DOI: 10.26346/1120-2726-240

Large Language Models and Minimalism: Theories, grammars, and computational modeling

Jason Ginsburg

Graduate School of Human and Environmental Studies, Kyoto University, Japan
<ginsburg.jasonrobert.2h@kyoto-u.ac.jp>

Chesi (*this issue*), responding to Piantadosi (2024), compares Large Language Models with work in Generative Linguistics. Chesi points out that syntactic tests exist for Large Language Models, but not for theories of Generative Linguistics, as well as that theories in Generative Linguistics lack proper formalization, which has led to Generative Linguistics becoming marginalized. In this paper, I take the position that Generative Linguistics develops theories of language, but Large Language Models are not theories. Also, while syntactic tests that examine the validity and scope of theories in Generative Linguistics could be useful, a number of large hurdles exist for their development.

KEYWORDS: large language models, theories, grammar, Minimalism, computer modeling.

1. Introduction

Large Language Models (LLMs) have been the focus of a great deal of attention in the media and in academia. This is most likely because they appear to be a huge leap forward in Artificial Intelligence. They can be used to create seemingly authentic language, and they may be able to pass some versions of the Turing Test (Turing 1950).[1] These models potentially have a wide variety of applications. Some applications may be helpful; e.g. responding to patients for health care (Ayers *et al.* 2023), helping with language learning (Klimova *et al.* 2024), increasing productivity on workplace writing tasks (Noy & Zhang 2023), etc. Some applications/ aspects of these models may be harmful; e.g. replacement of human jobs (Demirci *et al.* 2024), increase of plagiarism in academia and elsewhere (Chen *et al.* 2024; Kwon 2024), spread of authentic-looking misinformation (Raman *et al.* 2024; Spitale *et al.* 2023), use to commit fraud and deceive people (Park *et al.* 2024), misdiagnose illnesses (Barile *et al.* 2024), cause damage to the environment (Crawford 2024), etc.

Chesi (*this issue*), responding to a paper by Piantadosi (2024) that skewers Generative Linguistics, argues that Piantadosi is partially correct with respect to criticism of Generative Linguistics. In this paper, I focus on the following topics from Chesi's paper.[2] (Below, I use

(received December 2024)

Minimalist theories to refer to work in Minimalism generally following Chomsky (1995), which is the latest incarnation of work within the Generative Grammar framework, following a long tradition led by Noam Chomsky himself.)

Several claims from Chesi's paper:

A) LLMs are more observationally adequate than Minimalist theories, but they "lack explanatory adequacy" (p. 32). Compared to LLMs, Minimalist theories, on the other hand, may be superior with respect to descriptive adequacy.

B) Minimalist theories are unable to perform adequately in "comprehensive syntactic tests" (p. 5) and it is necessary to "adopt a modern approach to theory evaluation that relies on shared datasets and metrics" (p. 6). Furthermore, there is no "shared test/reference set" (p. 18) for Minimalist theories.

C) A problem for developing linguistic benchmarks for Minimalist theories and for testing how well Minimalist theories do on benchmarks is that Minimalism has not been properly formalized.

D) The lack of comprehensive linguistic benchmarks and the lack of a fully explicit theory has led to Minimalism becoming marginalized.

I attempt to address these issues in the following sections. In section 2, I address the issue of whether or not LLMs can be called theories and/or grammars, and how they compare with Minimalist theories. In section 3, I examine the issue of whether or not it is possible to create benchmarks for Minimalist theories.

2. LLMs as theories and grammars

Some researchers refer to LLMs as theories of language. Baroni (2023: 1) writes that "deep networks should be treated as theories making explicit predictions about the acceptability of linguistic utterances." Baroni further writes

> we can think of a deep net architecture, before any language-specific training, as a general theory defining a space of possible grammars, and of the same network trained on data from a specific language as a *grammar*, that is, a computational system that, given an input utterance in a language can predict whether the sequence is acceptable to an idealized speaker of the language. (Baroni 2023: 7)

Piantadosi (2024: 360) writes that LLMs "develop representations of key structures and dependencies", and that "[a]s argued by Baroni (2022), this means that language models should be treated as bona fide linguistic *theories*." He argues that "a space of possible theories is parameterized by the models and compared to data to find which theory is best in a formal sense (Baroni 2022: 360)."

A theory provides an explanation (typically a hypothesis) about some fact(s) of the natural world. Linguistic theories attempt to provide explanations for how language works. An LLM is a computer model that produces language output that is similar (although not identical) to the language input (its training data). Specifically, an LLM is a type of Generative AI which produces output that is similar to the input, and Generative AI can be used for tasks other than language. While there may be terminological issues regarding definitions of theories, I think that there are flaws to viewing an LLM as a theory. An LLM, as well as other neural net models, are not theories in the sense that they do not provide explanations for facts of the world. A computer model, such as an LLM, can be used to implement, test, and/or develop a theory, but an LLM is not a theory.

LLMs are engineering tools that produce language and there are theories behind the designs of LLMs, but these are not theories of language. The engineering methods used to construct LLMs do not take into account the structures of language. Rather, LLMs are designed to model the training data in optimal ways.

Assume, as Baroni and Piantadosi suggest, that an LLM produces language in a way that is similar to a human. Then the question arises of how it produces and 'understands' language. While it is possible to explain how an LLM is built, it is not clear exactly how the various parameters within an LLM are set to produce a given output. The inner workings of an LLM are often referred to as a black box, in that we do not know exactly what is going on within them (Dobson 2023, Fazi 2020, etc.). They can have billions of parameters that are connected in ways that we do not fully understand. Fazi (2021: 59) writes

> once a deep neural network is trained (or self-trained…), it can be extremely difficult to explain why it gives a particular response to some data inputs and how a result has been calculated. The strength of a deep neural network lies in its capacity to find non-linear patterns in large datasets and improve this extraction through iterative interactions.

An LLM finds patterns, but we do not necessarily know how it comes up with these patterns, and even if we were to develop theories to explain these patterns, these patterns do not necessarily correspond to how humans use language.

Fox & Katzier (2024: 72) write “We might be impressed by an LLM generating a Shakespearean sonnet or by LLM activity correlating with data from brain imaging, but unless these observations bear on theory selection, they are not going to tell us much about underlying machinery.” Linguists are interested in the underlying machinery of language. Looking at the output of an LLM is like looking at a corpus of language produced by humans (although there are likely some differences). Just because an LLM produces language that is generally identical to human language does not necessarily tell us anything unique about how language works.

Theories can be developed based on LLMs (and other neural networks). For example, Manning *et al.* (2020) examine how self-supervised neural nets represent linguistic structure. Lakretz *et al.* (2021) investigate how language models represent agreement. These works develop theories that attempt to account for how neural-net language models represent aspects of language. These theories may be useful for understanding aspects of the inner-workings of LLMs (and other types of similar models), and they might also be of use in building better LLMs in the future. If LLMs work in the same ways that humans brains do, then these theories could be useful for explaining how language works. However, it is not at all clear if LLMs work in the same way that human language does. Katzir (2023: 2) writes “Since LLMs were designed to be useful engineering tools, discovering that they teach us about how humans work would be startling indeed, akin to discovering that a newly designed drone accidentally solves an open problem in avian flight.”

LLM training is quite different from the human situation. LLMs are exposed to much more data than a human child is exposed to. As Piandadosi (2024: 354) notes, they are “trained on huge datasets of internet-based text to predict upcoming linguistic material.” According to Piantadosi (2024: 354), “a typical language model might be trained on hundreds of billions of tokens, estimated to cost millions of dollars in energy alone (Piantadosi 2024: 354).” Estimates are that children hear between 2 million to 11 million words per year (Warstadt & Bowman 2022, Hart & Risley 1992, Gilkerson *et al.* 2017). GPT-3 was reportedly trained on over 200 billion words (Brown *et al.* 2020, Warstadt & Bowman 2022). Piandadosi (2024: 357) writes that LLMs “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.” I note that I have never spoken with a child who has memorized much of the Internet. Also, as Piandadosi (2024: 358) notes, these models “are trained only on text prediction.” Humans are not trained on text prediction.

Presumably, due to their ability to generally produce correct language, Chesi points out that "the computational perspective appears to be leading in terms of observational adequacy, and possibly in terms of descriptive adequacy as well" (p. 19). Furthermore, compared with Minimalist theories, Chesi writes that LLMs "are observationally more adequate but lack explanatory adequacy" (p. 32). LLMs clearly lack explanatory adequacy, but I also think they lack observational adequacy and descriptive adequacy.

Chomsky (1964: 63) writes that "[a] grammar that aims for observational adequacy is concerned merely to give an account of the primary data (e.g. the corpus) that is the input to the learning device." If an LLM could be considered a grammar, then it might be accurate to consider it to be more observationally adequate than Minimalist theories. However, I do not think that an LLM can be considered a grammar that corresponds to grammar used by humans for language. Chomsky (1965: 6) writes that "by a generative grammar I mean simply a system of rules that in some explicit and well-defined way assigns structural descriptions to sentences." Carnie (2021: 27) defines a grammar as "the set of mental rules or procedures that generate a sentence." Radford (1997: 4) writes that "a grammar of a language is a model of the grammatical competence of the fluent native speaker of the language." Given these definitions, ChatGPT (or other LLMS) is not a grammar, or at least not a human language grammar. There is no clear evidence that it assigns structural descriptions to sentences, nor that it has components that directly correspond to the mental rules or procedures that humans use to generate (and parse) sentences. There is no evidence that it is an accurate model of the grammatical competence of a human, in the sense that, as far as we know, it does not produce sentences in the same way that a human does.

Regarding descriptive adequacy, Chomsky (1964: 63) writes that "a grammar that aims for descriptive adequacy is concerned to give a correct account of the linguistic intuition of the native speaker." If you look at the output of an LLM, it produces grammatically correct sentences. On the other hand, it is not clear if it has any intuitions. You can ask an LLM questions about grammar, and it might be able to answer correctly. But these answers result from patterns found in the training data. For example, I asked ChatGPT to discuss the grammaticality of the sentences in Table 1, which require that the main verb agree with the head noun of the subject, even though the head noun of the subject is not the closest noun to the main verb. The answer in Table 1 is correct. However, this correctness does not show that ChatGPT actually has a notion of subject-verb agreement. Instead, ChatGPT was able to come up with an appropriate answer due to what it ‘learned’ from patterns its training data, not from some internal grammar.

Discuss the grammaticality of these sentences.
a. The girl who likes her school friends studies hard.
b. The girl who likes her school friends study hard.

Summary
· Sentence (a) is grammatical because it correctly uses the singular verb *studies* with the singular subject *the girl*.
· Sentence (b) is ungrammatical because it incorrectly uses the base form *study* with the singular subject *the girl*.

Table 1. Sample discussion (abbreviated) with ChatGPT.

3. Comprehensive syntactic tests (benchmarks) for Minimalist theories

Chesi argues that a problem for Minimalist theories is an inability “to perform adequately in similarly complete and extensive benchmarks as the ones presented in SyntaxGym” (p. 6). SyntaxGym (<syntaxgym.org>, Hu *et al.* 2020) is an online platform that enables the performance of different computational language models to be compared. Having syntactic benchmarks that could be used to evaluate Minimalist theories could potentially be beneficial, but complex issues arise.

Note that the basic types of phenomena that Hu *et al.* (2020) compare with SyntaxGym come from research in generative linguistics. They write that they chose 16 “[o]f the 47 empirical phenomena reviewed in the summary sections at the end of each chapter (Hu *et al.*: 1727)” of the introductory syntax textbook Carnie (2013).[3] Some of these types of constructions are given in Table 2.

	Construction types	Examples
(a)	Agreement	The author that the senators hurt is/*are good.
(b)	Center Embedding	The painting that the artist painted deteriorated. *The painting that the artist deteriorated painted.
(c)	Garden path	The woman brought the sandwich from the kitchen fell in the dining room.
(d)	Subordination	*As the doctor studied the book. As the doctor studied the book, the nurse walked into the room.

(e)	Negative polarity licensing	No author that the senators liked has had any success.
(f)	Long distance dependencies (pseudo relative clause)	What the young man planted was the crops.

Table 2. Types of examples found in SyntaxGym.

Although Chesi correctly points out that the performance of Minimalist theories are not tested with benchmarks such as those given in Table 2, it is notable that all of these types of constructions can be accounted for with Minimalist theories. These types of constructions are taken from a syntax textbook, and thus it is not surprising that accounts of these types of constructions can be found in textbooks such as Radford (2016) and Carnie (2021). There is also a wide body of literature in the Generative linguistics literature (not necessarily just Minimalism) regarding these types of constructions.

Although Minimalist theories can likely account for the various constructions in SyntaxGym, there is no one agreed-upon Minimalist theory that accounts for all of these types of constructions, which is a valid point that I think Chesi is making. The question then arises of whether or not it is possible to test Minimalist theories on these types of phenomena so that they can be compared with computer models of language such as LLMs. I think that it may be possible, but there are several complicating issues.

It is possible to create a computer model based on Minimalist theories and to test the model's ability to parse/generate target syntactic constructions. However, I do not think that it is currently possible to create a model of a single over-arching Minimalist theory that all (or most) researchers working in the Minimalist framework would agree on.

There are at least two approaches that I know of to computational modeling of Minimalist theories. One is the Minimalist Grammar approach of Stabler (1997, 2011) and related work.[4] The other approach, which I have been directly involved with, is attempts to model the latest Minimalist theories with computer programs.

Beginning at least with Fong's (1991) Government and Binding Theory-based parser, there has been research attempting to computationally model theories of Generative Linguistics. Some representative works are the following. Fong & Ginsburg (2012) models constructions with pronouns and antecedents. Fong & Ginsburg (2014) models *tough*-constructions. Ginsburg (2016) models basic statements and *wh*-questions from the perspective of Labeling theory (Chomsky 2013). Fong & Ginsburg

(2019) discusses the architecture of a computational model based on Phase Theory (Chomsky 2001), and Ginsburg & Fong (2019) discusses how this single model accounts for a variety of basic syntactic phenomena and constructions including multiple agreement, constructions with expletives, thematization/extraction, the *that*-trace effect, subject *vs* object *wh*-movement, and relative clauses. Fong & Ginsburg (2023) presents a model that accounts for a wide-variety of English relative clause constructions. Ginsburg (2024) presents a model that permits Merge operations to apply with a limited amount of freedom to account for basic statements, control constructions, *wh*-questions, and yes/no-questions. These types of Minimalist theory-based computational models are possible, and their scope can be extended, but they face several hurdles.

One important issue for these models is that linguists working on Minimalist theories do not agree on one particular Minimalist theory. Among linguists, there are often varying views about particular phenomena, as well as about the core nature of the Minimalist program. This means that no matter which theory is implemented, there will be researchers who may not be pleased. For example, Ginsburg (2024) discusses a computational implementation of some of the latest work in Minimalism, in particular following recent proposals in Chomsky (2001, 2013, 2015, 2021b, 2024). This work is recent, and the majority of linguistics papers in linguistics journals do not make use of the latest notions from Minimalist theories.[5] In order to construct the computer model described in Ginsburg (2024), I had to make assumptions about a variety of controversial topics (see Ginsburg 2024 for references). Without making specific assumptions, the model would not have been able to sufficiently implement a Minimalist theory. For example, a large body of recent work in Minimalism assumes that features on T are inherited from C. But due to the complexity of feature inheritance, and what I see as a lack of strong evidence for a complex feature inheritance operation, I did not implement feature inheritance. Head movement has been the subject of numerous analyses. Some work argues that head movement is problematic and should not exist in the syntax, whereas a large body of work assumes that it occurs in the syntax. Of the work that makes use of syntactic head movement, there are a variety of differing proposals ranging from adjunction of one head with a higher head, internal pair-Merge of a head with a higher head, and movement of a head to a higher specifier position. There are also accounts that posit that some head movement occurs in the syntax and some head movement occurs post-syntactically, as well as accounts for purely post-syntactic head-movement. Following the latest work in Chomsky (2021b, 2024), I assumed that head-movement is a post-syntactic operation.

Furthermore, Chomsky (2021b) proposes a novel FormCopy relation that accounts for how two NPs can be given the same reference. In more recent work, Chomsky (2024) develops a new account of *wh*-movement and focus-movement effects. I incorporated these latest theories into my model, but these theories are not necessarily accepted (or even well-known) by the mainstream community working on Minimalist theories.

Chesi argues that Minimalist theories have not been properly formalized and that a fully explicit theory is lacking, writing that “I think the original sin of most generative linguists is that they have gotten used to incomplete pseudo-formalizations and data fragment explanations” (p. 40). I think that this is one reason why creating computer models for Minimalist theories is useful, as well as a problem. In an ideal world, there would be one Minimalist theory generally agreed upon by linguistics researchers, and this theory could be implemented. But that currently is not the situation. The purpose of linguistic research is to understand how language works. If linguists currently have not come to a consensus about how to account for language, then that reflects the state of our understanding. The development of theories that truly explain how language works could help lead to a more unified theory, and computational models, which can demonstrate how well a single theory accounts for a variety of diverse syntactic phenomena, can be useful for this purpose.

Another problem for development of computational models of Minimalist theories is the technical skills that are required. Most computer scientists are not well-versed in linguistics and most linguists are not well-versed in computer programming. Theoretical syntacticians typically develop theories about how language works. They do not implement their theories on a computer. The development of an easy-to-use software application that enables theoretical linguists to test their theories could help to deal with this problem. How exactly this software should be designed and what exactly it should do are open questions though. But something of this sort could potentially be extremely beneficial.

Chesi argues that Minimalism has become marginalized because of its lack of a fully explicit theory. This may be the case. Development of a fully explicit theory should be a goal of linguistic research. I do not know of a clear way to achieve this, but developing a better understanding of language could help. Linguists (not just those working in Generative Grammar) should be striving to find out how language actually works. The goal of linguistics research is not to come up with clever theories of new (and occasionally old) data. The goal is to actually understand the human faculty of language. Linguists should view theories that are overly complex with suspicion – too much complexity likely indicates a lack of understanding of a phenomenon. Linguists in different disciplines should talk to each other more and see if there is common ground. Common ground may lead to progress.

4. Conclusion

I have focused on a few of the issues that Chesi raises in his paper. Notably, I think that referring to LLMs as theories and/or as grammars is potentially problematic. A set of benchmarks for testing Minimalist theories could potentially be beneficial, but it faces obstacles, in that there is very little work done modeling Minimalist theories, and in that Minimalist theories at this point are not at all unified. Generative AI, and the LLMs that it produces, is a truly impressive technology (not necessarily in a good way) that is having a large impact (not necessarily good) on human society. However, notions that an LLM can account for how language works may be misguided. LLMs can be useful (as well as harmful), and they are easily accessible. Thus, they will overshadow theoretical work in linguistics. But theoretical work in linguistics gives us insight into how language works. In my opinion, LLMs do not.

Notes

[1] It is quite difficult (and often not possible) to reliably distinguish language produced by an LLM from language produced by a human.

[2] Due to lack of space I do not discuss some of the other important issues that Chesi raises.

[3] Hu *et al.* cite Carnie (2012), but I assume that they are referring to the 2013 edition.

[4] While this approach is influenced by Minimalism, it does not appear to have closely followed recent work in Minimalism.

[5] I did an informal survey of syntax-related articles to see how often the latest theories of Chomsky (Chomsky 2013 and later) are discussed. In *Glossa: A Journal of Linguistics*, I counted 19 articles from 2024 that discuss syntax. Of these, only 3 referenced some of the latest work by Chomsky. In the journal *Syntax*, for 2023, I counted 14 articles, of which only 2 cited some of the latest Chomsky work. It looks like the majority of syntax-related publications in my field are not focused on the latest theories of Chomsky (which is primarily what I focus on in my research). I think that these facts show that the field of Generative Linguistics, and linguistics in general, is not exactly unified.

Bibliographical References

See the unified list at the end of this issue.