DOI: 10.26346/1120-2726-241

The dose makes the poison: Chesi's vision and subregular syntax

Thomas Graf

Department of Linguistics, Stony Brook University, Stony Brook, NY, USA
mail@thomasgraf.net

In this reply to Chesi's *Is it the end of (generative) linguistics as we know it,* I argue that the specifics of his vision for generative syntax in the 21st century remain hazy. Depending on how one interprets Chesi's methodological desiderata, they may well have a chilling effect on novel approaches instead of fostering them. As a concrete example of this dynamic, I discuss the problems with Chesi's focus on benchmarks and Minimum Description Length and how it would undermine recent efforts in subregular syntax that are in fact closely aligned with Chesi's goals. I conclude that Chesi's vision has merit, but only in moderation.

KEYWORDS: Minimalist grammars, subregular syntax, syntactic benchmarks, minimum description length.

In *Is it the end of (generative) linguistics as we know it,* Chesi argues that Piantadosi's criticism of generative syntax contains more than just a grain of truth, and he proposes several changes to keep generative syntax competitive in this brave new world of LLMs. From the perspective of computational syntax, I agree with Chesi that there is a lot to criticize about how generative syntax, in particular Minimalism,[1] currently operates as a theory and as a field: published papers frequently require a fair amount of exegesis in order to arrive at a fully worked out analysis; there are systemic issues with what counts as data and how that data is collected, reported, and preserved; the analytical space is delineated by pre-theoretical claims about computation and learnability that have little grounding in the actual research on those topics. The reader can probably add a fair number of their own pet peeves to this list. But what isn't perfect isn't necessarily in dire straits. The *status quo* always falls short in comparison to a bold vision unencumbered by reality. This is why it is important for the critic to present their envisioned alternative with sufficient detail so that their audience can assess whether this alternative is SOUND and FEASIBLE.

Chesi's paper leaves key issues to the reader's imagination in this respect. One extreme interpretation, henceforth EI, is that Chesi wants the merit of generative proposals to be determined by EI1) their quantitative performance on a common test set such as SyntaxGym and EI2)

(received November 2024)

their Minimum description length (MDL). This interpretation is neither sound nor feasible. A much more moderate interpretation, henceforth MI, is that generative syntax should MI1) integrate corpora and databases into its empirical base, and MI2) adopt a broader methodology of theory comparison that also includes quantifiable metrics. This much more modest (detractors may say milquetoast) interpretation is not only feasible, it is already reality, e.g. through recent work that grows out of *subregular syntax* (Graf 2022a; Graf 2022b; Hanson 2025). In the following, I will first argue against EI (Sec. 1) and then sketch how subregular syntax instantiates the spirit of MI but would be severely handicapped by the adoption of EI (Sec. 2).

1. Against benchmarks and MDL

Under EI, Minimalism would incorporate non-generative methodologies in response to the LLM paradigm Piantadosi champions. Syntacticians should measure theories' observational adequacy over curated datasets, and they should formalize descriptive adequacy via Minimum Description Length (MDL). I will briefly touch on the issues with this position, many of which have already been discussed in the literature much more in-depth than is possible here. This abstract methodological discussion will form the backdrop against which the very concrete, practical problems with EI will emerge in Sec. 2.

For starters, it is far from obvious that Piantadosi is indeed arguing from a position of strength, which diminishes the motivation for EI. Kodner et al. (2023) provide a detailed criticism of Piantadosi's claims, and Lan et al. (2024) argue that the findings in Wilcox et al. (2023), which Chesi perceives as a watershed moment, paint an incomplete picture: "We examine the evidence further, looking in particular at parasitic gaps and across-the-board movement, and argue that current networks do not succeed in acquiring or even adequately approximating wh-movement from training corpora roughly the size of the linguistic input that children receive" (Lan et al. 2024, 1). Hence one should not be too hasty in adapting the panacea offered by EI. And indeed the veneration of benchmarks (Sec. 1.1) and MDL (Sec. 1.2) fails to appreciate the unique challenges of studying syntax.

1.1 The problem with benchmarks

Consider first the case of "shared benchmarks such as SyntaxGym, CoLa, or BLiMP" (p. 39). Chesi calls LLMs "observationally more adequate" (p. 40) than Minimalism because LLMs seem to "perform properly on shared benchmarks" (p. 39). But shared benchmarks cannot do the heavy lifting of observational adequacy.

Even in NLP, which is utterly dominated by the use of benchmarks, there is increasing awareness that they provide a very limited way of measuring a model's performance. Numerical scores tell us how often a model gives the correct answer, but they do not shed light on whether the correct answer was given for the correct reasons. Nor do they capture how badly the model gets things wrong when it gets them wrong. And most importantly, model X can greatly outperform model Y in the current benchmark yet underperform relative to Y when tested against a new dataset that controls for some confounds of the old one. In their discussion of the model BERT, Bender and Koller (2020: 5186) note that "BERT's unreasonably good performance on the English Argument Reasoning Comprehension Task [...] falls back to chance if the dataset is modified by adding adversarial examples that just negate one piece of the original" and that "BERT's performance on the English Multi-genre Natural Language Inference dataset [...] is predicated on its ability to leverage syntactic heuristics [...]. In a dataset carefully designed to frustrate such heuristics, BERT's performance falls to significantly below chance." Bender and Koller (2020: 5186) warn that a model's performance on a given benchmark may just be "a mirage built on leveraging artifacts in the training data".

One might object that syntactic theories would be less likely to employ such heuristics, but in order to capture observed behavior, they effectively have to. This is because in the realm of corpora and experiments, there is no such thing as grammaticality judgments, only acceptability judgments. But acceptability invariably involves factors that go beyond syntax, such as lexical frequency, semantic plausibility, and processing difficulty. A syntactic theory that does not account for these factors cannot hope to replicate attested acceptability judgments, while a theory that does take them into account has access to non-syntactic heuristics that could allow it to perform well in benchmarks despite getting the syntax wrong.

But perhaps this merely shows that we need to keep building better and better benchmarks of increasing sophistication, controlling for more and more confounds? This retort presupposes that there is a finite number of confounds, that we know them all, and that there is a way of addressing each confound without creating new ones. Odds are that at least one of those three isn't true. And none of this even considers the time and resources generative syntax would have to pour into this ill-motivated enterprise of corpora creation. Nor does it price in the cost of maintaining benchmarks, the potential long-term effects of transcription errors on theory building, and how the reliance on benchmarks marginalizes understudied, low-resource languages (all of which are well-known issues in NLP). A heavy focus on benchmarks would also reduce theoretical diversity because only a few approaches have enough practitioners to continuously tweak their theories for better benchmark performance. Every scientific field is subject to the Matthew effect, and in the short to medium term, a lousy theory with lots of manpower and resources is likely to outperform a good theory that only a few researchers are working on. Elevating benchmarks to the arbiters of observational adequacy would greatly exacerbate this. With no clear pay-off and many risks, the "benchmarkification" of observational adequacy looks like a methodological dead end.

1.2 The problem with MDL

The limitations of benchmarks and corpora also affect MDL, which Chesi presents as "a practical mathematical way to compare the 'descriptive adequacy' fit [...] of a theory" (p. 9). This vastly oversimplifies the intricacies of MDL. As a concrete example, consider Ermolaeva (2023), who uses MDL to compare multiple Minimalist analyses. To this end, she formalizes them in terms of Minimalist grammars (Stabler 1997; Stabler 2011) and then measures two components: how many bits it takes to encode each grammar (*grammar cost*), and how many bits are needed to encode her test corpus based on the structural descriptions provided by each grammar (*corpus cost*). In an exemplary display of scholarly transparency she carefully lays out all the limitations of her approach:

> MDL can disagree with a linguistic intuition on what constitutes a simpler explanation of the data, if some aspect of the analysis is not taken into account by the encoding scheme, or cannot be expressed by the chosen formalism, or requires an overhead cost that does not pay off in the case of the chosen corpus. [...] Some proposals in the linguistics literature are motivated by patterns that could only be translated into cost reduction under a sophisticated encoding scheme; [...] extremely small datasets can favor overfitting grammars, if the reduction in corpus cost provided by introducing syntactic generalizations is insufficient to justify the initial investment in the grammar. This also applies to large but repetitive datasets. [...] [W]ith a very large corpus of diverse sentences (which is a better representation of natural language as a whole) the MDL value is decided primarily by the corpus cost. (Ermolaeva 2023: 108-110)

MDL results aren't a universal arbiter of succinctness, they are dependent on the choice of corpus. This is problematic for all the reasons mentioned in the previous section, but the problem goes even deeper. Given the limitations and confounds of small corpora, generative syntax would have to consider fairly extensive corpora, which makes grammar cost negligible. But grammar cost is what is actually of interest to syntacticians. A simplicity metric that pays little attention to the simplicity of the grammar simply misses the mark.

Of course one could carefully engineer all the MDL parameters to get a more appropriate result, but that is exactly the problem: MDL is not an easy, objective way of quantifying simplicity, it is an elaborate modeling task that is ripe with linguistically arbitrary decisions. It is replete with formal parameters that differ in subtle ways yet yield different results. Based on what criteria should linguists trust one MDL comparison but disregard another? And since there are few researchers who know both MDL and syntactic theory well enough to combine the two in an insightful manner, an excessive focus on MDL would supercharge the Matthew effect and stifle theoretical diversity in the field. The malleability of MDL disqualifies it as the be-all and end-all of linguistic simplicity and theory comparison.

At this point some readers may object that my arguments against EI have the flavor of conceptual nitpicking that erroneously allows *perfect* to be the enemy of *good*. But quite to the contrary, it is EI that allows *perfect* to be the enemy of *good* by pursuing an empiricist pipe dream of fully quantifiable theory comparison (and by extension, theory construction). In a perfect world, this approach would work and greatly accelerate linguistic progress by allowing syntacticians to skip the challenging and time-intensive task of theory comparison. But NLP is living proof that EI1 is far from perfect, and the work that has been done on combining MDL and syntactic theory casts major doubt on the feasibility of EI2. Insisting on EI ignores these warning signs to the detriment of all the good alternatives that are already around. In the next section, I present subregular syntax as one example of such an alternative, discuss how it meets many of Chesi's desiderata, and explain why EI would have a chilling effect on this enterprise.

2. Subregular syntax: An answer to Chesi's vision?

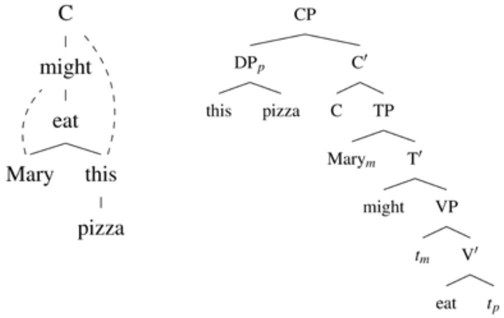
Subregular syntax is a program that combines generative syntax with notions from formal language theory (subregular complexity) that have also been fruitfully applied to phonology and morphology (Heinz 2010; Chandlee 2014; Aksënova et al. 2016; Jardine 2016; Chandlee 2017; Aksënova and Deshmukh 2018; Chandlee and Heinz 2018;

Graf and Mayer 2018; Heinz 2018; Mayer and Major 2018; Hao and Andersson 2019; Burness et al. 2021; Chandlee 2022; Aksënova et al. 2024; Burness et al. 2024). The central goal is to identify very restricted classes of dependencies and constraints that are powerful enough to capture a wide range of empirical phenomena, and to leverage these restrictions for learning, typology, and cognition. Since this isn't the place for a comprehensive discussion (see Graf 2022a,b), I will limit myself to a few brief examples of how this program opens up new analytical avenues for Minimalism (Sec. 2.1) and how it captures the spirit of Chesi's paper as expressed by MI (Sec. 2.2). After that, I will explain why this program could not thrive under EI, the extreme interpretation of Chesi's proposals (Sec. 2.3).

2.1 What is subregular syntax?

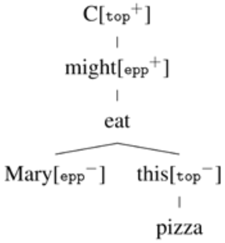
The central object of study in subregular syntax is the computations that underpin syntactic structure building, i.e. the syntactic derivation. A syntactic derivation is represented in a format similar to a dependency graph.

(1) Syntactic derivation and corresponding phrase structure tree for *this pizza, Mary might eat* (*v* omitted for brevity)



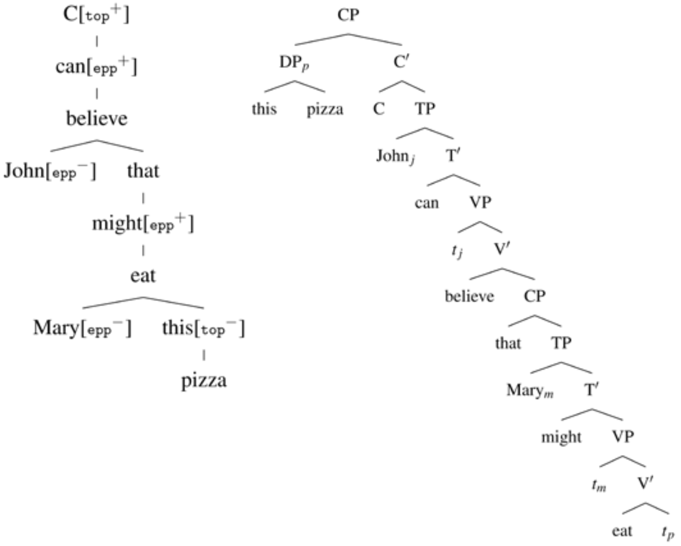
In Minimalist terms, solid branches in (1) indicate Merge steps (with the rightmost daughter of a node being merged as its complement and all other daughters being merged as specifiers). Dashed branches encode a long-distance dependency, usually movement. Since trees are mathematically easier to reason over than graphs, dashed branches are replaced with diacritics of opposite polarity such that minus (-) marks the dependent of a dashed branch and plus (+) its opposite end. The diacritics are usually chosen to reflect syntactic theory, e.g. `top` for topicalization and `epp` for subject movement.

(2) Syntactic derivation with dashed branches replaced by diacritics



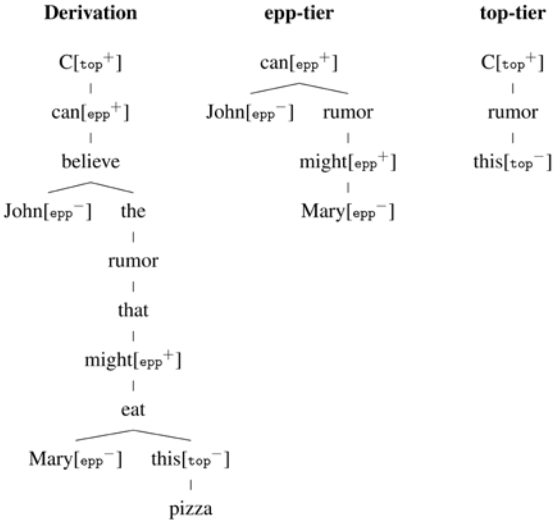
Note that Merge dependencies are strictly local over these representations: for any given head, it suffices to inspect its string of daughters to determine that the head-argument configurations are well-formed (correct number of arguments, each argument with the requisite category, and so on). But long-distance dependencies can span arbitrarily large domains, as in (3).

(3) Syntactic derivation and phrase structure tree for *This pizza, John can believe that Mary might eat*



However, strict locality does obtain if specific nodes can be ignored for calculating locality. This can be visualized as syntactic tiers. A tier contains a subset of a derivation's nodes, preserving their relative order.

(4) Syntactic derivation from (3) with epp-tier and top-tier



Over these tiers, it once again suffices to consider only mother-daughter configurations. In particular, if a node carries a plus-diacritic, say epp$^+$, then its string of epp-tier daughters has to contain exactly one instance of epp$^-$.[2] The difference between a Merge dependency and a long-distance dependency, then, is that the former puts restrictions on the string of DAUGHTERS whereas the latter puts restrictions on a string of TIER DAUGHTERS.

By expanding its focus beyond full structures to tiers, subregular syntax provides many new insights into syntax. Crucially, it does so in a manner that closely matches MI, the moderate interpretation of Chesi's paper.

2.2 How subregular syntax fits Chesi's vision

Subregular syntax avoids many of the criticisms Chesi levels against Minimalist syntax. It is rigorously formalized, grounded in computation and cognition, makes a connection to psycholinguistics (gradience, sentence processing, acquisition), allows for quantitative comparisons of analyses, and integrates corpora into its research methodology.

Subregular syntax grows out of Minimalist grammars and has a full formalization in terms of first-order logic (Graf 2023). This allows subregular syntax to draw from the rich body of computational work in that area (Stabler 2011; Graf to appear). Notably, subregular syntax is fully compatible with the Minimalist grammar work on sentence processing that Chesi mentions (Lee and De Santo; Kobele et al. 2013; Gerth 2015; Graf et al. 2017; Hunter et al. 2019; De Santo 2020; Pasternak and Graf 2021; Liu 2023). But even without the links to Minimalist grammars, subregular syntax is a rigorous enterprise in the sense that formal rigor is indispensable for some of its key findings.

Consider, for example, the status of tiers in the grammar. While they may look like a new kind of syntactic representation, the mathematics reveals that tiers are a visual metaphor for a specific kind of cognitive architecture. This is easier to illustrate with strings. Suppose that some language has CV syllables and a vowel harmony system without neutral vowels where *a* and *o* cannot occur in the same word as *i*. Hence *baboba* and *bibibi* would be well-formed, but not *babobi*. From a cognitive perspective, this system only requires enough working memory to store the last two symbols and the ability to check whether the current symbol is a valid continuation of the previous two symbols.

(5) Memory configurations for vowel harmony when processing *babobi*

MEMORY CELL 2	MEMORY CELL 1	CURRENT SYMBOL	PERMITTED?
-	-	b	Yes
-	b	a	Yes
b	a	b	Yes
a	b	o	Yes
b	o	b	Yes
o	b	i	**No!**

Tiers modify this architecture with a cognitive least-effort principle: memory cells are updated only if relevant symbols are encountered. For our vowel harmony system, only vowels need to trigger a memory update, which is the same as saying that vowels PROJECT onto their own tier. As a welcome side-effect, this also reduces the necessary working memory to just one cell (and if the harmony featured neutral vowels, those simply would not trigger a memory update).

(6) Memory configurations for vowel harmony when processing *babobi* with a vowel tier

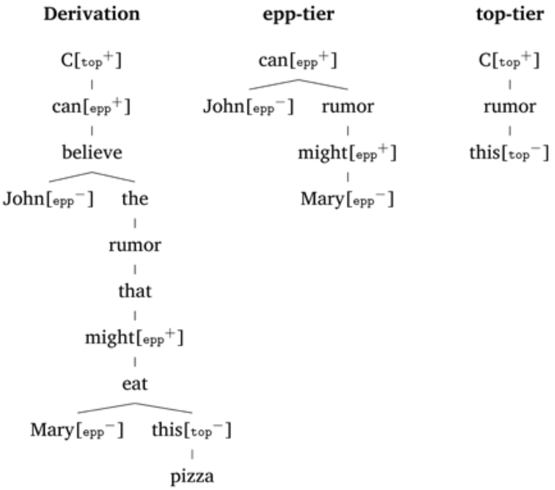
VOWEL MEMORY CELL	CURRENT SYMBOL	PERMITTED?
-	b	Yes
-	a	Yes
a	b	Yes

VOWEL MEMORY CELL	CURRENT SYMBOL	PERMITTED?
a	o	Yes
o	b	Yes
o	i	**No!**

Hence tiers represent a particular way of managing a finitely bounded amount of working memory. This insight can also be used to prove that tiers can only capture a very restricted subset of the regular (string/tree) languages. Yet this subset covers tremendous linguistic ground.

For example, island effects are unsurprising in the sense that they require no new cognitive resources beyond what is needed for long-distance dependencies. If *rumor* projects onto every movement tier as in (7), then `top`$^+$ on the matrix C-head won't have any instance of `top`$^-$ among its `top`-tier daughters and thus the derivation is ill-formed.

(7) Complex NP island *This pizza, John can believe the rumor that Mary might eat*



The same tier projection mechanism that allows for long-distance dependencies thus also allows for them to be disrupted by intervening elements, giving rise to island effects. In fact, this approach to islands can be pushed even further by using a probabilistic tier projection for *rumor* and similar heads, which makes it possible for the grammar to produce the kind of gradient acceptability judgments that have been observed in psycholinguistic experiments (Torres et al. 2023). And as explained in Graf

(2022a,c), the general idea of projecting non-movers onto movement tiers can also be leveraged for other phenomena such as Irish wh-agreement, extraction morphology, and German wh-copying. Tier projection thus ties together a diverse range of phenomena (movement, islands, extraction morphology, wh-copying, and more) in such a manner that the same structural analysis can produce categorical and gradient judgments.

Tiers also put restrictions on externalization that line up with the empirical facts. Graf (2023) shows that mapping a syntactic derivation to its bare phrase structure with copies requires no additional memory beyond what is provided by tiers, but identifying which of these copies should be pronounced is a harder problem. It can be solved with tiers only if syntax obeys a constraint similar to the Ban on Improper Movement, providing a computational motivation for the existence of such constraints. At the same time, tiers also allow us to distinguish between different implementations. A literal implementation of the Ban on Improper Movement requires *n* additional tiers, where *n* grows polynomially with the number of movement diacritics (`epp`, `top`, and so on). Probe horizons (Keine 2020), on the other hand, achieve the same effect without requiring any new tiers. Since each tier corresponds to a separate memory register, probe horizons are preferable to the Ban on Improper Movement because the latter imposes a polynomial memory load whereas the former has no additional cost at all. Tiers thus provide a theory-internal succinctness metric, and in contrast to MDL this metric is directly grounded in cognition.

There is also emerging work on how syntactic tiers can be learned from limited positive data (Swanson 2024), and how the Tolerance Principle (Yang 2016) could be applied to subregular learning (Hanson 2024). With recent findings that at least some neural architectures (LSTMs) have subregular biases (Torres and Futrell 2023), even a convergence of subregular syntax and neural networks is in the realm of possibilities.

On top of all that, subregular syntax also has principled uses for corpora. In Graf (2020), I conjectured that syntactic category systems are subregular in the sense that the syntactic category of a lexical item can be correctly inferred in a strictly local manner. Ongoing work by Kenneth Hanson and Logan Swanson suggests that this conjecture holds for *MGbank*, a fragment of the Penn treebank reanalyzed for use with Minimalist grammars (Torr 2017). Of course MGbank may be missing exactly those constructions that would disprove the conjecture, but this kind of corpus work nonetheless is an essential step in vetting subregular claims.

In sum, subregular syntax is interesting for the purposes of this reply because it instantiates a lot of what Chesi wants to see in a syntactic theory: rigorous formalization, computational grounding, connections to processing and learnability, modeling of experimental data, quantifiable notions of succinctness, and the use of corpora. Despite all that, though, subregular syntax wouldn't flourish under EI, the extreme interpretation of Chesi's proposal.

2.3 Why Chesi's vision doesn't fit subregular syntax

The methodological issues from Sec. 1 that made EI unsound are also the reason why this extreme interpretation of Chesis' paper isn't feasible, either. The case of subregular syntax illustrates this quite clearly.

First, the use of shared benchmarks presupposes a shared vision of what the relevant phenomena are. While subregular syntax draws a lot from the Minimalist literature in its empirical analyses, it is not beholden to them. For example, subregular analyses of binding (Graf and Abner 2012; Graf and Shafiei 2019) have computational reasons to treat binding as a distributional constraint on pronominal forms rather than a constraint on co-indexation. They do not address whether *John told Peter that Bill likes him* is ill-formed when *him* refers to *Bill*, they merely observe that the string contains one or more viable antecedents for *him*. If linguistic theories had to prove their worth on standardized benchmarks, one of them being a test suite of co-indexed binding sentences, subregular syntax would be an immediate failure simply for factoring interpretation out of syntactic binding.

Even for the phenomena where subregular syntax marches in lockstep with generative orthodoxy, benchmarking it would be a laborious task. The first step of bechmarking subregular syntax is to annotate the datasets with subregular tree structures. As Chesi notes (p. 5), SyntaxGym currently contains about 4,000 sentences, and this number is bound to grow within the next few years. Annotating all these sentences would be a herculean task, even with sophisticated tooling for semi-automatic annotation. It would also be theoretically dubious because many constructions that show up in these datasets haven't been investigated from a subregular perspective yet. And barely any of that work would include phenomena that truly challenge subregular syntax, e.g. closest conjunct agreement. Why should subregular syntax spend its limited resources on making itself benchmark-able if those benchmarks are utterly misaligned with the priorities of the program?

Succinctness as formalized via MDL is also a problematic criterion for subregular syntax. Our discussion of vowel harmony in (5) and (6) already showed that the increased expressivity of tiers also comes with increased succinctness. This is a general fact of computation (Savitch 1993). If we wanted to optimize for succinctness, we should not stop at tiers, we should move all the way up to finite-state (string/tree) automata. But then we lose many of the upsides of subregular syntax: learnability, restricted typology, and tiers as a cognitively grounded succinctness metric. MDL comparisons thus would penalize subregular syntax for the very thing it is built on: sacrificing some succinctness for the linguistic benefits of limited expressivity.

Both EI1 and EI2 thus would have the opposite effect of what Chesi seems to envision: rather than elevating innovative work like subregular syntax, EI would stop it dead in its tracks.

3. Conclusions

Chesi's paper presents a grand vision of generative syntax in the 21st century, but it remains unclear just what exactly this vision ought to look like in practice. The paper allows for many interpretations, from the moderate (MI) to the extreme (EI). While MI is fairly uncontroversial, EI is so strong that it would undermine even those enterprises that currently come closest to Chesi's vision. I discussed subregular syntax as a concrete example of this dynamic. Irrespective of how one feels about Chesi's vision, then, there is value to it, but only in moderation. The dose makes the poison.

Abbreviations

EI = extreme interpretation; LLM = Large Language Models; MDL = Minimum Description Length; MI = moderate interpretation.

Notes

[1] Piantadosi and Chesi each use the term *generative linguistics* in their papers even though they only consider the subfield of *generative syntax* and in particular Minimalist syntax. Notably, generative phonology had a very similar debate years ago prompted by the recent advances in NLP (Pater 2019, Rawski & Heinz 2019, a.o.), and many of the arguments made there carry over to the current conversation.

[2] Note that any system that can compute this "exactly one" requirement can also compute its weakened counterpart "at least one". Graf and Kostyszyn (2021) use this fact to explain the existence of multiple wh-movement.

Bibliographical References

See the unified list at the end of this issue.