What IS a Natural Language, so that Language Models could learn it (and cognitive scientists stayed sane)?

David J. Lobina
April 2023

The hype surrounding Large Language Models remains unbearable when it comes to the study of human cognition, no matter what I write in this Column about the issue – doesn't everyone read my posts? I certainly do sometimes.

Indeed, this is my fourth, successive post on the topic, having already made the points that Machine/Deep Learning approaches to Artificial Intelligence cannot be smart or sentient, that such approaches are not accounts of cognition anyway, and that when put to the test, LLMs don't actually behave like human beings at all (where? In order: here, here, and here).[1]

But, again, no matter. Some of the overall coverage on LLMs can certainly be ludicrous (a covenant so that future, sentient computer programs have their rights protected?), and even delirious (let's treat AI chatbots as we treat people, with radical love?), and this is without considering what some tech charlatans and politicians have said about these models. More to the point here, two recent articles from some cognitive scientists offer quite the bloated view regarding what LLMs can do and contribute to the study of language, and a discussion of where these scholars have gone wrong will, hopefully, make me sleep better at night.

One Pablo Contreras Kallens and two colleagues have it that LLMs constitute an existence proof (their choice of words) that the ability to produce grammatical language can be learned from exposure to data alone, without the need to postulate language-specific processes or even representations, with clear repercussions for cognitive science.[2]

And one Steven Piantadosi, in a wide-ranging (and widely raging) book chapter, claims that LLMs refute Chomsky's approach to language, and *in toto* no less, given that LLMs are *bona fide* (his choice of words) theories of language; these models have developed *sui generis* representations of key linguistic structures and dependencies, thereby capturing the basic dynamics of human language and constituting a clear victory for statistical learning in so doing (Contreras Kallens and co. get a tip of the hat here), and in any case Chomskyan accounts of language are not precise or formal enough, cannot be integrated with other fields of cognitive science, have not been empirically tested, and moreover…(oh, *piantala*).[3]

Piantadosi's paper was posted on *LingBuzz*, a repository of linguistic papers, just last month, and soon after a couple of responses were posted there too, with some of the material discussed in these papers applying to the Contreras Kallens article too. This short response points out that we cannot conclude, from the apparent fact that LLMs correlate with human behaviour, that such models are *ipso facto* theories of human capacities, as this would be a case of inappropriate causality, unfortunately too common an occurrence in the field.[4]

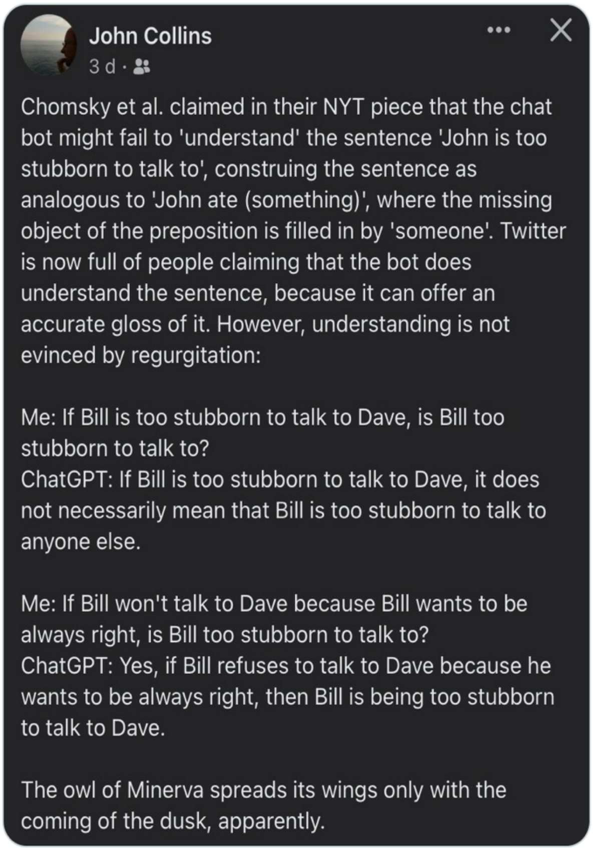
More tellingly, the linguist Roni Katzir, in a longer response, shows that LLMs cannot be said to have mastered the grammar of English, as they fail to detect some of the grammatical nuances that 12-year-olds and adults are unambiguously and effortlessly sensitive to – or said otherwise, that LLMs cannot be regarded as theories or models of human language because they do not capture some of the basic facts of our knowledge of language, in contrast to well-established theories in linguistics.

Katzir did so by providing ChatGPT with pairs of sentences and asking the chatbot to choose which sentence of each pair sounded better, just as I had done myself last month, obtaining a similar result – that is, the chatbot fails to correlate with human behaviour on various, uncontroversial judgements, and contrary to what the Contreras Kallens and Piantadosi papers both claim, the bot is certainly capable of producing ungrammatical sentences, even preferring ungrammatical sentences to grammatical ones in some cases – and reaching a similar conclusion – namely, that the chatbot doesn't demonstrate mastery of knowledge of language.[5]

I wouldn't want to exaggerate the significance of this approach, however; as stressed last month, I consider asking chatbots about the meaning of this or that sentence, or for grammatical judgements, to be quite absurd, for the simple reason that neither the chatbots nor the underlying LLMs have the right architecture to engage with such questions. *A fortiori*, some of the conclusions from the computational linguistics literature regarding what LLMs have learned about language, including hierarchical dependencies and the like, seem to me to be the result of projecting knowledge, and even mental states, to artificial networks that do not possess any of the relevant mental machinery, and on the basis of a methodology that rarely inspects an underlying LLM directly, in contrast to how connectionist networks from the 1980s and 90s were evaluated. Most scholars just ask the chatbots that sit on top of LLMs rather simple questions about this or that sentence, without probing too much – in this post, I briefly mentioned some of the relevant tests from the literature, which are very underwhelming – and this can only yield misleading views.

Still, it is hoped that the "grammatical exercises" Katzir and myself tested on ChatGPT – brief as both were, though somewhat more incisive than other tests from the literature – bring a bit of perspective to proceedings and perhaps even bring home the point that LLMs do not correlate with human behaviour on the linguistic front, not even in the weak sense of instantiating an input-output function (recall Pylyshyn's revenge).

Some cognitive scientists really should know better, though. At one point in his long tirade (31 pages, plus 16 pages of references), Piantadosi brings up a recent *New York Times* piece by Chomsky and colleagues in which they argue that because of how LLMs work, they are unlikely to properly interpret sentences such as *John is too stubborn to talk to*, only for Piantadosi to make light of this argument (he and numerous people on Twitter, actually) by stating that ChatGPT does understand this particular sentence if you ask the chatbot what it means. Unfortunately, Piantadosi misses the actual point by failing to keep in mind that knowledge of language is structured and intricate, and that core principles capture various facts about language, not least how a given sentence can be interpreted across different structural contexts.[6] Or as the philosopher John Collins aptly put it at the time:



John Collins
3 d ·

Chomsky et al. claimed in their NYT piece that the chat bot might fail to 'understand' the sentence 'John is too stubborn to talk to', construing the sentence as analogous to 'John ate (something)', where the missing object of the preposition is filled in by 'someone'. Twitter is now full of people claiming that the bot does understand the sentence, because it can offer an accurate gloss of it. However, understanding is not evinced by regurgitation:

Me: If Bill is too stubborn to talk to Dave, is Bill too stubborn to talk to?
ChatGPT: If Bill is too stubborn to talk to Dave, it does not necessarily mean that Bill is too stubborn to talk to anyone else.

Me: If Bill won't talk to Dave because Bill wants to be always right, is Bill too stubborn to talk to?
ChatGPT: Yes, if Bill refuses to talk to Dave because he wants to be always right, then Bill is being too stubborn to talk to Dave.

The owl of Minerva spreads its wings only with the coming of the dusk, apparently.

The put-down is quite right, and clearly deserved – the chatbot does behave like Chomsky et al. had hypothesised – but the point I want to make in this occasion is a much more fundamental and basic one: whatever LLMs have learned and mastered is not a natural language. And I mean this literally and absolutely. So, what is a natural language, then?

A typical characterisation will describe natural language as a system of communication (and/or of thought), composed of a vocabulary and a grammar, the latter a construct that includes, or ought to include, various components, in particular syntax, semantics, and phonetics and phonology. Indeed, a common way to define a natural language in linguistics, and possibly one that can be traced back to Ancient Greece,[7] is as a system that connects sound (or other externalisations of language, e.g., hand gestures or signs) and meaning – a thought-articulation pair, if you will (see the graphic heading this post, taken from this paper) – the connection between the two mediated by syntax. After all, phonological processes can only operate over a combination of words if syntax puts these words together, and the same goes for the semantic component in respect to the interpretation of a sentence.

Even though many scholars only have syntax in mind when talking of a "grammar", it should be stressed that for a given sentence to be grammatical, all the components I have mentioned need to align. In fact, I do not know of any theory for which this is not true, regardless of differences in focus, deviant sentences, and special cases. For some theories, syntax sends structures to phonology and semantics for further operations, and in this sense the syntactic engine is said to be constrained by various conditions that phonology and semantics impose, whilst for other theories syntax, semantics, and phonology operate in parallel, with various linking points during a derivation.[8] The overall picture yields all sorts of actual repercussions for the study of language (take note, Christiansen).

Consider some of the processes involved in how children acquire language, for instance. Children are sensitive to various phonological properties of the language or languages they are exposed to, basically from birth, and some of these features, such as intonation and stress, can help them identify categories such as determiners, work out where words start and end (in combination with some of the statistical patterns children can track), and even outline specific syntactic phrases (intonation can provide strong clues regarding when a syntactic phrase ends). Further, so-called semantic bootstrapping, according to which children can use semantic categories such as *objects* and *actions* as cues to syntactic categories such as *nouns* and *verbs*, respectively, can help children acquire the syntax of their language. And, in turn, so-called syntactic bootstrapping, which has it that children use "syntactic frames" to learn the meanings of words (a frame is like the argument structure that specifies that the verb *to ask* requires an *agent* doing the asking and a *topic*, or *theme*, to be asked about), can help children acquire the semantics of their language. *E così via.*

An intricate and systematic phenomenon, which results in an intricate and systematic knowledge of language, this state of affairs also applies, necessarily, to language comprehension, where intonation can help hearers resolve potential ambiguities in the input they receive during verbal communication, the meanings of words can activate potential follow-ups, canonical word orders can help predict what meaning is being communicated, *e così via*.[9] But this is just to restate the fact that by the grammar of a language, linguists mean, at the very least, syntax, semantics, and phonology.[10]

Given these facts about natural language, it is not a little misguided to claim that LLMs produce "grammatical" language; at one point in their paper, in fact, Contreras Kallens et al. explicitly state that they are not claiming that LLMs understand language, and this betrays not an insignificant disconnect in the entire approach. Knowledge of language and language understanding go hand in hand in human cognition (recall the John Collins quote, *supra*), and if your theory of language doesn't account for how we understand language, covering most if not all intricacies, then you do not have a theory of language at all.[11]

What language models produce is text, one word or token at a time, and for some rather specific purposes (engineering purposes, in fact), but it is actual humans who do the interpreting of the outputs these models produce and who identify them as meaningful and grammatical, though this really only makes sense from the perspective of the knowledge of language humans possess, and not from the viewpoint of what LLMs themselves do.

What Contreras Kallens et al. and Piantadosi mean by "language" cannot be a natural language, for no LLM has learned an actual natural language, but possibly, and I'm only speculating here based on past experience, "language" as this notion is often understood in formal language theory (and computer science) – namely, as a system that produces well-formed strings of "words". This is to reduce natural language to, as is often the case, string generation, and despite the decades-old protestation of myriad linguists that this is not all that relevant to the study of language, though it has certain uses (indeed, in LLMs, or in the speech recognisers Piantadosi mentions at the end of his paper; at least it ends with a good joke), this understanding of "language" is like a distant relative of natural language, at best, and more of a third cousin once removed than a brother, thus sharing rather little indeed.

What Contreas Kallens et al. should have claimed, thus, is that LLMs are an existence proof that the ability to produce human-like text can be achieved, not from exposure to data alone, but by exposing artificial networks that operate in parallel with loads and loads of attention layers to gigantic amounts data taken from the internet without attribution or consent, with little to no repercussions for cognitive science (and good luck to everyone claiming that natural language acquisition can be accounted for by exposure to data alone, and without language-specific processes or representations).

What Piantadosi should have written, then, is a very short chapter stating that LLMs are obviously not theories of language; such models certainly do not capture the basic dynamics of human language; they prove nothing regarding the statistical learning of natural languages; and they don't refute anyone's linguistic theory, Chomsky's or anyone else's (with a postscript confirming that the rest of the book has been put to the fire; *Stephen Hero* it won't be, I am sure, so no great loss here).

But the whole issue is moot, really. As I have stressed in this series of posts, in dealing with ChatGPT we are not evaluating the language model directly, which remains a closed system to outside researchers, in fact; we are probing the capabilities of the model via a chatbot that is adding quite a bit to the mix – the "prompt engineering" that directs the responses in a way that makes sense for a user, the "filters" that specify what a conversation or a story or a puzzle look like and how they ought to proceed, the reinforcement learning that aims at yielding appropriate and adequate responses, and possibly much more. A complex and impressive engineering feat, but clearly not a theoretical development to keep linguists (and cognitive scientists) honest.

[1] Once upon a time, I used to write about the role of common languages in fostering nationalist identities (including how Kiev has become Kyiv in the English-speaking world), the relationship between language and thought (including what James Joyce got wrong about the interior monologue, or inner speech), and Rudolf Rocker – much more interesting stuff.

[2] One of the authors of the article (a letter-article, actually), Morten Christiansen, a well-known connectionist, must have been part of dozens of papers exhibiting the formulation I have just described (as pointed in a previous post, by the way, connectionist networks are old-fashioned LLMs). To wit: 'this connectionist network offers an alternative to this theory of language [a Chomskyan theory, more often than not] and this has clear implications for our theories of [insert topic here]'. These papers would continue by describing networks that have been fed thousands of exemplars, oftentimes of infrequent structures that native speakers nonetheless master by early childhood, in order for the network to acquire a specific aspect of language, typically in isolation to other features of language, in contrast to how our knowledge of language is in fact organised. I read many of Christiansen's papers during my PhD half a generation ago or thereabouts, and without fail I would lose interest at the bit that said 'was fed thousands of sentences during training'.

[3] *Pianta la pianta e piantala, mala pianta*; apparently the word "piantadosi" may derive from "piantadoso", a word I didn't know and which means full of plants. More seriously, one dreads to think about what might follow Piantadosi's chapter; is a new edition of *The Anti-Chomsky Reader* about to come out? A truly astonishing paper, terribly tendentious, and even deranged at times (citing Behme is bad enough, but airing Postal's tantrums on *LingBuzz* is a bit much), I struggle to understand how someone as accomplished as Piantadosi – some of his work on Bayesian models of *the language of thought*, for instance, is excellent – could even consider making such a paper public at all – to my eyes, a quite embarrassing piece, and not a little humiliating. Starting with the passive-aggressive framing of Chomsky's influence in linguistics in terms of privilege (aren't we current with our vocabulary?), followed by page after page of citing all and any kind of paper that might be called upon for the cause, including studies that are not uncontroversial and have not gone uncontested in the literature (I can think of Evans & Levinson's papers on language universals, not to mention old Dan Everett's work on self-embedding), Piantadosi seems to be going the way of Geoffrey Pullum and Paul Postal – otherwise very good when attention to detail in methodology and analysis is required, but absolutely terrible when it comes to more conceptual stuff. One can only hope that he won't go full postal – embittered, not a little mad, and a contributor to a future edition of the Anti-Chomsky Reader.

[4] A case I have often encountered has to do with claims that some neuroimaging evidence force us to re-evaluate well-established cognitive theories and principles – this article, for instance, argues that neuroimaging data challenges the notion that there is a critical period to learn a native language, given that the same neural substrate is activated during first and second language learning (more

properly, the critical period hypothesis refers to the learning of a language *tout court*, though the general idea has also been applied to the difference in success and proficiency between learning a native language in infanthood and learning a second language in adulthood). But the principle that there is a critical period to learning a native language remains unchanged regardless of the neurological record – indeed, it remains a true fact of human development – and what needs to be established is why the neurological record is as it is given that there are reasons to believe it ought to be different (though the latter assumption needs to be taken with a grain of salt, considering that the fine-grained distinctions of many cognitive-science theories are not currently observable in the brain).

[5] I should add that in his paper Katzir surmises that pair examples from any textbook in linguistics would have worked in order to make the argument, and this is exactly the kind of data I used, which had furthermore been empirically corroborated. Katzir also does a good job of emphasising the difference between competence (knowledge of language) and performance (how this knowledge is put to use), as well as the distinction between theories of competence and theories of how language is acquired, where the latter domain usually allows for a family of processes and phenomena (e.g., statistical learning) that do not come under the purview of theories of competence per se.

[6] It could have been worse, I guess. Emily Bender, she of stochastic parrot fame, didn't particularly cover herself in glory by pointing to the bad AI coverage of *The New York Times* (and to the paper's supposed transphobia!) as a way to criticise Chomsky's piece, as if there were any direct relationship between the article and one's personal views about the NYT, only to direct readers to a profile of hers in the *New York Magazine* – not so much a 'read this piece of mine instead', but 'read about meeee instead' (the naked self-promotion, she's worse than me, or Dan Everett; well, maybe not the latter).

[7] Claude Panaccios's *Mental Language* offers some background on this, in addition to the orbiting question of how language may relate to *the language of thought*.

[8] A derivation is the linguist's term for how a sentence is "generated" by the language system, internally; the production of language would be the actual uttering or gesturing (or else) involved in externalising language.

[9] Learning a language and parsing a language are closely connected phenomena; Janet D. Fodor once wrote two papers on this, one called *Parsing to Learn*, the other *Learning to Parse*.

[10] As for the words of a language, it is more accurate to state that the primitives of grammar are lexical items – that is, bundles of syntactic, phonological, and semantic features. Not what your local LLM uses it, either, then.

[11] It is worth noting that most people working with LLMs do not actually claim that these models constitute theories or accounts of natural language, but engineering tools.