DOI: 10.26346/1120-2726-246

Rejecting a pessimistic assessment and instrumentalist prescription for linguistic theory

Edward P. Stabler

University of California, Los Angeles, USA <stabler@ucla.edu>

Chesi suggests that disagreements among leading researchers indicate that generative grammar is failing, and he proposes that large language models may provide a better kind of theory, one that rejects modularity in favor of using all aspects of surface linguistic context in every prediction. But the argument is not persuasive. Widely accepted and ongoing empirical advances suggest that alternative generative linguists agree on much more than recent programmatic disputes might suggest, and formal studies confirm this. While language models have significant instrumental, surface predictive power, as competent speakers do, these do not disconfirm claims of generative grammar or provide any alternative explanation of what human language is and why it has the properties it does.

KEYWORDS: generative linguistics, syntax, modularity.

One main thread through the argument in Chesi (*this issue*) has these six steps:

(1) TROUBLE IN GENERATIVE LINGUISTICS. In a 2012 meeting of leading researchers in linguistics, "it was practically impossible to present all the problems in a concise and coherent manner within a consistent framework," and it was hard to specify "the extension of the relevant empirical basis fitting [each] specific theory." Chesi blames the 'spectre' of the Minimalist program: "Thirty years on, it must be acknowledged that while the program began with commendable intentions, the emerging framework still lacks consistency." Rather than leading to a coherent and testable theory, he says minimalism has "turned out to be a sum of idiosyncratic interpretations."

(2) THE SUCCESS OF VERY LARGE LANGUAGE MODELS (vLLMs), on the other hand, Chesi says, might lead one to conclude that vLLMs are "genuine theories of language." He says,

> Their dimension might be an issue, but only when a smaller model would obtain a comparable level of accuracy on [a set of benchmark] tests. In this respect, these vLLMs are, in fact, really the best theories on the market, i.e. observationally more adequate than any [Minimalist Grammar].

 (received February 2025)

(3) ON THE RELEVANCE OF PSYCHOLINGUISTICS, Chesi thinks linguists are failing to make relevant connections with psychological research:

> Generative linguists are sitting on the bench, watching the game, laughing at some experimental results (…) But they remained in the background. As Piantadosi provocatively said, this is "what happens when an academic field isolates itself from what should be complementary endeavours".

(4) ON THE RELEVANCE OF MODULARITY, which posits simple, individually overgenerating, components to define language structure in their interaction, Chesi says, "an operation that overgenerates systematically is, computationally speaking, useless."

(5) THE DIAGNOSIS. He says his "primary concern is that the Minimalist Program's underspecification of key concepts (…) has become untenable (…) [T]he original sin of most generative linguists is that they have gotten used to incomplete pseudo-formalizations and data fragment explanations"

(6) THE PRESCRIPTION he offers is: a formalization that connects theory to data, and a shared test/reference data set "that encompasses all the relevant contrasts we aim to capture." Lacking these "has caused generative linguistics to lose its footing and become marginalized in the contemporary landscape."

The perspective in (1-6) is not uplifting. I think a more positive perspective, sketched here in (1'-6'), better fits the facts. I invite the reader to consider which alternative is better supported.

(1') GENERATIVE LINGUISTICS is a vibrant, active field, especially recently with enlarged perspectives coming from understudied languages, with more sophisticated empirical grounding of abstractions, and with surprising convergences among competing traditions.

It is a mistake to regard disagreements among leading researchers as a sign of failure. Leading researchers always disagree. That is part of what being a leader in science is. But more importantly, deep disagreements about foundations and evidence lurk in even the most successful fields. This can be verified with even a quick glance at whether the conflicts between quantum mechanics and relativity can be reconciled, whether string theory is empirically testable, whether group selection is needed in

evolutionary biology, what evidence bears on the puzzling emergence of ribosomes in early life forms. In the sciences, the significant points of consensus are not in the foundations or on the frontiers, but in the intermediate and often approximate empirical generalizations to be accounted for.

A better assessment of the vitality and success of a science is provided by considering whether important ideas are evolving, whether they are being revised and overturned by new and increasingly comprehensive empirical research. This measure, I think, supports a very much more positive assessment than Chesi's. Even staying close to consensus views, very many things that seemed to be true when I was a student have turned out to be false. And theoretical developments are bringing new kinds of phenomena into focus. For example, among things that I have been considering recently:

- it is now clear that surface constituent order universals will have to be rather abstract;
- even the rather narrow surface order hypothesis of the 'final over final condition' seems to be violated by certain particles;
- some constituent displacements seem to be non-syntactic or, at least, quite unlike most of the rest of syntax (e.g. Irish pronoun displacement, hyperbaton in Latin and Greek);
- various and perhaps all exceptions to the mirror principle are plausibly attributable to post-syntactic operations;
- remnant movement is possible in a variety of conditions;
- hyper-raising is also possible in a number of languages;
- like CP with φ features, it seems TP with φ features blocks A-movement. Developments like these constantly update the 'new normal' in syntax.

Some basic things stay mainly the same:

- hierarchical structure is implicated in selection, movement, agreement, and case – often relating non-adjacent positions in pronounced strings;
- prosodic and phonetic processes show sensitivity to covert syntactic structure;
- some of the languages thought to be most unlike English – e.g. Warlpiri, Niuean, Salish languages like St'át'imcets, Algonquian languages like Innu-aimûn – reveal surprising regularities shared with English and other languages;
- one abstract kind of surface constituent order universal is that the sets of pronounced strings (of any category of any language) are mildly context sensitive – as evidenced by converging analyses of tree adjoining grammar, combinatory categorial grammar, minimalist grammar, and significant parts of lexical functional grammar;

- since mildly context sensitive grammar derivations have finite state definitions, it now looks like most, possibly all linguistic structures may have tier-based strictly local grammars;
- to a good first approximation, elements that are syntactically unique – in the sense that they are constants of all structure preserving maps – seem to be semantically unique in the sense that their values are invariant under permutations of the domain of interpretation.

Every active linguist could make a list like this, mixing new discoveries, major theoretical developments, and fundamental, long-standing working hypotheses. Generative linguistic theory is active and moving in directions that could not have been imagined at its origins.

Why is Chesi's assessment so different? Well, notice that the hypotheses that he focuses on are much more sweeping and programmatic: relativized minimality, the 'cartography' of linguistic structure, Kayne's linearization, the poverty of stimulus argument, the T-model. These are very high level proposals, touching on virtually everything in generative syntax. It is not a surprise that leading researchers would like to talk about such things, and it is equally unsurprising that there are big disagreements about them! Those disagreements should inspire new research, not get you depressed.

(2') Very Large Language Models (vLLMs) have enormous potential in the sciences, but characterizations of what they can successfully learn are not available, nor are there good high-level characterizations of how they compute.

In particular, vLLMs have not been the source of any of the theoretical developments or working hypotheses in linguistics listed in (1'). One reason is certainly that so little is understood about what and how vLLMs learn. The technology was developed largely by trial-and-error, leading one award-winning Google scientist to call that technology 'alchemy'. A widely quoted, appropriate response disagrees with the tone but not the substance of this charge, and points out that "engineering artifacts have almost always preceded the theoretical understanding."[1] The fact that these devices are so poorly understood is also one of the reasons that they are hazardous in critical applications (Bengio *et al.* 2024).

The main linguistic idea that Chesi seems to think vLLMs challenge is the innateness hypothesis, the hypothesis that human language learners must have language-specific biases in order to explain how they get from the kinds of linguistic data they have available to their grammars. Chesi says this is "a cornerstone of generative linguistics." But notice

that it seems to have no bearing on any of the empirical issues I mentioned in the previous section. It is discussed, but only rather rarely and usually tangentially in mainstream syntax journals. It is not mentioned in any of the 8 or so introductory syntax textbooks on my shelf, except for one paragraph in van Riemsdijk & Williams (1986).

How is this a 'cornerstone' of generative grammar? There are actually (at least) two ideas that should be distinguished here, between which there is some tension. The first, from the 1960's:

PoS (POVERTY OF THE STIMULUS). Language is so complex that it is implausible that it could be inferred from the data that suffices for human language learners, without some distinctively linguistic bias.

The second idea – the one suggested by the van Riemsdijk & Williams (1986: 303) textbook – is:

SoL (SIMPLICITY OF LANGUAGE). When a feature in the grammar "predicts the language's behavior with respect to a wide variety of constructions, and predicts as well how that language will differ with respect to those constructions from a language that makes the opposite choice," then this indicates that human languages may be simpler than they originally appeared, and so language acquisition could be much easier than might have been assumed.

The assumption that SoL holds very generally underpins 'parameter setting' theories of language acquisition.

It is certainly possible that vLLMs could challenge PoS. Chesi proposes that we could put this challenge to the test by carefully evaluating the relative success of vLLMs *vs* linguistic theories on a large data set. I agree, if the data set really is similar to what a human would require, and if the vLLMs actually acquire languages that are really like the ones we have. Both of those criteria are difficult to assess, especially the latter, since we lack a good characterization of what human languages are. Chomsky (1975: Ch. 1) famously emphasizes that prerequisite question, and points this out as a goal of generative grammar. This is almost always a surprise to students, who think that it is perfectly obvious what a human language is – but if you try to be precise and correct, it is not obvious at all. Precise characterizations of what is to be learned are also a focus of machine learning theory.[2] Some psycholinguists working on acquisition also agree about that prerequisite question, as we see in the first sentence of Crain & Thornton (2021), for example. And Zipf's law guarantees that finite, testable corpora will always miss significant parts of the language.

Ongoing research is pursuing precise and even beautiful algebraic descriptions of distributed, high-dimensional (and possibly asynchonous) computation, to replace what Piantodosi calls "a mess of billions of weighted connections between sigmoids" (Smolensky 1990; Plate 1994; beim Graben & Potthast 2014; Kleyko *et al.* 2023). If we knew what to look for, we might find it, but we currently lack appropriate conceptual tools. That is why Chesi and Piantadosi provide no examples except PoS. Upsetting PoS would be interesting, but it would still be interesting to understand how languages work, and what human languages have in common – returning us to issues in (1'). As for SoL, vLLM research has nothing (yet) to offer, as we have seen in its inability to address any of the sorts of issues in (1').

(3') PSYCHOLINGUISTICS, with its precise, quantified measurement of human behavior in experimental settings, reveals abundant evidence of the linguistic structure posited by generative linguistics.

Although there is significant evidence of structures in language, it is difficult to get a very clear view of them. For example, it is very rare that precise, quantitative evidence can decide among competing theories in syntax. This must be what Chesi means when he says linguists are isolated. One challenge has been teasing apart the many factors that influence human linguistic performance, so that we can control for things not related to particular mechanisms that we are interested in. Eye tracking during reading clearly shows sensitivity to linguistic structure, but it is difficult to fully disentangle the linguistic processes from effects of attention and eye movement control (Clifton *et al.* 2015); effects of structure but also audition and attention are clear in speech comprehension (Fodor & Bever 1965; Holmes & Forster 1972; Cecchetti *et al.* 2023); event-related brain potentials (ERPs) and functional magnetic resonance imaging (fMRI) are sensitive to syntax but also to many other things (Stowe *et al.* 2018; Brennan *et al.* 2016; Li *et al.* 2022).

Another complicating factor is that, at the more abstract level adopted by most linguistic theories, competing theories often share more than is obvious at first. Though they look very different, recursive and iterative processes are equivalent (Turing 1937). Depending on many other factors, top-down, bottom-up and the various hybrid analytical engines (parsers, etc.) can all have similar performance profiles, especially on short inputs (Kaplan 1987).

Given all these factors, it is no surprise that many different proposals emerge before the understanding of how they all relate to each other. I think the right strategy here is not to insist on the engagement of theories that we do not yet understand how to engage. The best that can be

done is to allow theories to emerge as we come to understand what psycholinguistic methods are most revealing and what regularities can be identified, isolated and developed.

Chesi suggests, on the other hand, that reference to poorly understood non-linguistic mechanisms is too often used by linguists to excuse empirical and theoretical failures, calling this the 'dust under the carpet' strategy. But he offers nothing to challenge the examples mentioned above, examples of the relevance of domain-general visual and auditory mechanisms, and wide ranging ERP and fMRI sensitivity – methods that provide abundant evidence of linguistic structure. A better name for what is happening might be 'dust on the lens': With current methods and current understanding, it is a challenge to clearly distinguish the traces of linguistic processes in the massively parallel and asynchronous computation of the human mind.

Chesi mentions two examples that he thinks really show an unsound 'dust under the carpet' strategy. First, he says in §2.2 that we see this methodological mistake in Chomsky *et al.* (2023) when they do not go into difficult judgements about subject islands. It is hard to take this criticism too seriously when Chesi does not mention any mistaken claim on their part, and the literature he cites describes efforts underway to integrate and extend experimental results in this domain with broader Chomskian theories of language structure and processing. Chesi also claims that hypotheses about modules in the language faculty are instances of the unsound 'dust under the carpet' methodology. In particular, in §3.1.2 he argues against the hypothesis that quantifier scope and interpretation are "a matter for the LF component" in the sense that quantifier scope and interpretation do not influence the pronounced sentences. His remarks on this difficult topic are brief and, I think, unpersuasive, but more importantly, they reveal some confusions about the role of modularity in science and in computation.

Chesi suggests that whether some component should be included in a theory (or some component of a theory) should be assessed with respect to the succinctness of the theory and data, with and without that component. While succinctness can be a clue when assessed with respect to the motivated vocabulary of a successful theory, it does not support fine, finite comparisons because of its sensitivity to the encoding. In science, components are usually identified instead by the relative independence of causes and effects. Indeed, Chesi's own argument against separating quantifier scope from syntax is causal, not succinctness-based. In any case, the modular strategy is one of the oldest in science, famously mentioned in Plato's *Phaedrus* (265e). Computer scientist Edsger Dijkstra says, "when faced with different concerns, we should

separate them as completely as possible, and deal with them in turn" (Dijkstra 1982: §636). So let us consider (4').

(4') MODULARITY AND COMPOSITION. In constraint-based programming, the interaction of general constraints picks out a desired solution. In standard computing, multi-purpose functions (like 'add 1 to the register') are composed in specific, restricted ways. To compute linguistic structures, strings and trees are often defined with large sets of transducers, each of which is general, simple and overgenerating, but when composed they can compute the effect of all of them in one traversal (Engelfriet *et al.* 2009).[3] It is not an exaggeration to say, as Milewski (2020: §1.3) does: "composition is the essence of computing."

It is a puzzle why Chesi would make claim (4), which, on the face of it, is a plain mistake. It might derive in part from Piantadosi's argument that vLLMs "refute Chomsky's approach to language," not because they tell us anything about any of the things generative linguists are working on, but because that whole framework is misconceived, misconceived because syntax is radically non-autonomous. To support this view, Piantadosi does not state any particular syntactic hypothesis and show how it is really a semantic or statistical claim, nor does he state any general non-modular claim and show that it somehow subsumes syntactic ones. Instead, he observes that syntax, semantics, and world knowledge are not deliberately separated in most vLLM training data. But that observation provides no support, since the same is true for human language learners.

Consider some of the more careful arguments for non-separability in the long literature on this topic. McGee & Blank (2024) show that, in transformer-based vLLMs, 'attention heads' that seem most attuned to syntactic features are nevertheless also sensitive to semantic plausibility. And since what we are interested in is human autonomy of syntax, it is perhaps relevant that fMRI studies similarly fail to find regions of the brain that are exclusively syntax-specific (Blank 2016). This kind of thing is no surprise. Many years ago, Cutler & Fodor (1979) showed that even phoneme recognition is sensitive to discourse cues, and that study was recently replicated by Beier & Ferreira (2022). Do any of these sorts of results show that it is a mistake to try to define regularities in syntax that are independent of what is meant or what is plausible or what was just said?

It is interesting that even after Cutler & Fodor (1979) and many other similar results, Fodor was convinced that linguistic processes are separate, modular in a relevant sense (Fodor 1980, 1983, 2010). He carefully notes that there are many senses in which two processes can be 'separate'. So here, suppose we grant that plausibility in context has an

effect on the analysis of a phoneme or a syntactic constituent. Does that mean that phonetics and syntax need to involve in-context-plausibility? Certainly not! If the factors relevant to assessing plausibility are relatively independent from those defining the acoustic cues for phonemes, or the cues for syntactic constituency, they should be defined separately, even if IN PROCESSING they are always deployed together. Watching the runs of a sorting program, you may never see evidence of ordering checks except when elements are permuted, but ordering and permutation are very different things. They should be defined separately.

(5') CLARITY AND PARSIMONY. As empirical generalizations are noticed, they should be formulated as precisely as possible, tested, and reformulated as necessary.

Linguistic field work does not require an understanding of how extended multi tree transducers can be composed, any more than biological field work requires an understanding of Fisher's contributions to genetics. In linguistics, as in every science, there is often an uneasy tension between the mathematicians and the experts in the lab. Both are needed, and it is important that they talk to each other. I do not agree with Chesi that the degree of under-specification in generative grammar, or in the minimalist tradition specifically, has become "untenable." For example, I think there are quite clear, well-informed and well-situated discussions of how certain clitics challenge the final-over-final condition, in spite of the fact that the principle has not, as far as I know, been part of any larger formalization of syntax. Similarly for the arguments that we really do see hyper-raising, or any of the other things listed in (1'). Peer-review is not consistent and not perfect in many ways, but it is our best guide to whether a theoretical contribution has formulated its question and its response well enough to be worth your attention. Select (at least some of) your reading from the best conferences and best journals, with the best referees. Working on and citing a paper provides a vote for others to consider reading it too. With those forces at work, it is hard to imagine how anyone would think that some kind of overarching emphasis on formal rigor is going to benefit the science.

(6') SHARED DATA, SHARED ANALYTIC TOOLS, CLEAR HYPOTHESES, AND CLEAR CONNECTIONS WITH COMPETING AND COMPLEMENTARY PROPOSALS – of course these are desirable.

I have not used the data sets that Chesi mentions, but, in my career, I have made a number of requests for data, and in almost every case, it was

provided to me. One advantage of this informal kind of access is that it (often) connects me directly with the original field researcher, setting the stage for any needed clarification about methods of collection or aspects of the data not reported. Shared but small collaborations like this have, I think, always been happening, and new communications technologies make collaboration especially easy. And large shared data sets are increasingly helpful in psycholinguistics (as in Ozaki *et al.* 2024, for example).

One other thing that should be noticed here is that some of the most valuable data is quite hard to get. For some endangered and understudied languages, the best access available comes through the careful work of a handful of dedicated researchers. I know I am not alone in being grateful to linguists with the skills and situations that enable them to do this work. Often, those linguists report not only speaker judgments, indicating appropriate care about how these were obtained, but they also provide initial, sometimes very well-informed interpretations of how their results fit with what is known about other languages. That allows them to give particular attention to surprising results, often with the consequence of realizing that those results are actually not completely unexpected theoretically.

In computer science, and in computational linguistics especially, where the analytic tools can be very substantial, requiring years of work to assemble, replication became a serious problem many years ago. But (as mentioned in the video cited in note 1) virtually all major publications and conference papers are now expected to make relevant code available, if their claims depend on that code. Many researchers working on vLLMs realize that the resources required for development and experimentation with those systems are prohibitive for many, and members of some of the wealthier corporate research labs have worked hard to make not just code and data but also computing resources widely available when that is feasible.

In my own work, I have focused on key and challenging parts of linguistic theory, and I have never felt that I would learn more from a large scale formalization/implementation of current ideas, or that I needed such a formalization to decide which ideas look promising. I have plenty of serious problems on my plate already. And in linguistics quite generally, the role of large scale formalization and data collection in major theoretical developments has not been central. I think this is true not only in Chomskian linguistics but also in linguistic traditions where formalization and implementation has been more prominent: TAG, CCG, type-logical grammar, HPSG, LFG. That could change, but I do not see it changing anytime soon.

Formalization of small, parsimonious fragments of linguistic theory, with an eye on studies of human linguistic performance, is an interest I share with Chesi, so it is no surprise that we agree on the value of this work. It is most exciting for me when formalization leads to theoretically valuable conceptual clarification, but it can also be useful for showing students the challenges of getting from grammar to the intended structures, and for psycholinguists interested in testing how the calculation of structure might relate to measures of human linguistic performance.

Abbreviations

ERPs = event-related brain potentials; fMRI = functional magnetic resonance imaging; PoS = Poverty of Stimulus; SoL = Simplicity of Language; vLLM = very Large Language Models.

Acknowledgements

Thanks to Dominique Sportiche and Kristine Yu for discussion.

Notes

[1] Ali Rahimi compares vLLMs to alchemy in presenting an award winning paper, now available as a short video that can provide a rough feeling for the nature of the research, even for those without a deep background in machine learning: <www.youtube.com/watch?v=x7psGHgatGM>. Yann LeCun, Meta Chief Scientist, Turing award winner, and NYU professor, responded on his blog: <www.facebook.com/yann.lecun/posts/10154938130592143>. That was 2017, but the issues raised so clearly there have not gone away – see note 2 and see Bengio *et al.* (2024).

[2] There are precise characterizations of what can be learned in the limit, what is PAC learnable, and what is learnable in the sense that empirical risk tends to zero (Poggio *et al.* 2004). But as mentioned in the video cited in note 1, these results are part of what made the sudden success of vLLMs so surprising! New ideas about successful generalization are emerging: Belkin *et al.* (2019); Zhang *et al.* (2021); Martinetz *et al.* (2024), *inter alia.*

[3] In programming, this kind of step is sometimes called 'deforestation' because not only can it avoid overgeneration, but can sometimes eliminate the need to build intermediate trees (Wadler 1990). The potential for this kind of composition in defining human linguistic processing has been observed before (Chen & Hale 2010; Stabler 1991; Gorman 2016*, inter alia*).

Bibliographical References

See the unified list at the end of this issue.