DOI: 10.26346/1120-2726-239

The Generative enterprise is alive

Sandiway Fong

University of Arizona, USA <sandiway@arizona.edu>

We critically evaluate Piantadosi's claim that deep neural networks have obsoleted linguistic theory. The Generative Enterprise seeks to explain why human language exhibits discrete infinity, yet is not unrestricted. In fact, all languages appear to obey the same basic underlying properties. Through the lens of the Strong Minimalist Thesis (SMT), driven by evolutionary considerations, inquiry has been focused on maximally simple operations such as Merge for structure, and Minimal Search for establishing structural relations. By contrast, the computationally expensive setting of billions of parameters in current deep neural networks perform provides no biologically plausible explanation for human language. Moreover, we show through simple examples that the performance of current systems turn out to be highly overrated.

KEYWORDS: Strong Minimalist Thesis (SMT), discrete infinity, Merge, Minimal Search

The generative enterprise presupposes that language is a 'computational system', expressible abstractly through formalism, and concretely by the so-called language 'organ'. An organ is an entity that has dedicated function, and the idea is that there are functional components of the brain specialized for thought and human language, much like there are areas (uncontroversially) dedicated to vision and olfaction in nearly all animals. It almost goes without saying that the organic brain is a finite resource, yet there is an infinite number (and variety) of possible sentences. We, therefore, must be capable of generating this 'discrete infinity' using a 'finite' computational system, as infinity cannot possibly be memorized. Even if the huge parameter space of the latest Large Language Models (LLMs), recently estimated at 1.8 trillion in the case of GPT-4 could be utilized (source: G. Holz, reported in Benesty 2023), humans are continually inventing new words and novel sentences.

The goal of the generative enterprise is to provide an explanation for the functionality and limits of the language organ, i.e. why language is the way it is and not some other way. Piantadosi writes:

> The state of the art in virtually all computational language tasks makes use of deep neural networks. (Piantadosi 2023)

We will return to consider both what ‘state of the art’ means and what counts as an appropriate ‘language task’, but the scientific task of explaining the human language organ is largely ignored by the artificial neural network community. There are several reasons why this might be the case. First, it is incorrect to assume that by merely adopting the artificial counterpart of the human neuron, one has provided a ready explanation. LLMs are trained on vast amounts of data requiring huge amounts of power.[1] The human brain operates on an estimated 20W, and it is generally acknowledged that a child understands language from almost no exposure.[2] Any neural network architecture that requires incalculable human lifetimes to power up is no explanation.[3] There have been attempts to use smaller amount of training data, as a cursory nod to science, but the general trend is clear: onwards and upwards.[4] A second reason stems from general ignorance in the neural network research community of some of the fundamental properties of human language. For example, given the discrete infinity property, it seems at first rather counterintuitive that language has many limits. For example, we know that language does not make use of the simplest mechanisms that cognition makes available. Chomsky has made this point repeatedly in print:

(1) The simplest operation is certainly within the cognitive repertoire. A child has no problem picking the first bead on a string. (Chomsky 2021)

Linear operations such as picking the first or nearest item are part of the primitive cognitive toolkit, yet language makes no use of this to compute relations between phrases. Chomsky proposes that linear operations are simply unavailable to language, see (2).[5] The surprising ‘basic property’ of language is that relations are structural, and that kids know this as early as they can be tested, so it is not learnt.[6] In other words, instead of a shallow scan of the relevant words in a sentence, we take the considerable trouble of computing (sometimes deep) phrasal structure before conducting a search for the closest relevant phrase.

(2) The puzzle is that from infancy on we ignore 100% of what we hear (linear order) and reflexively use only structures that we never hear but that our mind constructs, with non-trivial computations. (Chomsky 2021)

The examples in (3a-b), from Chomsky (2021), illustrate that English subject-verb number (NUM) agreement operates under phrasal (not linear) search.

(3) a. the $\textit{bombing}_{\textit{sg}}$ of the $\text{cities}_{\text{pl}}$ $\textit{was}_{\textit{sg}}$/*$\text{were}_{\text{pl}}$ criminal
    b. the $\textit{bombings}_{\textit{pl}}$ of the city_{sg} $\textit{were}_{\textit{pl}}$/*$\text{was}_{\text{sg}}$ criminal

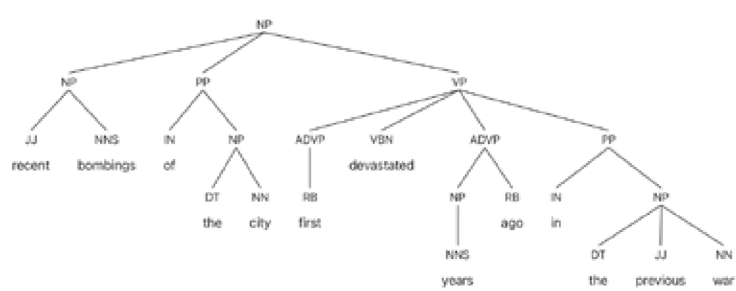
c. {bombing$_{[sg]}$, (of) {cities$_{[pl]}$}}
d. {{bombing$_{[sg]}$, (of) {cities$_{[pl]}$}}, {INFL$_{\varphi}$, {v$_{pst}$, {be …}}}}
e. recent *bombings$_{pl}$* of the *city$_{sg}$* first devastated *years$_{pl}$* ago in the previous *war$_{sg}$*

There is no way to avoid computing noun phrase (NP) structure such as in (3c).[7] In (3d), the subject NP from (3c) has been attached to a phrase headed by INFL$_{\varphi}$, viz. inflection, containing φ-features, e.g. person and number (NUM), that will be spelled out on the verb *be*. The functional head INFL$_{\varphi}$ triggers a simple search for the NUM feature, finding the value SG (singular) associated with *bombing*, the first relevant value in a top-down φ-feature search of the NP (3c). This primitive operation of Minimal Search (MS) occurs time and time again in language, not just for relation computation, but also in the case of Merge, the primitive operation that recursively builds phrase structure. MS identifies the first compatible item in phrase structure, and goes no further. This is the simplest possible search procedure.[8] In fact, Chomsky states:

(4) Right now I don't see any reason why any operation should be exempt from MS. (Chomsky, personal communication)

Piantadosi, and Chesi (*this issue*), cites the SyntaxGym test suites as containing relevant examples to test linguistic theories.[9] But correctly computing phrase structure (so we can then apply MS) is by no means a simple task (even for labeled data-trained statistical parsers). For example, consider the case of the complex NP given in (3e), parsed as in (5) below using the *Berkeley Neural Parser* (Kitaev *et al.* 2019). If MS is applied, we will correctly find *bombings* (tagged NNS, i.e. plural common noun), but details of the generated structure are incorrect.[10] In fact, without appropriate context, the Berkeley parser initially parses (3e) as a sentence headed by *devastated* (not NP).[11] Yet humans accomplish this structural decoding as a matter of simple reflex.

(5)



To take one more example of MS from Chomsky (2021), consider the case of adverb-verb construal as in examples (6a-d) with *carefully*. There are two verbs to choose from, *fix* and *pack*. (For clarity, the linearly closest verb in each case has been enclosed with '[ ]'.)

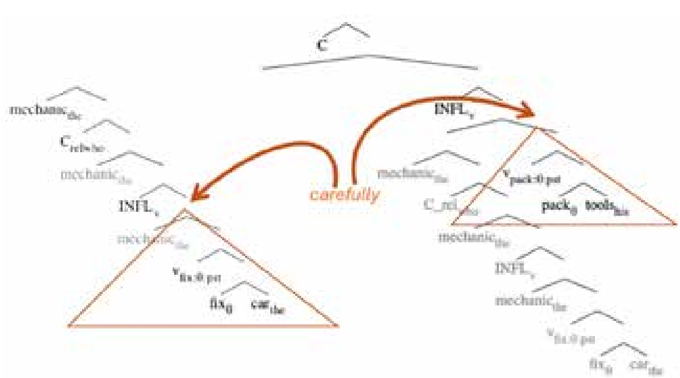
(6) a. the mechanic who *fixed* the car *carefully* [*packed*] his tools
b. *carefully*, the mechanic who [*fixed*] the car *packed* his tools
c. the mechanic who *fixed* the car [*packed*] his tools *carefully*
d. the mechanic who *carefully* [*fixed*] the car *packed* his tools

Chomsky states the task:

(7) Language task: construal
The adverb *carefully* seeks a verb [to modify], but it cannot use the simplest computation: pick the linearly closest verb. (Chomsky 2021)

Let us dispense with the simple cases. Examples (6c) and (6d) straightforwardly obey the linear constraint, selecting *pack* and *fix*, respectively. Example (6b) is a counterexample to linearity, as *carefully packed his tools* is the only possible reading. Finally, example (6a) is a bit of a conundrum. Clearly, *carefully* can modify either *fix* or *pack*. How can we reconcile this with the MS idea of the adverb *carefully* finding only the first (and nearest) verb, then stopping? The answer is quite simple if we assume phrase structure as in (8) below.[12] *Carefully* is positioned between the verb phrases *fix the car* and *pack his tools*. If attached to the former, MS finds *fix*. If attached to the latter, the MS finds the verb *pack*.[13] Therefore, linguistic theory predicts both structures are available.

(8)



I believe Piantadosi must agree that the Chomsky ‘language task’ stated in (7) is a reasonable one. In fact, he has no choice, as it is one that must be computed if one is to understand the sentence. As OpenAI’s ChatGPT is opaque with respect to internal representation, we must instead probe it through careful questioning. With this in mind, consider the performance of ChatGPT on example (6a), reproduced below in (9a-b).

(9) a.

the mechanic who fixed the car carefully packed his tools. What did he do carefully?

The mechanic carefully packed his tools. The action he did carefully is "packed."

b.

Did he carefully fix the car?

Based on the sentence provided, it only states that the mechanic **carefully packed his tools**. It does not specify how he fixed the car, so we cannot conclude that he fixed the car carefully from this information alone.

While (9a) is a fine response, (9b) notably falls short of human performance, despite offering ChatGPT a second chance with explicit hinting. Piantadosi waxes lyrically on ChatGPT’s generative abilities, see (10) below, but this kind of hype is misplaced and misleading as the language task is defined to suit the core characteristic of a LLM, i.e. to generate readily comprehensible follow-on text.

(10) Ignore whatever theoretical persuasions you have for a moment and just recognize how remarkable that response is. (Piantadosi 2023)

Actually, since the dawn of the treebank statistical parsing era, beginning with the venerable million word Penn Treebank (Marcus *et al.* 1994), the language task has been defined with automated machine testing in mind. Under this scenario, it is natural to (randomly) divide up a large corpus into non-overlapping training and test sets. Evalution on the (withheld) test set must be automatically score-able. Instead of a simple correct/incorrect response to a machine-generated parse, or an identifiable task such as (7), easily calculated bracketing similarity is used to smoothly assign partial credit. For many years, chasing an improved F-score with respect to bracketing was the language task *de jure*, see (11) below.[14]

(11) from (Cahill 2008):

	≤40 words			≤100 words		
	LP	**LR**	**F-Score**	**LP**	**LR**	**F-score**
Collins (1999) M1	88.2	87.9	88.0	87.7	87.5	87.6
Collins (1999) M2	88.7	88.5	88.6	88.3	88.1	88.2
Collins (1999) M3	88.7	88.6	88.6	88.3	88.0	88.1
Charniak (2000)	90.1	90.1	90.1	89.6	89.5	89.5
Charniak and Johnson (2005)						91.0
McClosky et al. (2006a)						92.1
Bod (2003)				90.8	90.7	90.7
Petrov and Klein (2007)	90.7	90.5	90.6	90.2	89.9	90.0

The move to automated testing of this kind naturally biases evaluation and development of parsing systems in favor of sentences that occur frequently in the *Wall Street Journal*, as in the case of the *Penn Treebank*, whether a theoretical treatment of the data exists or not. For engineering purposes, this move makes sense: it provides a level playing field for the evaluation of competing parsers. However, as language has a very long tail with respect to sentence variety, a consequence of discrete infinity, if something is not attested in the *Penn Treebank*, with scoring now on a corpus frequency basis, it essentially becomes irrelevant to engineering. In theoretical linguistics, sentences, irrespective of frequency (and particular language), may be critical probes into syntactic (and semantic) theory.[15] As is natural in scientific discourse and general theory development, there are multiple (competing) theories, and plenty of unresolved puzzles exist, so critical examples come and go. Piantadosi claims:

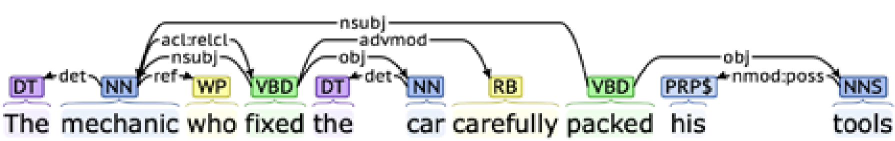
(12) First, [LLMs] are precise and formal enough accounts to be implemented in actual computational systems, unlike most parts of generative linguistics. (Piantadosi 2023)

This criticism is clearly off the mark. ‘Precise and formal’ is only a technical issue if the implementor fails to correctly understand and translate the underlying theory. Substantial implementations of various syntactic theories have been built, e.g. the XLE platform (Kaplan *et al.* 2002) for Lexical Functional Grammar (LFG), the DELPH-IN consortium for HPSG, and in the Chomskyan tradition, perhaps Fong (1991) and

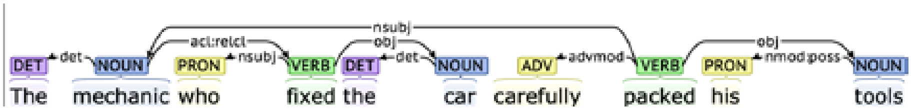
Ginsburg & Fong (2019) represent the most extensive working systems. However, as theories are quite incomplete, the scale of these computational systems must be limited, unless the implementor is willing to forgo theory and also adopt descriptive methods to fill any gaps. In the case of grammatical formalisms that lend themselves to statistical modification, e.g. as in the case of context-free and dependency grammars in the statistical parsing era, or perhaps more recently, in the Minimalist Grammar (MG) framework, broader coverage may be achievable.[16]

There are a few notable disadvantages of integrating statistical data into grammars, perhaps first noted by Chomsky (1956), many decades before probabilistic context-free grammar (PCFG) parsing systems came to the fore in the 1990s. One concerns ungrammaticality (to which we will return later), the second disadvantage concerns structural ambiguity. Intuitively, as the number of training examples grows, the amount of statistical evidence for a structural rule also grows, therefore one might expect statistical parsers to converge upon similar analyses given enough data. However, statistical parsers construct (and also discard) many possible parses for a given sentence, as there are many possible combinations of structural rules, usually reporting only the top-ranked parse. Typically, there are only minute differences in the logprob values of the top ranked parses.[17] As a result, different parsers trained on similar data can surface different structures. For example, (13a) and (13b) represent the output of Stanford's popular state-of-the-art *CoreNLP* and *Stanza* dependency parsers, respectively, on Chomsky's ambiguous sentence (6a).

(13) a. CoreNLP dependency parse

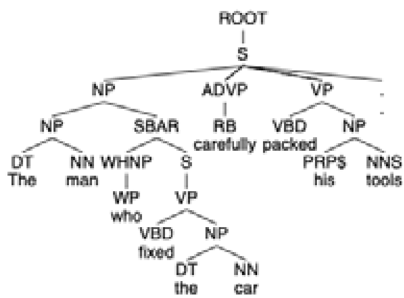


b. Stanza dependency parse

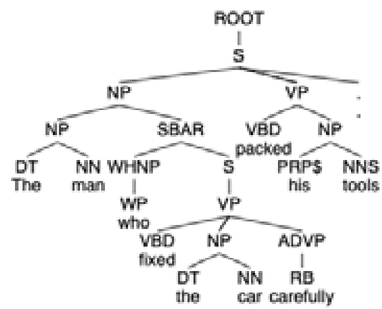


The dependency relation *advmod* indicates that the adverb *carefully* (RB/ADV) modifies *fix* in (13a), but *pack* in (13b). For comparison, the classic Stanford PCFG parser (Klein & Manning 2003) produces structurs (14a) and (14b) as ranked parse #3, #5 and #1, #2 and #4, respectively.[18]

(14) a.



b.



(Not shown here for reasons of brevity, but the *Berkeley Neural Parser*, like *Stanza*, selects *pack* on the same sentence.)[19] Consider now the obvious conundrum facing the poor language engineer: which system should he trust and pick? Recall also from the response in (9b), ChatGPT is no help in this case.

Next, let us turn to consider ungrammaticality. Note that predicting ungrammaticality is not generally considered to be a 'language task' in the engineering NLP community, yet distinguishing grammatical from ungrammatical sentences has been core to the generative enterprise from the earliest days. As anticipated by Chomsky (1956), reproduced in (15) below, inherently, statistical systems run into difficulty on this task.

(15)

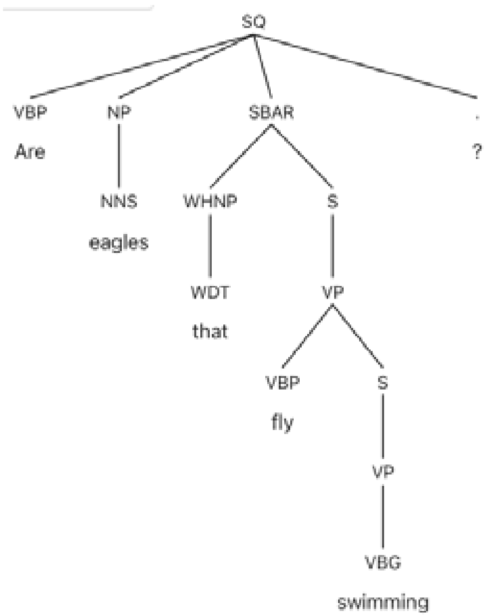
> In § 2.4 we argued that there is no significant correlation between order of approximation and grammaticalness. If we order the strings of a given length in terms of order of approximation to English, we shall find both grammatical and ungrammatical strings scattered throughout the list, from top to bottom. Hence the notion of statistical approximation appears to be irrelevant to grammar. In § 2.3 we pointed

Consider the examples in (16a-b), taken from (Chomsky 2013):

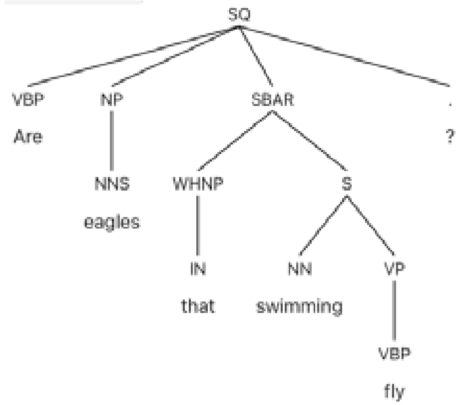
(16) a. eagles that fly are swimming
b. eagles that are swimming fly
c. are eagles that fly swimming?
d. *are eagles that swimming fly?

English generally permits the formation of an interrogative sentence from a declarative one by fronting an auxiliary verb such as *be*, as in (16c), based on (16b).[20] Despite the fact that interrogative (16d), based on (16b), is ungrammatical, the *Berkeley Neural Parser* will happily recognize and parse both sentences (16c) and (16d) as (17a) and (17b), respectively.[21] On the plus side, the Penn Treebank label SQ is an indication that an inverted yes/no sentence has been recognized.[22] On the minus side, however, the syntactic structures for both are clearly incorrect.

(17) a. parse for question (16c)

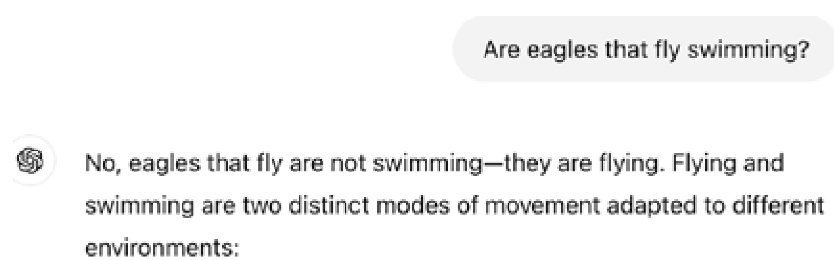


b. parse for ungrammatical question (16d)

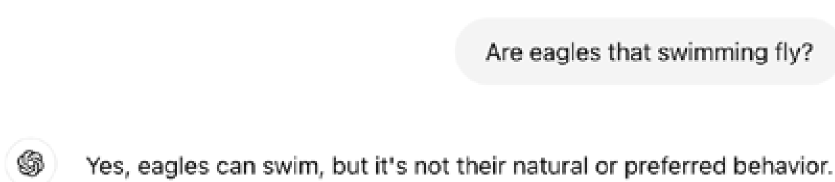


With the advent of representationally-opaque LLMs, as structural descriptions cannot be part of the answer, the language task has been redefined again to suit the framework and narrative. Instead, we must query the LLM, e.g. as in (18a-b) below.

(18) a.

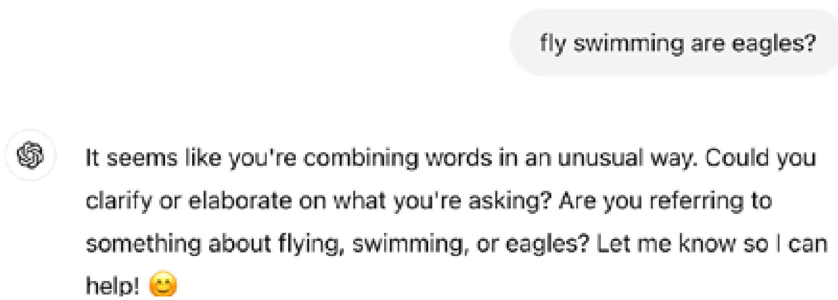


b.



ChatGPT's answer for (18a) simply misinterprets the question. Although (18b) is illicit in English, ChatGPT, with its usual authoritative tone of voice, answers what it thinks is the question, rather than detecting ungrammaticality and asking the user to re-phrase the question, something that it is perfectly capable of doing: see, for example, (19) below. In any case, (18a-b) hardly justifies Piantadosi's 'remarkable response' remark given earlier in (10).

(19)



One must agree with one portion of another one of Piantadosi's 'remarkable' remarks, shown in (20) below, viz. that one has to interact with LLMs to get a sense of their competency. Note Piantadosi's use of the term 'like talking to a child' is perhaps misplaced, given the context of this discussion. Certainly, from what we know, a child demonstrates substantial syntactic competency from an early age.[23]

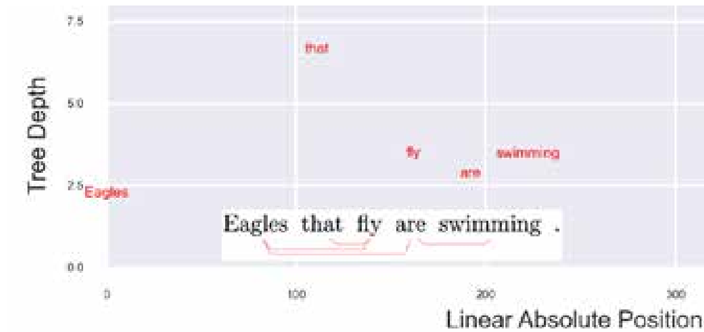
(20) It is somewhat difficult to convey *how remarkable* the models are currently. You just have to interact with them. They are imperfect, to be sure, but my qualitative experience interacting with them is like talking to a child … (Piantadosi 2023)

Syntax is necessary as the 'basic property of language', mentioned earlier (and demonstrated with Chomsky's examples), is that relations are fundamentally structural. Indeed, Piantadosi cites Hewitt & Manning's (2019) work, henceforth H&M, in regard to (21).

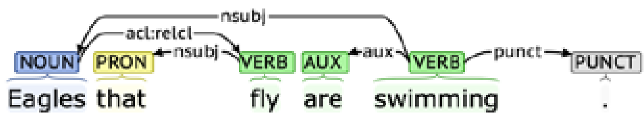
(21) These models discover structure – including hierarchical structure – from their training (Piantadosi 2023)

Let us consider whether this is true for the simple grammatical cases (16a) and (16c), presented to ChatGPT above. (22a) below gives the H&M depth map and structure obtained for the declarative sentence (16a). The H&M unlabeled dependency structure (computed from the depth map) reasonably corresponds to Stanza's labeled dependency structure, given in (22b), though punctuation (.) is unconnected.[24]

(22) a.

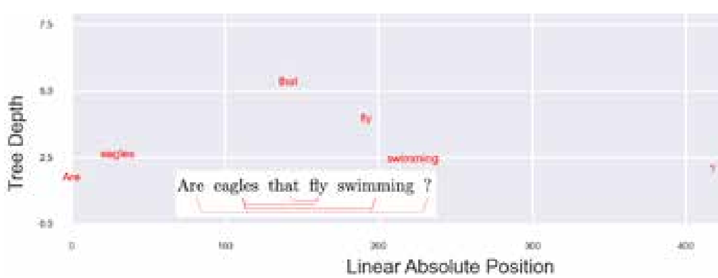


b.

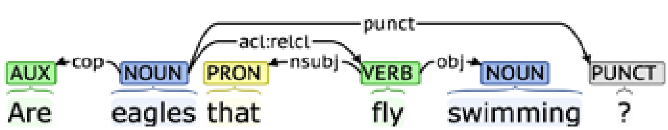


Let us now turn to sentence (16c), the yes/no interrogative counterpart of (16a). We can ask whether H&M's method assigns a reasonable unlabeled dependency structure in this case. The result is a rather unconvincing (23a). The fronted auxiliary *be* is connected only to the punctuation (?), and not to the head verb *swimming*. There also seems to be a double link between *eagles* and *swimming*. More alarming is the misparse from 'state-of-the-art' Stanza, shown in (23b). Here, swimming is analyzed as the object of the embedded verb *fly*.[25]

(23) a.



b.



Returning to the H&M analysis, we would be remiss if we did not point out the parse can be 'saved' by retroactively removing the question mark (?) from the input: this will map *be* and *swimming* together (being closest in computed tree depth). However, this mapping mistake is not generally present in the H&M model, compare (23a) with (24b) in the case of yes/no interrogative sentence (24a).

(24) a. Are humans conscious/alive?
b.

Adverb-verb construal is the other example of the basic property of language discussed earlier. The H&M model performs well with the examples in (6a-d), except with respect to the ambiguity in example (6a) that humans readily detect. However, (25a), in which *carefully* is rendered unambiguous (to the human reader) by the introduction of a sentence-initial *carefully*, is misanalyzed, as shown in (25b).

(25) a. Carefully, the mechanic who fixed the car carefully packed his tools
b.

(The second occurrence of *carefully* must modify *fix*.)

To summarize, I hope I have demonstrated (in the space permitted) that not only is structure clearly important for language analysis, but that structure is difficult to recover, even when using the tools that 'state-of-the-art' machine learning methods have provided. In other words, syntax remains an unsolved, or more accurately, a partially-solved, problem.

Piantadosi also misunderstands Chomsky's position on statistical modeling. To Chomsky, core language is I-Language, the part unaffected by experience (and therefore, statistics):

(26) Note that statistical information is irrelevant to I-language as a matter of principle, though as has always been assumed in the generative enterprise (see Chomsky 1957), it can be highly relevant to processing and acquisition. (Chomsky 2021, footnote 14)

Finally, I would like to address Chesi's (intentionally) pessimistic title *Is it the end of* (*generative*) *linguistics as we know it?*. Back in 2002, the prominent linguist Fukui asked Chomsky the following question:

(27) How do you characterize the development of linguistic theory in the last 20 years? Has there been what might be called “the third conceptual shift” in the field since the original The Generative Enterprise interview took place some twenty years ago? What are the continuities and departures from earlier works in generative grammar? (Fukui, in Chomsky 2004)

Chomsky’s reply, in (28), illustrates not only the scope of recent work, but also just how narrowly Piantadosi has construed the generative enterprise in his paper.[26]

(28) I think the most dramatic change that’s taken place in the past twenty years is not really in linguistic theory, but just in the scope and depth of linguistic work. There’s been just an enormous explosion of work on the widest imaginable range of typologically variable languages of every conceivable type in very much greater depth than ever before. Also [there have been] explorations in new areas, like a lot more work on formal semantics and its various aspects, a lot of work on properties of prosodic systems and metrical systems and others that’s new. The field’s just exploded in scale, quite apart from theoretical changes. (Chomsky 2004)

Recent theoretical developments since the 2002 interview has placed research on a new and exciting footing. In particular, the Strong Minimalist Thesis (SMT), as discussed in Chomsky 2021 and Chomsky 2024, has sharpened and narrowed the scope of the Minimalist Program considerably. For example, structure-building is now severely constrained, ideally limited to the simplest possible formulation of phrasal construction, viz. Simplest Merge, with evolutionary considerations in mind. Maximal simplicity means computational formulation becomes trivial (combinatorics aside). Principles of nature and computational efficiency have also come to the fore, as part of the so-called ‘Third Factor’ considerations (Chomsky 2005), the ‘First’ and ‘Second Factors’ being genetic endowment and experience, respectively. Maximal efficiency now plays a role in determing the scope of possible I-Language operations such as Minimal Search, which underpins how structure must be read. For syntax, earlier accounts for syntactic puzzles must pass the stringent SMT test, and new solutions must be found. This is by no means unusual as theories evolve, and is, in fact, in accordance with standard scientific practice. With respect to parsing, Chomsky believes Nature has optimized I-Language for thought, not communication.

(29) If that makes expressions harder to process and even makes some thoughts impossible to express without circumlocution, too bad. Nature doesn’t care. (Chomsky 2021)

To be clear, this does not mean communication is impossible, it is a fact that modern humans can effectively parse and interpret E-language.

(See note 5 for the distinction between I- and E-Language.) Parsing under SMT guidelines and practical implementation both depend on taming the combinatorics of Merge for perception. In summary, there has been considerable progress made through the years, current work is exciting, and I geninuely believe the future is promising, though, of course, as always, despite the ever-present lure of shortcuts promised by machine learning, much work remains to be done.

Abbreviations

H&M = Hewitt & Manning (2019); LLMs = Large Language Models; MS = Minimal Search; PCFG = Probabilistic Context-Free Grammar; SMT = Strong Minimalist Thesis.

Notes

1 Although basically considered a trade secret, we know the environmental impact and power requirements of the latest models are simply staggering. For example, it has been reported that xAI in Memphis TN requires 70MW to run 100,000 GPUs concurrently, power that the local utility was initially unable to supply. As a result, 18 natural gas generators (2.5MW each) had to be parked outside, worsening the local smog (Kerr 2024). Similarly, Google has ordered six or seven small nuclear reactors also for its AI datacenters according to Lawson (2024).

2 Various sources estimate the human brain consumes about 20W, not all of which is available to language. See references cited in Ling (2001).

3 For example, according to Nvidia *n.d.*, Meta's Llama 3.1 LLM (405B parameters) took nearly 31 million GPU hours of compute time. 31 million hours is approximately 35.4 centuries. On a human scale, 3,500 years ago would be around the time of construction of Cleopatra's Needles in Egypt.

4 For example, Hosseini *et al.* (2024) estimate that the human child hears 100 million words by the age of 10, therefore, it is reasonable to train models limited by this 'developmentally plausible' yardstick. However, as Chomsky has pointed out, the child does not have to learn to compute using structure, see note 6. See also the discussion on training data size in Chesi *this issue*.

5 We must be careful to distinguish between I-Language (the core system) and externalized E-Language when we say 'unavailable for language'. In Chomsky's view, externalization is the bridge between the language organ and sensorimotor system, a more ancient system in evolutionary terms. Linearity is imposed by the mapping to the primitive sound subsystem.

6 Chomsky (2021) cites the relevant figure as 30 months of age.

7 For structural simplicity, we assume here the *of*-insertion rule of Chomsky (1986) and that determiners represent spellout of certain features, e.g. definiteness and φ-features, of the head noun.

8 Search crucially does not involving finding two items, say α and β, then doing a comparison, e.g. $\alpha < \beta$, $<$ an ordering operator, and picking the best one, e.g. α. Comparison-based search, the basis for computer sorting algorithms and optimality-

theoretic models, is more computationally expensive than Chomsky's simplest MS. Therefore, sorting is not part of the language toolkit.

9 <syntaxgym.org>.

10 For example, the reduced relative represented by the verb phrase *(VP) first devasted years ago* should modify *the city*, not *recent bombings*. Stanford's CoreNLP (<corenlp.run>) also makes the same mistake. Another mistake concerns the attachment of the prepositional phrase (PP) *in the previous war*.

11 The *Berkeley Neural Parser* is not alone in this behavior. Stanford's *Stanza* (<stanza.run>) also makes the same mistake. See also note 10. In fact, when examined closely, none of the parsers mentioned here, i.e. Berkeley, Stanza or CoreNLP, correctly decode (3e).

12 Parse in (8) is generated by the SMT Parser. See <sandiway.arizona.edu/smt-parser>.

13 Technically, we have in (8) that *carefully* attaches to a vP and therefore c-commands the relevant verb. Given MS see also note 8, the other verb is not even considered.

14 In (11), LP and LR stand for Labeled Precision and Recall, respectively. The (bracket) F-score, an overall metric, is the harmonic mean of LP and LR.

15 Indeed, they can be essentially 'novel' sentences, i.e. not generally noticed before. Notable examples highlighted in the literature to illuminate theoretical points include Icelandic transitive expletive constructions (not grammatical in English), and unattested data except in dialects of English, e.g. Belfast English in the case of relative clauses.

16 For example, see discussion of statistics and MG in Chesi *this issue*.

17 The *Stanford PCFG* parser is perhaps the rare beast that can be easily tweaked to report the top-*N* parses along with the logprob scores. See (13a-b).

18 For completeness, the reported logprob scores for the top-5 (to 6 sf.) are -76.5597, -77.1335, -77.1615, -77.1933 and -77.3231.

19 The curious reader is invited to try the systems mentioned in this paper online. By varying the sentences, it is possible to get all systems to agree and disagree in all possible combinations.

20 Or by inserting the dummy verb *do* when no appropriate auxiliary exists, as in *do eagles that fly swim*?

21 Not shown here, but Stanford's *Stanza* essentially computes the same parses as the *Berkeley Neural Parser*.

22 Stanford's CoreNLP fails to recognize a yes/no question in either case.

23 There is a large literature on the subject of syntactic knowledge and language acquisition. See, for example, Fisher's reply to Tomasello, a well-known critic of the innateness hypothesis, in Fisher (2002), and Lidz & Gleitman (2004) for some relevant discussion.

24 With the addition of some lexical knowledge, we can probably assign similar (to Stanza) labels to the unlabeled H&M structure.

25 Note that Stanford CoreNLP makes the same mistake, viz. *swimming* is positioned as the direct object of deeply embedded *fly*.

26 For instance, in the section *Syntax is integrated with semantics*, Piantadosi contrasts vector models with the Chomskyan framework, yet, for example, the term 'semantics' is mentioned 13 times in Chomsky (2021). Needless to say, basic syntactic topics such as structural ambiguity, scope and argument structure, are clearly relevant to both syntax and meaning. No generative linguist excludes meaning from the study of syntax.

Bibliographical References

See the unified list at the end of this issue.