembering that Chomsky's approach to linguistics isn't so much about his particular beliefs—many linguists who work in his tradition disagree with him on specific points. And there is nothing wrong with being wrong: Chomsky's own theories elegantly articulated many deep, compelling ideas that have motivated linguistic and cognitive research. But modern language models highlight the weaknesses of his methodological prescriptions. One cannot seek principles without justifying them with rigorous experiments and comparisons; one can't reason about complex systems without implementations; one can't discover mechanisms without formally testing completely different approaches (like neural networks); one can't proclaim simplicity, optimality, or "perfection" without seriously connecting these terms to formal notions. These arguments worked on many linguists and consequently prevented Chomsky's tradition from developing anything close to the wild successes of modern language models.

Frederick Jelinek's quip "Every time I fire a linguist, the performance of the speech recognizer goes up" (Jelinek 1988) was a joke among linguists and computer scientists for decades. I've even seen it celebrated by academic linguists who think it elevates their abstract enterprise over and above the dirty details of implementation and engineering. But, while generative syntacticians insulated themselves from engineering, empirical tests, and formal comparisons, engineering took over. And now, engineering has solved the very problems the field has fixated on—or is about to very soon. The unmatched success of an approach based on probability, internalization of constructions in corpora, gradient methods, and neural networks is, in the end, a humiliation for everyone who has spent decades deriding these tools.

But now we can do better.

Acknowledgements

I am grateful to Benjamin Hayden, Ev Fedorenko, Geoffrey Pullum, Kyle Mahowald, Shimon Edelman, and Dan Everett, for detailed comments and suggestions on this paper. This paper benefited greatly from Felix Hill's Twitter presence (@FelixHill84), especially on topics like emergence and earlier connectionist work.

References

Adger, David. 2018. The autonomy of syntax. In *Syntactic Structures After 60 Years*, 153–175.

Ahmed, Nur, Muntasir Wahed & Neil C Thompson. 2023. The growing influence of industry in AI research. *Science* 379(6635). 884–886.

Anderson, Philip W. 1972. More is different: broken symmetry and the nature of the hierarchical structure of science. *Science* 177(4047). 393–396.

Baroni, Marco. 2022. On the proper role of linguistically oriented deep net analysis in linguistic theorising. In *Algebraic Structures in Natural Language*, 1–16. CRC Press.

Barrett, David, Felix Hill, Adam Santoro, Ari Morcos & Timothy Lillicrap. 2018. Measuring abstract reasoning in neural networks. In *International Conference on Machine Learning*, 511–520.

Bates, Elizabeth & Brian MacWhinney. 1982. Functionalist approaches to grammar. *Language acquisition: The state of the art*. 173–218.

Behme, Christina. 2012. A Potpourri of Chomskyan Science.

Behme, Christina. 2014. A 'Galilean' Science of Language. *Journal of Linguistics* 50(3). 671–704.

Belinkov, Yonatan & James Glass. 2019. Analysis methods in neural language processing: a survey. *Transactions of the Association for Computational Linguistics* 7. 49–72.

Bender, Emily M, Timnit Gebru, Angelina McMillan-Major & Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM conference on Fairness, Accountability, and Transparency*, 610–623.

Bender, Emily M & Alexander Koller. 2020. Climbing towards NLU: on meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198.

Berwick, Robert C & Noam Chomsky. 2016. *Why Only Us: Language and evolution*. MIT press.

Bhatia, Sudeep & Russell Richie. 2022. Transformer networks of human conceptual knowledge. *Psychological Review*.

Blasi, Damián, Antonios Anastasopoulos & Graham Neubig. 2021. Systematic Inequalities in Language Technology Performance across the World's Languages. *arXiv preprint arXiv:2110.06733*.

Block, Ned. 1986. Advertisement for a Semantics for Psychology. *Midwest Studies in Philosophy* 10. 615–678.

Block, Ned. 1998. Conceptual role semantics.

Bod, Rens, Remko Scha & Khalil Sima'an. 2003. *Data-oriented Parsing*. The University of Chicago Press.

Boden, Margaret A. 2008. *Mind as machine: A history of cognitive science*. Oxford University Press.

Bommasani, Rishi et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Bowman, Samuel, Christopher Potts & Christopher D Manning. 2015. Recursive neural networks can learn logical semantics. In *Proceedings of the 3rd workshop on continuous vector space models and their compositionality*, 12–21.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems* 33. 1877–1901.

Carnie, Andrew. 2021. *Syntax: A Generative Introduction*. John Wiley & Sons.

Caucheteux, Charlotte, Alexandre Gramfort & Jean-Rémi King. 2022. Deep language algorithms predict semantic comprehension from brain activity. *Scientific Reports* 12(1). 16327.

Chater, Nick, Florencia Reali & Morten H Christiansen. 2009. Restrictions on biological adaptation in language evolution. *Proceedings of the National Academy of Sciences* 106(4). 1015–1020.

Chater, Nick & Paul Vitányi. 2007. 'Ideal learning' of natural language: Positive results about learning from positive evidence. *Journal of Mathematical Psychology* 51(3). 135–163.

Chen, Stanley F & Joshua Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language* 13(4). 359–394.

Chi, Ethan A, John Hewitt & Christopher D Manning. 2020. Finding universal grammatical relations in multilingual BERT. *arXiv preprint arXiv:2005.04511*.

Chomsky, Noam. 1956. Three models for the description of language. *IRE Transactions on Information Theory* 2(3). 113–124.

Chomsky, Noam. 1957. Syntactic Structures.

Chomsky, Noam. 1959. *Chomsky, N. 1959. A review of BF Skinner's Verbal behavior. Language, 35 (1), 26–58.*

Chomsky, Noam. 1969. Quine's empirical assumptions. *Synthese* 19(1). 53–68.

Chomsky, Noam. 1995. *The Minimalist Program*. MIT Press.

Chomsky, Noam. 2002. *On Nature and Language*. Cambridge University Press.

Chomsky, Noam, Ian Roberts & Jeffrey Watumull. 2023. Noam Chomsky: The False Promise of ChatGPT. *The New York Times*. https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html.

Christiansen, Morten H & Nick Chater. 1999. Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science* 23(2). 157–205.

Christiansen, Morten H & Nick Chater. 2015. The language faculty that wasn't: A usage-based account of natural language recursion. *Frontiers in Psychology* 6. 1182.

Clark, Alexander. 2000. Inducing syntactic categories by context distribution clustering. In *Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*.

Clark, Alexander & Shalom Lappin. 2010. *Linguistic Nativism and the Poverty of the Stimulus*. John Wiley & Sons.

Collins, Chris & Edward Stabler. 2016. A formalization of minimalist syntax. *Syntax* 19(1). 43–78.

Collins, Katherine M, Catherine Wong, Jiahai Feng, Megan Wei & Joshua B Tenenbaum. 2022. Structured, flexible, and robust: Benchmarking and improving large language models towards more human-like behavior in out-of-distribution reasoning tasks. *arXiv preprint arXiv:2205.05718*.

Contreras Kallens, Pablo, Ross Deans Kristensen-McLachlan & Morten H Christiansen. 2023. Large language models demonstrate the potential of statistical learning in language. *Cognitive Science* 47(3). e13256.

Coupé, Christophe, Yoon Mi Oh, Dan Dediu & François Pellegrino. 2019. Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science Advances* 5(9). eaaw2594.

Cover, Thomas M. 1999. *Elements of Information Theory*. John Wiley & Sons.

Crain, Stephen & Paul Pietroski. 2001. Nature, nurture and universal grammar. *Linguistics and Philosophy* 24. 139–186.

Crain, Stephen & Paul Pietroski. 2002. Why language acquisition is a snap. *The Linguistic Review* 19(1). 163–183.

Croft, William. 1995. Autonomy and Functionalist Linguistics. *Language*. 490–532.

Crutchfield, James P. 1994a. Is anything ever new? Considering emergence. In *Santa Fe Institute Studies in the Sciences of Complexity*, vol. 19, 1–15.

Crutchfield, James P. 1994b. The calculi of emergence: computation, dynamics and induction. *Physica D: Nonlinear Phenomena* 75(1-3). 11–54.

Dąbrowska, Ewa. 2015. What exactly is Universal Grammar, and has anyone seen it? *Frontiers in Psychology* 6. 852.

Dale, Rick & Gary Lupyan. 2012. Understanding the origins of morphological diversity: The linguistic niche hypothesis. *Advances in Complex Systems* 15(03n04). 1150017.

Dasgupta, Ishita, Andrew K Lampinen, Stephanie CY Chan, Antonia Creswell, Dharshan Kumaran, James L McClelland & Felix Hill. 2022. Language models show human-like content effects on reasoning. *arXiv preprint arXiv:2207.07051*.

Delétang, Grégoire, Anian Ruoss, Jordi Grau-Moya, Tim Genewein, Li Kevin Wenliang, Elliot Catt, Marcus Hutter, Shane Legg & Pedro A Ortega. 2022. Neural networks and the Chomsky hierarchy. *arXiv preprint arXiv:2207.02098*.

Denić, Milica, Shane Steinert-Threlkeld & Jakub Szymanik. 2022. Indefinite Pronouns Optimize the Simplicity/Informativeness Trade-Off. *Cognitive Science* 46(5). e13142.

Edelman, Shimon. 2007. Bridging language with the rest of cognition. *Methods in Cognitive Linguistics*. 424–445.

Edelman, Shimon. 2019. Verbal behavior without syntactic structures: beyond Skinner and Chomsky. *arXiv preprint arXiv:4783474*.

Edelman, Shimon & Morten H Christiansen. 2003. How seriously should we take minimalist syntax? A comment on Lasnik. *Trends in Cognitive Science* 7(2). 60–61.

Edelman, Shimon & Heidi Waterfall. 2007. Behavioral and computational aspects of language and its acquisition. *Physics of Life Reviews* 4(4). 253–277.

Elman, Jeffrey L. 1990. Finding structure in time. *Cognitive Science* 14(2). 179–211.

Elman, Jeffrey L. 2004. An alternative view of the mental lexicon. *Trends in Cognitive Sciences* 8(7). 301–306.

Elman, Jeffrey L, Elizabeth A Bates & Mark H Johnson. 1996. *Rethinking innateness: A connectionist perspective on development*. MIT press.

Evans, Nicholas & Stephen C Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences* 32(5). 429–448.

Everaert, Martin BH, Marinus AC Huybregts, Noam Chomsky, Robert C Berwick & Johan J Bolhuis. 2015. Structures, not strings: Linguistics as part of the cognitive sciences. *Trends in Cognitive Sciences* 19(12). 729–743.

Everett, Caleb, Damián E Blasi & Seán G Roberts. 2015. Climate, vocal folds, and tonal languages: Connecting the physiological and geographic dots. *Proceedings of the National Academy of Sciences* 112(5). 1322–1327.

Everett, DanielL. 2005. Cultural constraints on grammar and cognition in Pirahã: Another look at the design features of human language. *Current Anthropology* 46(4). 621–646.

Fedorenko, Evelina, Michael K Behr & Nancy Kanwisher. 2011. Functional specificity for high-level linguistic processing in the human brain. *Proceedings of the National Academy of Sciences* 108(39). 16428–16433.

Fedorenko, Evelina & Rosemary Varley. 2016. Language and thought are not the same thing: Evidence from neuroimaging and neurological patients. *Annals of the New York Academy of Sciences* 1369(1). 132–153.

Fodor, Jerry A. 1975. *The Language of Thought.* Harvard University.

Fodor, Jerry A & Zenon W Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition* 28(1-2). 3–71.

Frankle, Jonathan & Michael Carbin. 2018. The lottery ticket hypothesis: finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635.*

Futrell, Richard & Michael Hahn. 2022. Information Theory as a Bridge Between Language Function and Language Form. *Motivations for Research on Linguistic Complexity: Methodology, Theory and Ideology.*

Futrell, Richard, Kyle Mahowald & Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences* 112(33). 10336–10341.

Futrell, Richard, Laura Stearns, Daniel L Everett, Steven T Piantadosi & Edward Gibson. 2016. A corpus investigation of syntactic embedding in Pirahã. *PLOS ONE* 11(3). e0145289.

Futrell, Richard, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros & Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. *arXiv preprint arXiv:1903.03260.*

Gauthier, Jon, Jennifer Hu, Ethan Wilcox, Peng Qian & Roger Levy. 2020. SyntaxGym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 70–76.

Ge, Ruifang & Raymond Mooney. 2005. A statistical semantic parser that integrates syntax and semantics. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, 9–16.

Geiping, Jonas & Tom Goldstein. 2022. Cramming: Training a Language Model on a Single GPU in One Day. *arXiv preprint arXiv:2212.14034.*

Gibson, E. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In *Image, language, brain: Papers from the first mind articulation project symposium*, 95–126.

Gibson, Edward, Leon Bergen & Steven T Piantadosi. 2013. Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences* 110(20). 8051–8056.

Gibson, Edward, Richard Futrell, Steven P Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen & Roger Levy. 2019. How efficiency shapes human language. *Trends in Cognitive Sciences* 23(5). 389–407.

Gold, E.M. 1967. Language identification in the limit. *Information and Control* 10(5). 447–474.

Goldberg, Adele E. 2003. Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences* 7(5). 219–224.

Goldberg, Adele E. 2006. *Constructions at Work*. Oxford University Press.

Goldberg, Adele E. 1995. *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Goldberg, Yoav. 2019. Assessing BERT's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

Goldstein, Ariel, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. 2022. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience* 25(3). 369–380.

Goldstein, Jeffrey. 1999. Emergence as a construct: History and issues. *Emergence* 1(1). 49–72.

Gonen, Hila, Shauli Ravfogel, Yanai Elazar & Yoav Goldberg. 2020. It's not Greek to mBERT: inducing word-level translations from multilingual BERT. *arXiv preprint arXiv:2010.08275*.

Goodman, Nelson. 1965. The new riddle of induction. In *Fact, Fiction, and Forecast*.

Goodman, Noah D, Joshua B Tenenbaum & Tobias Gerstenberg. 2014. *Concepts in a probabilistic language of thought*. Tech. rep. Center for Brains, Minds and Machines (CBMM).

Goodman, Noah D, Tomer D Ullman & Joshua B Tenenbaum. 2011. Learning a theory of causality. *Psychological Review* 118(1). 110.

Grand, Gabriel, Idan Asher Blank, Francisco Pereira & Evelina Fedorenko. 2022. Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature Human Behaviour* 6(7). 975–987.

Gulordava, Kristina, Piotr Bojanowski, Edouard Grave, Tal Linzen & Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. *arXiv preprint arXiv:1803.11138*.

Hahn, Michael, Dan Jurafsky & Richard Futrell. 2020. Universals of word order reflect optimization of grammars for efficient communication. *Proceedings of the National Academy of Sciences* 117(5). 2347–2353.

Harris, Randy Allen. 1993. *The Linguistics Wars*. Oxford University Press.

Hauser, Marc D, Noam Chomsky & W Tecumseh Fitch. 2002. The faculty of language: what is it, who has it, and how did it evolve? *Science* 298(5598). 1569–1579.

Heavey, Christopher L & Russell T Hurlburt. 2008. The phenomena of inner experience. *Consciousness and Cognition* 17(3). 798–810.

Hewitt, John & Christopher D Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4129–4138.

Hill, Felix, Kyunghyun Cho, Sébastien Jean & Yoshua Bengio. 2017. The representational geometry of word meanings acquired by neural machine translation models. *Machine Translation* 31. 3–18.

Hockett, Charles F. 1959. Animal "languages" and human language. *Human Biology* 31(1). 32–39.

Hoover, Jacob Louis, Morgan Sonderegger, Steven T Piantadosi & Timothy J O'Donnell. 2022. The plausibility of sampling as an algorithmic theory of sentence processing. *PsyArXiv*.

Hosseini, Eghbal A, Martin A Schrimpf, Yian Zhang, Samuel Bowman, Noga Zaslavsky & Evelina Fedorenko. 2022. Artificial neural network language models align neurally and behaviorally with humans even after a developmentally realistic amount of training. *bioRxiv*. 2022–10.

Hu, Jennifer, Jon Gauthier, Peng Qian, Ethan Wilcox & Roger P Levy. 2020. A systematic assessment of syntactic generalization in neural language models. *arXiv preprint arXiv:2005.03692*.

Hutter, Marcus. 2004. *Universal Artificial Intelligence: Sequential decisions based on algorithmic probability*. Springer.

Jackendoff, Ray. 2013. Constructions in the Parallel Architecture. In *The Oxford Handbook of Construction Grammar*.

Jelinek, Frederick. 1988. Applying information theoretic methods: Evaluation of grammar quality. In *Workshop on Evaluation of Natural Language Processing Systems, Wayne, PA*.

Johnson, K. 2004. Gold's theorem and cognitive science. *Philosophy of Science* 71(4). 571–592.

Kanwisher, Nancy, Meenakshi Khosla & Katharina Dobs. 2023. Using artificial neural networks to ask 'why' questions of minds and brains. *Trends in Neurosciences*.

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu & Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Kemp, Charles & Terry Regier. 2012. Kinship categories across languages reflect general communicative principles. *Science* 336(6084). 1049–1054.

Kemp, Charles & Joshua B Tenenbaum. 2008. The discovery of structural form. *Proceedings of the National Academy of Sciences* 105(31). 10687–10692.

Kemp, Charles, Yang Xu & Terry Regier. 2018. Semantic typology and efficient communication. *Annual Review of Linguistics* 4. 109–128.

Kim, Najoung, Roma Patel, Adam Poliak, Alex Wang, Patrick Xia, R Thomas McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, et al. 2019. Probing what different NLP tasks teach machines about function word comprehension. *arXiv preprint arXiv:1904.11544*.

Kim, Taeuk, Jihun Choi, Daniel Edmiston & Sang-goo Lee. 2020. Are pre-trained language models aware of phrases? Simple but strong baselines for grammar induction. *arXiv preprint arXiv:2002.00737*.

Kirby, Simon, Tom Griffiths & Kenny Smith. 2014. Iterated learning and the evolution of language. *Current Opinion in Neurobiology* 28. 108–114.

Kocijan, Vid, Ernest Davis, Thomas Lukasiewicz, Gary Marcus & Leora Morgenstern. 2022. The defeat of the Winograd schema challenge. *arXiv preprint arXiv:2201.02387*.

Kocijan, Vid, Thomas Lukasiewicz, Ernest Davis, Gary Marcus & Leora Morgenstern. 2020. A review of Winograd schema challenge datasets and approaches. *arXiv preprint arXiv:2004.13831*.

Koza, John R. 1994. Genetic programming as a means for programming computers by natural selection. *Statistics and Computing* 4. 87–112.

Kwiatkowski, Tom, Sharon Goldwater, Luke Zettlemoyer & Mark Steedman. 2012. A probabilistic model of syntactic and semantic acquisition from child-directed utterances and their meanings. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 234–244.

Lake, Brenden M & Gregory L Murphy. 2021. Word meaning in minds and machines. *Psychological Review*.

Lake, Brenden M, Ruslan Salakhutdinov & Joshua B Tenenbaum. 2015. Human-level concept learning through probabilistic program induction. *Science* 350(6266). 1332–1338.

Lake, Brenden M, Tomer D Ullman, Joshua B Tenenbaum & Samuel J Gershman. 2017. Building machines that learn and think like people. *Behavioral and Brain Sciences* 40. e253.

Langley, Pat, Gary L Bradshaw & Herbert A Simon. 1983. *Rediscovering chemistry with the BACON system*. Springer.

Lasnik, Howard. 2002. The minimalist program in syntax. *Trends in Cognitive Sciences* 6(10). 432–437.

Lasnik, Howard & Jeffrey Lidz. 2016. The argument from the poverty of the stimulus.

Laurence, Stephen & Eric Margolis. 2001. The poverty of the stimulus argument. *The British Journal for the Philosophy of Science* 52(2). 217–276.

LeCun, Yann, Yoshua Bengio & Geoffrey Hinton. 2015. Deep learning. *Nature* 521(7553). 436–444.

Legate, Julie Anne & Charles D Yang. 2002. Empirical re-assessment of stimulus poverty arguments. *The Linguistic Review* 19(1-2). 151–162.

Legg, Shane & Marcus Hutter. 2007. Universal intelligence: A definition of machine intelligence. *Minds and Machines* 17. 391–444.

Liang, P., M.I. Jordan & D. Klein. 2009. Learning semantic correspondences with less supervision. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, vol. 1, 91–99.

Lin, Yongjie, Yi Chern Tan & Robert Frank. 2019. Open Sesame: getting inside BERT's linguistic knowledge. *arXiv preprint arXiv:1906.01698*.

Linzen, Tal & Marco Baroni. 2021. Syntactic structure from deep learning. *Annual Review of Linguistics* 7. 195–212.

Liu, Nelson F, Matt Gardner, Yonatan Belinkov, Matthew E Peters & Noah A Smith. 2019. Linguistic knowledge and transferability of contextual representations. *arXiv preprint arXiv:1903.08855*.

Liu, Yingtong, Elodie Winckel, Anne Abeillé, Barbara Hemforth & Edward Gibson. 2022. Structural, functional, and processing perspectives on linguistic island effects. *Annual Review of Linguistics* 8. 495–525.

Lupyan, Gary & Rick Dale. 2010. Language structure is partly determined by social structure. *PLOS ONE* 5(1). e8559.

Lupyan, Gary & Rick Dale. 2016. Why are there different languages? The role of adaptation in linguistic diversity. *Trends in Cognitive Sciences* 20(9). 649–660.

Maas, Andrew, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng & Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, 142–150.

Mahowald, Kyle, Isabelle Dautriche, Mika Braginsky & Ted Gibson. 2022. Efficient communication and the organization of the lexicon. In *The Oxford Handbook of the Mental Lexicon*.

Mahowald & Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum & Evelina Fedorenko. 2023. Dissociating language and thought in large language models: A cognitive perspective. *arXiv preprint arXiv:2301.06627*.

Manning, Christopher & Hinrich Schutze. 1999. *Foundations of Statistical Natural Language Processing*. MIT press.

Manning, Christopher D, Kevin Clark, John Hewitt, Urvashi Khandelwal & Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences* 117(48). 30046–30054.

Marblestone, Adam H, Greg Wayne & Konrad P Kording. 2016. Toward an integration of deep learning and neuroscience. *Frontiers in Computational Neuroscience*. 94.

Marvin, Rebecca & Tal Linzen. 2018. Targeted syntactic evaluation of language models. *arXiv preprint arXiv:1808.09031*.

McCauley, Stewart M & Morten H Christiansen. 2019. Language learning as language use: A cross-linguistic model of child language development. *Psychological Review* 126(1). 1.

McClelland, James L, Felix Hill, Maja Rudolph, Jason Baldridge & Hinrich Schütze. 2020. Placing language in an integrated understanding system: Next steps toward human-level performance in neural language models. *Proceedings of the National Academy of Sciences* 117(42). 25966–25974.

McClelland, James L & Karalyn Patterson. 2002. Rules or connections in past-tense inflections: What does the evidence rule out? *Trends in cognitive sciences* 6(11). 465–472.

McClelland, James L, David E Rumelhart & PDP Research Group. 1986. *Parallel distributed processing*. MIT Press.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado & Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems* 26.

Mollica, Francis, Geoff Bacon, Noga Zaslavsky, Yang Xu, Terry Regier & Charles Kemp. 2021. The forms and meanings of grammatical markers support efficient communication. *Proceedings of the National Academy of Sciences* 118(49). e2025993118.

Mollica, Francis & Steven T Piantadosi. 2019. Humans store about 1.5 megabytes of information during language acquisition. *Royal Society open science* 6(3). 181393.

Murty, Shikhar, Pratyusha Sharma, Jacob Andreas & Christopher D Manning. 2022. Characterizing Intrinsic Compositionality In Transformers With Tree Projections. *arXiv preprint arXiv:2211.01288*.

Nadel, Lynn & Massimo Piattelli-Palmarini. 2003. What is Cognitive Science. In *Encyclopedia of cognitive science*.

Newmeyer, Frederick J. 1991. Functional explanation in linguistics and the origins of language. *Language & Communication* 11(1-2). 3–28.

Newport, Elissa L. 1990. Maturational Constraints on Language Learning. *Cognitive Science* 14(1). 11–28.

Norvig, Peter. 2012. Colorless green ideas learn furiously: Chomsky and the two cultures of statistical learning. *Significance* 9(4). 30–33.

Packard, Norman H, James P Crutchfield, J Doyne Farmer & Robert S Shaw. 1980. Geometry from a time series. *Physical Review Letters* 45(9). 712.

Papadimitriou, Isabel, Ethan A Chi, Richard Futrell & Kyle Mahowald. 2021. Deep subjecthood: Higher-order grammatical features in multilingual BERT. *arXiv preprint arXiv:2101.11043*.

Papadimitriou, Isabel & Dan Jurafsky. 2020. Learning music helps you read: Using transfer to study linguistic structure in language models. *arXiv preprint arXiv:2004.14601*.

Patel, Roma & Ellie Pavlick. 2022. Mapping language models to grounded conceptual spaces. In *International Conference on Learning Representations*.

Pater, Joe. 2019. Generative linguistics and neural networks at 60: Foundation, friction, and fusion. *Language* 95(1). e41–e74.

Pavlick, Ellie. 2022. Semantic structure in deep learning. *Annual Review of Linguistics* 8. 447–471.

Pearl, Lisa. 2022. Poverty of the stimulus without tears. *Language Learning and Development* 18(4). 415–454.

Pereira, Fernando. 2000. Formal grammar and information theory: together again? *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* 358(1769). 1239–1253.

Perfors, Amy, Joshua B Tenenbaum & Terry Regier. 2011. The learnability of abstract syntactic principles. *Cognition* 118(3). 306–338.

Petersen, Erika & Christopher Potts. 2022. Lexical semantics with large language models: a case study of english break.

Piantadosi, Steven T & Edward Gibson. 2014. Quantitative standards for absolute linguistic universals. *Cognitive Science* 38(4). 736–756.

Piantadosi, Steven T & Felix Hill. 2022. Meaning without reference in large language models. *arXiv preprint arXiv:2208.02957*.

Piantadosi, Steven T & Robert A Jacobs. 2016. Four problems solved by the probabilistic language of thought. *Current Directions in Psychological Science* 25(1). 54–59.

Piantadosi, Steven T, Harry Tily & Edward Gibson. 2012. The communicative function of ambiguity in language. *Cognition* 122(3). 280–291.

Pinker, Steven & Alan Prince. 1988. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition* 28(1-2). 73–193.

Pires, Telmo, Eva Schlinger & Dan Garrette. 2019. How multilingual is multilingual BERT? *arXiv preprint arXiv:1906.01502.*

Postal, Paul. 2012. Two case studies of Chomsky's play acting at linguistics. *Lingbuzz.*

Prince, Alan & Paul Smolensky. 1997. Optimality: From neural networks to universal grammar. *Science* 275(5306). 1604–1610.

Pullum, Geoffrey K. 1989. Formal linguistics meets the boojum. *Natural language & Linguistic theory*. 137–143.

Pullum, Geoffrey K. 2007. The evolution of model-theoretic frameworks in linguistics. *Model-Theoretic Syntax at 10*. 1–10.

Pullum, Geoffrey K. 2013. The central question in comparative syntactic metatheory. *Mind & Language* 28(4). 492–521.

Pullum, Geoffrey K & Barbara C Scholz. 2002. Empirical assessment of stimulus poverty arguments. *The Linguistic Review* 19(1-2). 9–50.

Pullum, Geoffrey K & Barbara C Scholz. 2009. For universals (but not finite-state learning) visit the zoo. *Behavioral and Brain Sciences* 32(5). 466–467.

Quilty-Dunn, Jake, Nicolas Porot & Eric Mandelbaum. 2022. The best game in town: The re-emergence of the language of thought hypothesis across the cognitive sciences. *Behavioral and Brain Sciences*. 1–55.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1(8). 9.

Reali, Florencia & Morten H Christiansen. 2005. Uncovering the richness of the stimulus: Structure dependence and indirect statistical evidence. *Cognitive Science* 29(6). 1007–1028.

Reali, Florencia & Morten H Christiansen. 2009. On the necessity of an interdisciplinary approach to language universals. *Language Universals*. 266–77.

Redington, Martin, Nick Chater & Steven Finch. 1998. Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science* 22(4). 425–469.

Reed, Homer B. 1916. The existence and function of inner speech in thought processes. *Journal of Experimental Psychology* 1(5). 365.

Richards, Blake A, Timothy P Lillicrap, Philippe Beaudoin, Yoshua Bengio, Rafal Bogacz, Amelia Christensen, Claudia Clopath, Rui Ponte Costa, Archy de Berker, Surya Ganguli, et al. 2019. A deep learning framework for neuroscience. *Nature Neuroscience* 22(11). 1761–1770.

Roebuck, Hettie & Gary Lupyan. 2020. The internal representations questionnaire: Measuring modes of thinking. *Behavior Research Methods* 52. 2053–2070.

Rogers, Anna, Olga Kovaleva & Anna Rumshisky. 2021. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics* 8. 842–866.

Rogers, Timothy T, James L McClelland, et al. 2004. *Semantic cognition: A parallel distributed processing approach.* MIT press.

Ross, John Robert. 1967. *Constraints on variables in syntax.* (Doctoral dissertation).

Rule, Joshua S, Joshua B Tenenbaum & Steven T Piantadosi. 2020. The Child as Hacker. *Trends in Cognitive Sciences* 24(11). 900–915.

Rumelhart, David E & James L McClelland. 1986. On learning the past tenses of English verbs. In *Parallel Distributed Processing, volume 2.*

Schmidt, Michael & Hod Lipson. 2009. Distilling free-form natural laws from experimental data. *Science* 324(5923). 81–85.

Schrimpf, Martin, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum & Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences* 118(45). e2105646118.

Schrimpf, Martin, Jonas Kubilius, Michael J Lee, N Apurva Ratan Murty, Robert Ajemian & James J DiCarlo. 2020. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron* 108(3). 413–423.

Shain, Cory, Clara Meister, Tiago Pimentel, Ryan Cotterell & Roger Philip Levy. 2022. Large-scale evidence for logarithmic effects of word predictability on reading time. *PsyArXiv*.

Shannon, Claude E. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27(3). 379–423.

Siegelmann, Hava T & Eduardo D Sontag. 1995. On the computational power of neural nets. *Journal of Computer and System Sciences* 50(1). 132–150.

Siskind, J.M. 1996. A Computational Study of Cross-Situational Techniques for Learning Word-to-Meaning Mappings. *Cognition* 61. 31–91.

Smolensky, Paul & Géraldine Legendre. 2006. *The Harmonic Mind: From neural computation to optimality-theoretic grammar.* MIT press.

Socher, Richard, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng & Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1631–1642.

Solan, Zach, David Horn, Eytan Ruppin & Shimon Edelman. 2005. Unsupervised learning of natural languages. *Proceedings of the National Academy of Sciences* 102(33). 11629–11634.

Solomonoff, Ray J. 1964. A formal theory of inductive inference. Part I. *Information and Control* 7(1). 1–22.

Spelke, Elizabeth S. 2003. What makes us smart? Core knowledge and natural language. *Language in mind: Advances in the study of language and thought.* 277–311.

Steedman, Mark. 2001. *The Syntactic Process.* Cambridge MA: MIT Press.

Steinert-Threlkeld, Shane. 2020. Quantifiers in natural language optimize the simplicity/informativeness trade-off. In *Proceedings of the 22nd Amsterdam Colloquium*, 513–522.

Steinert-Threlkeld, Shane & Jakub Szymanik. 2019. Learnability and semantic universals. *Semantics and Pragmatics* 12. 1–39.

Steinert-Threlkeld, Shane & Jakub Szymanik. 2020. Ease of learning explains semantic universals. *Cognition* 195. 104076.

Tabor, Whitney. 2000. Fractal encoding of context-free grammars in connectionist networks. *Expert Systems* 17(1). 41–56.

Tabor, Whitney, Cornell Juliano & Michael K Tanenhaus. 1997. Parsing in a dynamical system: An attractor-based account of the interaction of lexical and structural constraints in sentence processing. *Language and Cognitive Processes* 12(2-3). 211–271.

Takens, Floris. 1981. Detecting strange attractors in turbulence. In *Dynamical systems and Turbulence*, 366–381. Springer.

Tenenbaum, Joshua B, Charles Kemp, Thomas L Griffiths & Noah D Goodman. 2011. How to grow a mind: Statistics, structure, and abstraction. *Science* 331(6022). 1279–1285.

Tenney, Ian, Dipanjan Das & Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. *arXiv preprint arXiv:1905.05950.*

Tenney, Ian, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316.*

Thompson, Bill, Simon Kirby & Kenny Smith. 2016. Culture shapes the evolution of cognition. *Proceedings of the National Academy of Sciences* 113(16). 4530–4535.

Tomasello, Michael. 2000. The item-based nature of children's early syntactic development. *Trends in Cognitive Sciences* 4(4). 156–163.

Tomasello, Michael. 2005. *Constructing a language: A usage-based theory of language acquisition*. Harvard university press.

Turing, AM. 1950. Computing Machinery and Intelligence.

Udrescu, Silviu-Marian & Max Tegmark. 2020. AI Feynman: A physics-inspired method for symbolic regression. *Science Advances* 6(16). eaay2631.

Ullman, Tomer D, Noah D Goodman & Joshua B Tenenbaum. 2012. Theory learning as stochastic search in the language of thought. *Cognitive Development* 27(4). 455–480.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser & Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* 30.

Warstadt, Alex & Samuel R Bowman. 2020. Can neural networks acquire a structural bias from raw linguistic data? *arXiv preprint arXiv:2007.06761*.

Warstadt, Alex & Samuel R Bowman. 2022. What artificial neural networks can tell us about human language acquisition. In *Algebraic Structures in Natural Language*, 17–60. CRC Press.

Warstadt, Alex, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, et al. 2019. Investigating BERT's knowledge of language: Five analysis methods with NPIs. *arXiv preprint arXiv:1909.02597*.

Warstadt, Alex, Leshem Choshen, Aaron Mueller, Adina Williams, Ethan Wilcox & Chengxu Zhuang. 2023. Call for Papers–The BabyLM Challenge: Sample-efficient pretraining on a developmentally plausible corpus. *arXiv preprint arXiv:2301.11796*.

Warstadt, Alex, Amanpreet Singh & Samuel R Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics* 7. 625–641.

Wei, Jason, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Weissweiler, Leonie, Taiqi He, Naoki Otani, David R Mortensen, Lori Levin & Hinrich Schütze. 2023. Construction grammar provides unique insight into neural language models. *arXiv preprint arXiv:2302.02178*.

Wexler, Kenneth & Peter W Culicover. 1980. *Formal Principles of Language Acquisition*, vol. 76. MIT Press.

Wilcox, Ethan, Roger Levy, Takashi Morita & Richard Futrell. 2018. What do RNN language models learn about filler-gap dependencies? *arXiv preprint arXiv:1809.00042*.

Wilcox, Ethan Gotlieb, Richard Futrell & Roger Levy. 2022. Using computational models to test syntactic learnability. *Linguistic Inquiry*. 1–88.

Yang, Yuan & Steven T Piantadosi. 2022. One model for the learning of language. *Proceedings of the National Academy of Sciences* 119(5). e2021865119.

Ye, Hao & George Sugihara. 2016. Information leverage in interconnected ecosystems: Overcoming the curse of dimensionality. *Science* 353(6302). 922–925.

Zaslavsky, Noga, Terry Regier, Naftali Tishby & Charles Kemp. 2019. Semantic categories of artifacts and animals reflect efficient coding. *arXiv preprint arXiv:1905.04562*.

Zhang, Chiyuan, Samy Bengio, Moritz Hardt, Benjamin Recht & Oriol Vinyals. 2021. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM* 64(3). 107–115.

Zhang, Yian, Alex Warstadt, Haau-Sing Li & Samuel R Bowman. 2020. When do you need billions of words of pretraining data? *arXiv preprint arXiv:2011.04946*.

Zipf, George Kingsley. 1965. *The psycho-biology of language: An introduction to dynamic philology*. MIT Press.