Adopting Large Language Models as a theory of language *does* refute Chomsky (but not like you think). A response to Ambridge & Blything (2024), Piantadosi (2023), and similar claims.

Charles Reiss
Concordia University
charles.reiss@concordia.ca

Veno Volenec
University of Zagreb & Concordia University
vvolenec@m.ffzg.hr

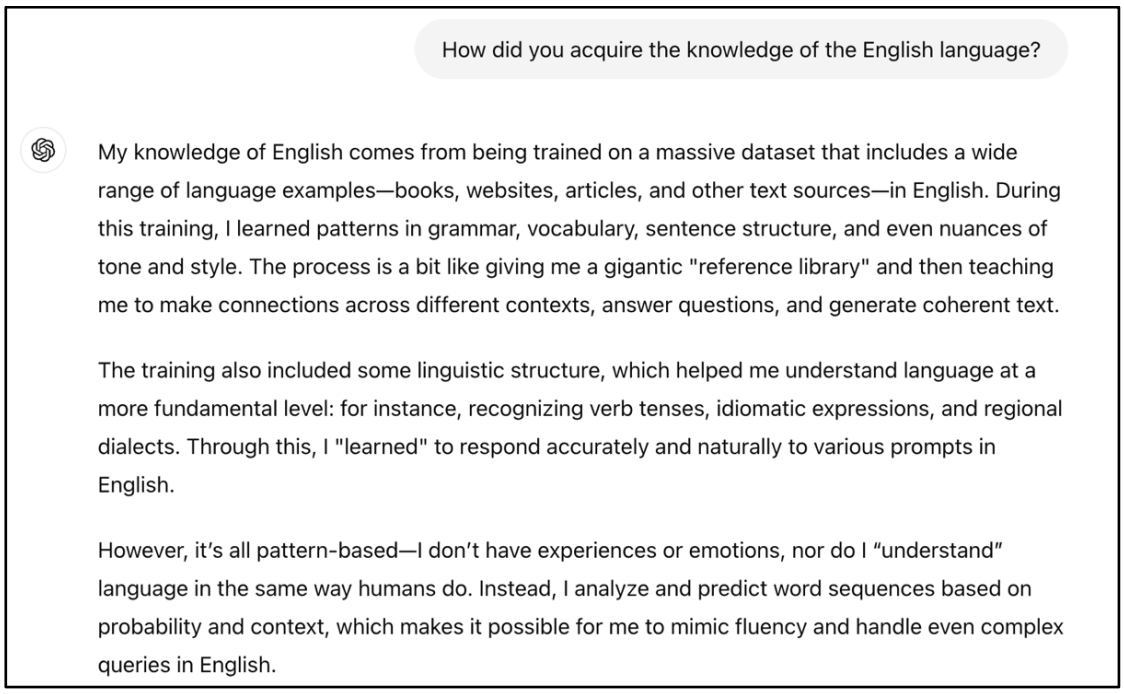
A recent paper with a self-acknowledged attention-seeking clickbait title — Ambridge & Blything (2024; henceforth A&B) — bombastically proclaims that "theoretical linguistics is dead" (p. 45). The alleged cause of its death are Large Language Models (LLMs) such as GPT-4o. LLMs are claimed to offer a better scientific theory of language "by a country mile", while "traditional linguistic theories don't come close" (p. 45). To prove this point, A&B examine how LLMs perform on acceptability judgement tasks that focus on "the only domain that [they] happen to know quite a bit about: verbs' argument structure privileges" (p. 44). Because LLMs perform as good as humans on this task, they conclude that LLMs explain language. Specifically, A&B take LLMs to be "the best currently available theories of speakers' representation and learning" (p. 34) of linguistic phenomena. The only linguistic approach that comes close to the scientific success of LLMs is the "exemplar-, input- and construction-based" approach (p. 33). In contrast, "traditional linguistic theories" (i.e., any theory except usage-based constructivism) are inferior because they "are not specified at anything close to the level of detail that would be required for them to make precise *quantitative predictions regarding the relative grammatical acceptability* of individual sentences" (p. 35; emphasis added). The image that comes to mind after reading the paper is that of two hunters standing over a meerkat they have just shot, proudly declaring that they have slain a notorious mighty lion. Hunters should at least know their prey.

The first and foremost problem with A&B is that they don't understand what "theoretical linguistics" of the sort they're criticizing is supposed to explain. (Whence the hunters' delusion.) Theoretical linguistics is *not* supposed to explain how people (let alone machines) perform on acceptability judgement tasks. So, the fact that, say, generative linguistics doesn't "make precise quantitative predictions regarding the relative grammatical acceptability of individual sentences" is a good thing. We don't want it to.[1] (We want it to capture grammaticality, not acceptability – a crucial distinction.) Theoretical linguistics *is* supposed to explain, however, how the human language faculty works. Humans use their language faculty, of course, on acceptability judgement tasks, but at the same time we use other cognitive, sensory and motor systems on such tasks, and these other

[1] We *do not* claim that quantitative research methods are not legitimate within generative linguistics, but that the theory itself shouldn't make quantitative predictions about the likelihood of some particular linguistic *behaviors*.

factors that contribute to how a person judges a sentence's acceptability are obviously not under the purview of theoretical linguistics. In other words, theoretical linguistics is about language knowledge (its content and acquisition), and not about the use of that knowledge on acceptability-measuring tasks. If this looks like we are alluding to the infamous, and to many people heretical, distinction between *linguistic competence* and *linguistic performance*, that is because we are. It is precisely because of their failure to understand this distinction that A&B's argument doesn't hold. The reasons for the necessity and validity of the competence-performance dichotomy have been stated *ad nauseam* by now (e.g., Chomsky 1965; 1980; 1986; Boeckx 2010; Isac & Reiss 2013; Smith & Allot 2016; Volenec & Reiss 2020; Firestone 2020; Dupre 2021) and are obvious given even the most basic observations of human behavior, so we won't repeat them here. We will just point out that an *infinite array of intractable random factors* influences linguistic performance, including performance on acceptability judgement tasks. It's as random as, say, a poet with a playful attitude toward language use performing much differently than a literal-minded person. Even the same speaker sometimes gives different judgements for the same sentence on different occasions (see Vaux & Cooper 2003: 116ff for details and examples). For that reason alone, one cannot draw *direct* conclusions about mental grammar based on acceptability judgements tasks. A *productive* approach is to use acceptability judgements as a *source of evidence* in the painstaking and indirect process of inferring the properties of mental grammar, carefully trying to control the variables and to disentangle the various contributing factors. An *unproductive* approach is to use acceptability judgements to argue that linguistic competence doesn't exist, that mental grammar (as opposed to behavior) is probabilistic, and that theoretical linguistics doesn't work. A&B's inflammatory statement that "large language models are better than theoretical linguists at theoretical linguistics" (p. 33) reveals a profound ignorance: they don't know what the object of inquiry of theoretical linguistics is, and therefore they don't know what they're criticizing.

The second reason why A&B are wrong in claiming that LLMs are better at explaining language than theoretical linguistics is that LLMs and human brains (where the actual object of study of linguistics is located) have nothing relevant in common (loose metaphors about 'neural networks' notwithstanding), especially in the context of how knowledge of language is acquired. Ironically, a very good argument for this comes from an LLM itself. Here is how ChatGPT-4o described its own 'language acquisition' (on October 28, 2024):



"Massive data-set", "gigantic reference library", "training", "teaching [it] to make connections", "[it does not] understand language in the same way humans do", "[it] mimics fluency". It even put scare quotes around 'learned'! It is clear that this is *nothing* like what goes on in the brain of a human infant. The human brain doesn't have access to a "gigantic reference library" in its infancy, and even if it did, it wouldn't be able to handle it. Notice also that LLMs acquire the ability to mimic human language fluency *solely* through reading, while *no child* has ever acquired language through reading (solely or even largely). Thus, LLMs' path toward the ability to produce human-like expressions is completely different from a child's path. A&B even describe how they optimized an LLM to perform better on acceptability judgement tasks: "Following a training session designed to familiarize the model with the task of rating sentences on a 5-point acceptability scale [...] the model is given (counterbalanced) prompts" (p. 38). What does that have to do with human language? Do A&B seriously think that babies undergo training sessions to acquire language? In line with A&B's reference to Popper's criterion of falsifiability (p. 34), even one baby that acquired language without explicit training is sufficient to refute A&B's 'theory' of language acquisition. In fact, every baby that ever existed refutes it. Humans engineered LLMs to be as human as possible in their overt communicative behavior; it is precisely in achieving that goal that LLMs have revealed how different they are because *their path to achieving it was profoundly inhuman*.

A&B did, however, get one important point right. They've admitted that even LLMs require *a priori*, *predetermined*, *in-built* principles and mechanisms that make the learning process possible upon exposure to data. In other words, they agree that even LLMs, which were trained on corpora containing "hundreds of billions of words" according ChatGPT's estimate, cannot function without some equivalent of Universal Grammar. In A&B's own words:

> ""We" – or at least the software engineers who built LLMs – made hundreds of decisions about the precise architecture and learning mechanisms that should be used. These engineers could have made different choices; and – depending on those choices – the models would have simulated human acceptability judgments either better or worse. These choices – fossilized in thousands of lines of computer code – are a theory of human language acquisition." (p. 42)

It is gratifying to see they've finally reached the conclusion that UG is inevitable after the misguided attempt to explain "why universal grammar doesn't help" in child language acquisition (Ambridge et al. 2014).[2]

It is also fitting that A&B would claim that "the construction-based approach is supported by findings from LLMs" (p. 37). As far as we can tell, the main finding from LLMs that is relevant for linguistics is that LLMs do not model human linguistic competence or its acquisition. Rather, LLMs give machines an ability to imitate human communicative behavior. So, if something that is not a theory of language lends support to the construction-based approach, what does that say about the construction-based approach?

Let us move on now to the syntactic phenomenon discussed in A&B's paper, namely the "domain that [they] know something about: learning and representing verbs' argument structure privilege" (p. 33). Unfortunately, they do not seem know much about that domain either, because their identification of the domain is sloppy and self-contradictory. The problem again derives from a lack of understanding of the object of study of theoretical linguistics, individual I-languages and the Human Language Faculty. Linguists often talk informally about 'languages' like English, and in that context, it is perfectly reasonable to talk about **the** verb *roll* in the sentences *The ball rolled* and *Someone rolled the ball* appearing with two different sets of arguments. But linguists also sometimes consider that the two verb forms might contain unpronounced differences, differences that are sometimes reflected in the pronunciation of such intransitive transitive pairs as *lie / lay*. Deciding whether the two printed words spelled *rolled* are actually type-identical is a legitimate question discussed by linguists (Borer 1994, 2004; Levin & Rappaport Hovav 2004), but we are concerned here with a much simpler issue.

First, let's agree that a verb, or any other morpheme, must be understood as having at least a tripartite structure. There must be a lexical meaning — expressing things like the distinction between, say, *rolling, spinning* and *crying.* There must also be a phonological representation — relating to things like the difference in pronunciation among *fleeing, flying* and *floating*. Finally, there must be a syntactic representation that indicates that a word is a verb, and perhaps, depending on one's theory, what its argument structure is, what arguments it takes. If our two tokens of *rolled* do not encode a

[2] Of course, the need to build 'priors' into machine learning systems is a truism, whether engineers explicitly acknowledge it or not (Versace et al. 2018; Rawski & Heinz 2019; Wolpert 2021). See Volenec & Reiss (2020) and Reiss & Volenec (2022) for a more detailed discussion.

difference in argument structure, there is a chance that they are type-identical. However, if they do encode such a difference, then by definition, they cannot be tokens of the same word. As we said, this is a legitimate issue that morphologists and syntacticians can discuss.

With that in mind, consider the following sentences of 'English':

(1) a. We are not allowed to run here.
b. We are not allowed running here.

(2) a. We are done with our homework.
b. We are done our homework.

Most people say that the (a) sentences are well-formed, and that is what ChatGPT-4o tells us, too. And most people, like ChatGPT-4o, will tell us that the strings in (b) are not well-formed, they are not sentences of English.

However, English speakers in Montreal, and many other speakers of Canadian English produce and accept the (b) examples as perfectly grammatical. To simplify our discussion, let's assume that the Canadian speakers use only the (b) forms and reject the (a) sentences. (They do not, probably because they have internalized more than one dialect – they are typically not aware that the (b) forms are not used by other English speakers.) For (1a) we might say that 'the past participle verb *allowed* occurs with an argument structure that has an infinitival complement,' whereas for (1b) we might say that 'the past participle verb *allowed* occurs with an argument structure that has a gerund complement'.[3] And in (2a) we might say that 'the verb *done* takes a prepositional phrase complement', whereas in (2b) 'the verb *done* takes a noun phrase complement'. However, in both cases, we would be talking nonsense. Unlike the case of *rolled* above, there is not even a remote possibility that we are talking about **the** past participle of **the** verb *allow* or **the** verb *do.*

In each case, a Canadian speaker has a morpheme that has a similar or identical semantic component as a morpheme of an American speaker, and the two have similar or identical phonological representations for the corresponding morphemes. However, the difference in argument structure in each case must be distinctly encoded. Canadian *allow* has different properties from American *allow*, so by Leibniz's Law concerning the Identity of Indiscernibles, the two must be different.

[3] The argument structure of a verb uses primitives such as 'external argument' and 'internal argument'. Arguments are typically nominal, but they could also be clausal. Whether an internal argument is nominal or clausal is captured in the subcategorization feature of that verb, not in the argument structure. Even if we assume a laxer version, in which the argument structure of a verb does specify whether that argument is clausal or not, the argument structure will certainly *not* specify morpho-syntactic differences between various types of nominals or clauses, i.e., whether an argument is a definite or indefinite nominal constituent, or whether an argument is a finite or non-finite clause, or what type of non-finite clause it is. This is captured in the selectional or subcategorization feature of that verb.

Similarly, within a single speaker's lexicon, *couch* and *sofa* share a common syntax and semantics, but they cannot correspond to the same morpheme because they differ phonologically. By the same reasoning, the word spelled *dog* by the two of us cannot be tokens of the same morpheme across our lexicons, because for one of us, *dog* rhymes with *morgue* but not for the other: they differ phonologically. If *any* aspects of our morphemes differ, they cannot be the same. We need to distinguish everyday talk that we all engage in — where we might say that English *mutton* and French *mouton*, or English *hound* and German *Hund* are 'the same word' — from scientific discourse that treats language as a branch of cognitive psychology (and ultimately biology of the brain). Despite everyday informal talk, only the 'Chomskyan' generative I-language approach, which sees 'language' as a mind-**i**nternal, **i**ndividual, **i**ntensional unconscious knowledge, can handle such facts. LLMs are fed *written* sources that obscure pronunciation differences as well as many, many other differences like the Canadian vs. 'Standard English' examples above. Human learners acquire lexicons and grammars that reflect the data to which they are exposed. An LLM might assign a string like *We are not allowed running here* a very low probability of occurrence, but that does not reflect what is actually going on.

In brief, A&B tacitly adopt a mystical pseudo-platonic idea that a verb exists out in the world, outside of human minds (for them it is located on paper or in computer files), and thus demonstrate that they do not know much about how people learn and represent argument structure privileges after all. Strings of text on paper or in computer files are related to the morphemes in human minds in an exceedingly indirect manner. A&B can't be better than theoretical linguists at analyzing language because they don't know what language is or even what a verb is for a linguist.

The problems reflected in A&B's paper were already dealt with in Chomsky's (1986) *Knowledge of Language* and elsewhere. In addition to falling into the platonic P-language trap that Chomsky identifies, they also fall prey to the confusion of the E-language conception that takes external artifacts like corpora or communicative behavior to be the object of study of linguistics. A further variant of E-language that Chomsky discusses is the socio-political notion of a language, one related to identity, history, tradition and political considerations. A&B, and virtually all work that relies on corpora falls prey to some basic errors. How do A&B decide what counts as English? Should we include Trump's speeches, Shakespeare's sonnets, Chaucer's *Canterbury Tales, Beowulf* and so on? The findings of theoretical linguistics are not affected by the outcome of the civil war that broke up Yugoslavia — we never believed that an entity called 'Serbo-Croatian' existed in the world, but rather, we believed in a bunch of more or less similar I-languages. LLM researchers have to make an arbitrary decision on whether to train a model on data that is called, say, 'Croatian', or additionally on data that is called 'Serbian'. Whatever decision is made will generate a set of probabilities that has no bearing on how the mental grammars of the individual speakers in those regions work.

A&B misleadingly cite a paper by one of us: "In phonology, for example, the once-mainstream idea of a universal set of phonological features (analogous to the categories assumed in

the domain of verb argument structure) is 'very much a minority position today, even among phonologists trained in the generative tradition' (Reiss 2023: 9)" (p. 44). In fact, the passage from Reiss (2023) continues thus: "For example, Reiss & Volenec (2022) is the sole contribution to a recent volume on phonological primes that adopts and defends the nativist position for features expressed by Chomsky and Halle." Aside from clarifying the point that a minority position is not necessarily the wrong position (if it were, everyone with a new idea would necessarily be wrong!), we argued that the *only* coherent way in which two linguistic forms can be considered as type-identical (phonologically, semantically, and/or syntactically) is in terms of a universal, innate set of representational primitives (e.g., phonological features). No one who does phonology invents features from scratch for every language or for every analysis, as pointed out by Chomsky & Halle (1965). A&B's reliance on spelling and arbitrarily delimited corpora is a non-starter for scientific inquiry.

In the generative linguistics literature, it is a commonplace that all entities, such as particular verbs, syntactic categories, phonological segments, syllables, etc., are mental constructs of individual I-languages. Even in the more 'concrete' domains of phonetics and phonology we find statements such as: "it should be perfectly obvious by now that segments do not exist outside the human mind" (Hammarberg 1976). This view extends to all levels of linguistics analysis: "Language, as far as I can tell, is all [mental] construction." (Jackendoff 1992: 164). And notably, as Chomsky puts it:

> "No one is so deluded as to believe that there is a mind-independent object corresponding to the internal syllable [ba], some construction from motion of molecules perhaps, which is selected when I say [ba] and when you hear it" (Chomsky 2015: 126).

Ambridge and Blything's analysis, which is fundamentally dependent on the existence of mind-independent verbs with argument structure, gives the lie to Chomsky's claim. To refer to the title of Piantadosi's (2023) recent viral paper, modern language models do indeed refute at least one of Chomsky's views: such delusions do exist.

Acknowledgements

We are very grateful to Gabe Dupre, Dana Isac, Elizabeth Tobyn and Jon Rawski for their valuable feedback on earlier drafts of the paper.

References

Ambridge, B., Blything, L. (2024). Large language models are better than theoretical linguists at theoretical linguistics. *Theoretical Linguistics* 50(1–2), 33–48.

Ambridge, B., Pine, J. M., Lieven, E. V. (2014). Child language acquisition: Why universal grammar doesn't help. *Language* 90(3), 53–90.

Boeckx, C. (2009). *Language in Cognition. Uncovering Mental Structures and the Rules Behind Them.* Malden: Wiley-Blackwell.

Borer, H. (1994). The projection of arguments. *University of Massachusetts occasional papers in linguistics* 17(20), 19–48.

Borer, H. (2004). The grammar machine. *The unaccusativity puzzle: Explorations of the syntax-lexicon interface* 5, 288–331.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge: MIT Press.

Chomsky, N. (1980). *Rules and Representations*. Oxford: Basil Blackwell.

Chomsky, N. (1986). *Knowledge of Language. Its Nature, Origins, and Use.* New York: Praeger.

Chomsky, N. (2015). *What kind of creatures are we?*. New York: Columbia University Press.

Chomsky, N., Halle, M. (1965). Some Controversial Question in Phonological Theory. *Journal of Linguistics* 1, 97–138.

Dupre, G. (2021). (What) Can deep learning contribute to theoretical linguistics?. *Minds and Machines* 31(4), 617–635.

Firestone, C. (2020). Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences U.S.A.* 117(43), 26562–26571.

Hammarberg, R. (1976). The metaphysics of coarticulation. *Journal of Phonetics* 4, 353–363.

Isac, D., Reiss, C. (2013). *I-Language. An Introduction to Linguistics as Cognitive Science*. Second edition. Oxford: Oxford University Press.

Jackendoff, R. (1992). *Semantic Structures.* Cambridge: MIT Press.

Levin, B., Rappaport Hovav, M. (2005). *Argument Realization.* Cambridge: Cambridge University Press.

Piantadosi, S. (2023). Modern language models refute Chomsky's approach to language. Lingbuzz Preprint: [https://lingbuzz.net/lingbuzz/007180].

Rawski, J., Heinz, J. (2019). No free lunch in linguistics or machine learning: Response to Pater. *Language* 95(1), 125–135.

Reiss, C. (2023). Research methods in Armchair linguistics. Lingbuzz Preprint: [https://lingbuzz.net/lingbuzz/007568].

Reiss, C., Volenec, V. (2022). Conquer primal fear: Phonological features are innate and substance-free. *Canadian Journal of Linguistics* 67(4), 581–610.

Smith, N., Allott, N. (2016). *Chomsky: Ideas and Ideals*. Cambridge: Cambridge University Press.

Vaux, B., Cooper, J. (2003). *Linguistic Field Methods*. Eugene: Wipf and Stock Publishers.

Volenec, V., Reiss, C. (2020). Formal Generative Phonology. *Radical: A Journal of Phonology* 2, 1–65.

Versace, E., Martinho-Truswell, A., Kacelnik, A., Vallortigara, G. (2018). Priors in animal and artificial intelligence: where does learning begin? *Trends in cognitive sciences* 22/11, 963–965.

Wolpert, D. H. (2021). What is important about the no free lunch theorems?. In *Black box optimization, machine learning, and no-free lunch theorems*, 373–388). Springer International Publishing.