Ben Ambridge* and Liam Blything

Large language models are better than theoretical linguists at theoretical linguistics

https://doi.org/10.1515/tl-2024-2002

Abstract: Large language models are better than theoretical linguists at theoretical linguistics, at least in the domain of verb argument structure; explaining why (for example), we can say both *The ball rolled* and *Someone rolled the ball*, but not both *The man laughed* and **Someone laughed the man*. Verbal accounts of this phenomenon either do not make precise quantitative predictions at all, or do so only with the help of ancillary assumptions and by-hand data processing. Large language models, on the other hand (taking text-davinci-002 as an example), predict human acceptability ratings for these types of sentences with correlations of around $r = 0.9$, and themselves constitute theories of language acquisition and representation; theories that instantiate exemplar-, input- and construction-based approaches, though only very loosely. Indeed, large language models succeed where these verbal (i.e., non-computational) linguistic theories fail, precisely because the latter insist – in the service of intuitive interpretability – on simple yet empirically inadequate (over)generalizations.

Keywords: large language models; causatives; grammaticality judgments

Sorry about the clickbait title. But now that we've got your attention, large language models are better than theoretical linguists at theoretical linguistics; at least in the domain that we know something about: learning and representing verbs' argument structure privileges. The phenomenon is that while – for example – some English verbs can occur in both the intransitive-inchoative (1) and transitive-causative construction (2), others can appear in the former (3) but not the latter (4), and instead form only periphrastic causatives with *make* (5).

Some of the original research described here received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no 681296: CLASS). Ben Ambridge is Professor in the International Centre for Language and Communicative Development (LuCiD) at The University of Manchester. The support of the Economic and Social Research Council [ES/L008955/1] is gratefully acknowledged.

***Corresponding author: Ben Ambridge**, University of Manchester, Manchester, UK; and ESRC International Centre for Language and Communicative Development (LuCiD),
E-mail: ben.ambridge@manchester.ac.uk
Liam Blything, University of Manchester, Manchester, UK; and ESRC International Centre for Language and Communicative Development (LuCiD)

(1) *The ball rolled*

(2) *Someone rolled the ball*

(3) *The man laughed*

(4) **Someone laughed the man*

(5) *Someone made the man laugh*

While our focus here is on **English causatives** (e.g., Ambridge and Ambridge 2020; Ambridge et al. 2008, 2009, 2011; Bidgood et al. 2021), similar (as we will see, probabilistic) restrictions hold for causative constructions across at least 40 other languages (Haspelmath 1993; Levshina 2016; Shibatani and Pardeshi 2002). In both English and other languages, similar (probabilistic) restrictions exist for, among other constructions, the **dative/distransitive** (double-object/prepositional) (Ambridge et al. 2012a, 2014), the **figure-/ground-locative** (Ambridge et al. 2012b; Bidgood et al. 2014; Twomey et al. 2014, 2016, for English; Ambridge and Brandt 2013, German; Aguado-Orea et al. 2016, Spanish), **verbal *un*-prefixation** (Ambridge 2013; R. Blything et al. 2014) and **passives** (Aryawibawa and Ambridge 2018, Indonesian; Ambridge et al. 2023, Hebrew; Darmasetiyawan and Ambridge 2022, Balinese; Liu and Ambridge 2021, Mandarin; Ambridge et al. 2016; Bidgood et al. 2020; Ambridge et al. 2021, English).

Our argument is that large language models (LLMs) are already the leading current theories of how speakers learn and represent these restrictions. Of course, they are not *perfect* theories – far from it – but they're *better* theories than any others that have been proposed.

First, just what is a theory, and what makes one theory better than another? There is little agreement on this question between philosophers of science (for an accessible summary, see https://plato.stanford.edu/entries/science-theory-observation/). But hopefully at least most of us can agree that a good theory (a) explains all (most?) of the relevant facts already observed, (b) predicts what we will see for new cases, (c) generates predictions that are testable (even, for Popperians, falsifiable), and (d) does so using *theoretical constructs*; entities whose meaning can be defined only with reference to other parts of the theory and which are (usually?) not directly observable (e.g., "force" in Newtonian physics; "merge" or "construction" in theoretical linguistics). By this – hopefully not-too-controversial – metric, LLMs are, we will argue, the best currently available theories of speakers' representation and learning of verbs' argument structure privileges and restrictions.

What are the rival theories? Under Transformational Grammar/Government and Binding approaches (e.g., Chomsky 1965, 1981), each item in the lexicon (e.g., *roll, laugh*) is categorized for its syntactic class (here, verb), and subcategorized for its valence (e.g., intransitive and transitive for *roll,* intransitive for *laugh*). These

subcategorization frames are also known as valence frames or thematic/theta-θ-grids. Similarly, under minimalism (based on the analysis in Adger 2003: 87–90) these privileges are captured by verbs' syntactic (c-selection) and semantic (s-selection) features. Under Head Driven Phrase Structure Grammar (based on the description in Müller 2023: 272) each verb's lexical entry contains (specifier and) complement valence features (e.g., ⟨NP[nom]⟩ for *laugh*; ⟨NP[nom], NP[acc]⟩ for *roll*). Though the representation is different, the situation is similar for Categorial Grammar (again, based on the description in Müller 2023), in which valence is again encoded in the lexicon (e.g., "vp/np" for a transitive verb; where "vp/np stands for something that needs an np in order for it to form a vp" (p. 248). Indeed, similar assumptions hold for all lexicalist approaches, including Lexical Functional Grammar (see Müller and Wechsler 2014a, 2014b, and other papers in *Theoretical Linguistics* Volume 40 Issue 1–2 for a summary and debate).

How do these theories stack up against the informal criteria set out above? Pretty well. They explain all of the relevant facts, at least at a broad-brush level (e.g., that transitive uses of *laugh* are ungrammatical in a binary sense), they generalize to new cases (e.g., uses of *laugh* with previously-unattested NP arguments), they generate testable new predictions (e.g., that two-argument uses of *laugh* with previously-unattested NPs will be ungrammatical) and do so using non-observable theoretical constructs (e.g., "valence", "lexical entry"). Neither is it necessarily the case that these theories fail to explain probabilistic acceptability judgments. As Müller and Wechsler (2014b: p. 211) note, lexical rules can (and, in their view, should) contain a probabilistic component. There also exist well-worked-out proposals for determining the valence of unfamiliar or novel verbs on the basis of their semantics (e.g., Pinker's 1989, semantic verb class account, which assumes a version of Lexical Functional Grammar).

The problem (and the reason that we described them as doing only "pretty well" is that current lexicalist approaches 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.

Construction-grammar approaches – in particular, those set out by Goldberg (1995, 2006, 2011, 2019) – fare better here, in that they include a detailed mechanism that accounts for probabilistic degrees of (un)acceptability (Goldberg 2011: 135):

(6) The probability of CxB statistically preempting CxA for a particular verb, $verb_i$:
P(CxB | a discourse context in which the learner might expect to hear CxA [$verb_i$])
This probability is equivalent to the following:
P(CxB | a discourse context at least as suitable for CxA, and $verb_i$)

For example, the probability of the periphrastic causative construction (CxB) statistically pre-empting the transitive causative construction (CxA) for the verb *laugh* (e.g., *Someone made the man laugh* pre-empting **Someone laughed the man*) is the probability of *X made Y laugh* in a discourse context in which a learner might expect to hear **X laughed Y*. A simple proxy for the number of suitable discourse contexts is the total input frequency of *laugh* in either the transitive- or periphrastic-causative. Thus, statistical preemption predicts that the greater the frequency of *X made Y laugh* relative to **X laughed Y*, the lower the acceptability of the latter in a judgment task.

Indeed, quite a few studies (e.g., Ambridge et al. 2015, 2018; Boyd and Goldberg 2011; Goldberg 2011; Perek and Goldberg 2017; Robenalt and Goldberg 2015) have found support for this prediction: the greater the frequency with which a verb appears in the periphrastic- versus transitive-causative in a relevant corpus, the lower the relative acceptability of that verb in the transitive-causative in a graded judgment task (and likewise for other constructions).

Yet even the preemption account does not, *on its own*, yield quantitative predictions of sentence acceptability. This is not to single out preemption (or the construction-based approach more generally) for criticism. Indeed, in specifying an equation for its calculation, preemption is unusually *well*-specified for a verbal theory (as opposed to a computational model). Yet, as with all verbal theories, deriving these (ultimately well-supported) predictions from the preemption account requires a number of ancillary assumptions. For a start, exactly what mechanism is "calculating" preemption, and using what equation? (Goldberg 2011, uses simple probability, plus a confidence measure based on natural log frequency; Stefanowitsch 2008, uses Fisher's exact test; Ambridge et al. 2018, 2020, 2022, use a log-transformed chi-square statistic). More fundamentally, how exactly do we know what counts as the use of the relevant verb in the relevant construction? Do all transitive uses count (including, for example, cognate objects, as in 7), or only transitive *causative* ones? What about resultatives (8), and does it make a difference if the object is reflexive (9)?

(7) *She laughed a hearty laugh*

(8) *She laughed him out of town*

(9) *She laughed herself stupid*

In general, construction grammar approaches assume that these more specific transitive constructions are daughters of a more general transitive construction (e.g., Diessel 2023), but it is not clear how to factor this notion of inheritance into the preemption calculations. Similarly, for periphrastic causatives, do we count only uses with canonical word order, or (given that they inherit certain properties from active constructions) passives too?

(10) *She was made to laugh* (*by someone*)

The situation is even more complex for verbs that have multiple meanings. Clear-cut homophones (11)/(12) presumably do not contribute to verb-in-construction counts.

(11) *She decided to ring him to pass on the good news*

(12) *The general decided the army needed to ring the city*

But what about polysemous verbs, which have multiple *related* meanings? How many of the following count as transitive (-causative) uses of *drop* when calculating preemption?

(13) *She dropped the ball*

(14) *She dropped the price*

(15) *She dropped her guard*

(16) *She dropped the player from the squad*

(17) *She dropped* (*him*) *a hint* [do ditransitives contribute to transitive counts?]

(18) *She dropped a new video on her channel*

How do we factor in the fact that the identity of the NPs seems to matter too? For the transitive causative, participants give higher ratings (Ambridge et al. 2009) to sentences with more direct causers (19) > (20)

(19) *The magician's spell vanished Lisa* >

(20) *The magician vanished Lisa*

And what about the wider context? The following sentence (21) is more acceptable if it is clear that *disappear* is being used as a synonym for "kill", rather than, for example, "usher out of the room"

(21) *The dictator disappeared his enemies*

Again, none of this is intended to criticize preemption in particular, which is specified in more detail than probably any of the rival verbal accounts in this domain; and, as we will see shortly, in very general terms, the construction-based approach is supported by findings from LLMs. The shortcoming is not of preemption or constructions, but of verbal (i.e., non-computational) linguistic theories in general: It's just not possible to specify quantitatively all of the relevant factors, and how they interact.

Indeed, *simple* computational models (as opposed to complex state-of-the-art LLMs) fare little better than verbal theories. I (the first author) have built a number of

simple computational models that make graded predictions of the acceptability of particular verbs in particular constructions (Ambridge and Blything 2016, for English datives; Ambridge et al. 2020, 2022, for causatives in English, Hebrew, Hindi, Japanese and K'iche; Ambridge, Aryawibawa et al. 2021, for Balinese causatives; Liu and Ambridge 2021, for Mandarin passives). The models were successful, in that their predictions correlated well with human acceptability judgments. But to get them up and running, we had to set up the models with pre-existing knowledge of both the relevant constructions and the construction-relevant (and no other!) semantic properties of each verb. And far from leaving the models to sort out the messy details of what "counts" as a use of a particular verb in a particular construction (see discussion above), we created the (toy) training sets using by-hand counts from relevant corpora.

In summary, then, previous accounts of how speakers learn and represent verbs' argument structure privileges either (a) don't make graded quantitative predictions of human acceptability judgments at all or (b) make these predictions only once the experimenter has made a lot of ancillary assumptions that aren't part and parcel of the theory, and engaged in a lot of by-hand data-pre-processing.

Large language models, in contrast, just work. As part of a larger project investigating the ability of LLMs to simulate human grammatical acceptability judgments (based on Mahowald 2023) we asked GPT 3 text-davinci-002 to rate the grammatical acceptability of the English transitive- and periphrastic causative sentences from Ambridge et al. (2020). Following a training session designed to familiarize the model with the task of rating sentences on a 5-point acceptability scale (details available at https://osf.io/ctvqy/), the model is given (counterbalanced) prompts such as the following:

Sentences A and B below are descriptions of the same event.
A: *Someone barked the dog*.
B: *Someone made the dog bark*.
Question: How grammatically acceptable is sentence A [or, on another trial, sentence B] as a description of the scene?

Figure 1 shows, for each verb, preference for (negative values) periphrastic over transitive causatives (e.g., for *Someone made the man laugh* over *Someone laughed the man*), on the five-point scale (or, for positive values, the reverse), for text-davinci-002 (Y axis) and adult raters from Ambridge et al. (2020) (X axis). The correlation between the human and model ratings is $r = 0.92$, $p < 0.001$ (95 % CI [0.87, 0.95]), with a rank order correlation (Kendall's tau) of $\tau = 0.80$, $p < 0.001$.

These "difference scores" (i.e., preference for periphrastic- over transitive-causative uses of each verb) allow for arguably the cleanest test of the model, since

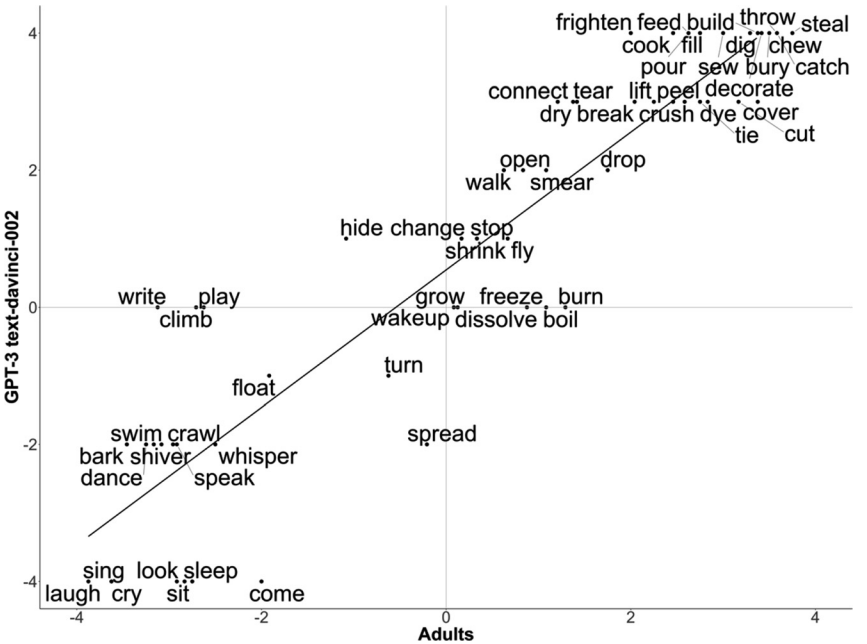

Figure 1: Preference for (negative values) periphrastic over transitive causatives (or, for positive values, the reverse), for text-davinci-002 (Y axis) and adults from Ambridge et al. (2020) (X axis).

they control out any general dispreferences that the model and/or human raters may show for (for example) low-frequency words. But similar values are found if we instead run the correlations over raw ratings for periphrastic causatives (Figure 2; r = 0.84, p < 0.001 [95 % CI = 0.75, 0.90], τ = 0.65, p < 0.001) or transitive causatives (Figure 3; r = 0.90, p < 0.001 [95 % CI = 0.84, 0.94], τ = 0.67, p < 0.001).

OK so the model makes the right predictions[1] but – we hear you ask – where is the theory? That's the point: the model is the theory. Text-davinci-002, or any other large language model, **is** a theory of (amongst other things) the representation and acquisition of verb argument structure (e.g., Frank 2023a). "But", critics object, "we have no idea what it's doing" (e.g., Kodner et al. 2023;[2] Milway 2023). Quite the opposite: Unlike

1 Indeed, there may soon come a point at which the "acceptability judgments" of LLMs are taken as the gold standard, with human ratings subject to "error", or at least uncontrolled individual differences. In fact, quite possibly, we are at that point already: It is not clear why the grammatical intuitions of 48 University of Liverpool students (Ambridge et al. 2020) should take precedence over the distilled wisdom of millions of English speakers worldwide (davinci002).

2 Kodner et al. (2023) cite Ambridge (2020a; *Against stored abstractions: A radical exemplar model of language acquisition*) as one example of "linguistic theories based on emergence and self-

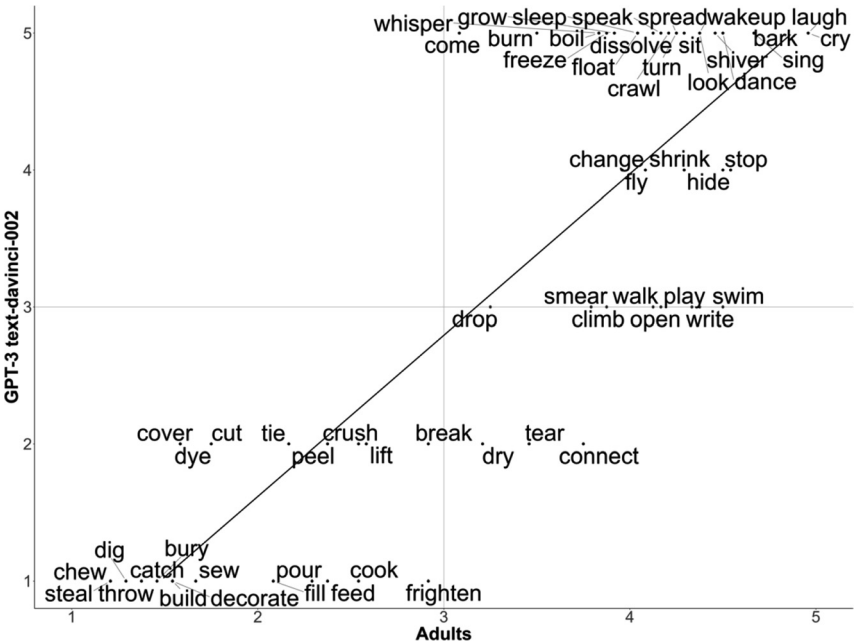

Figure 2: Raw acceptability ratings (5-point scale) for periphrastic-causatives for text-davinci-002 (Y axis) and adults from Ambridge et al. (2020) (X axis).

for traditional linguistic theories, every last detail of the model's assumptions and operation is written out in black and white, in thousands of lines of computer code.[3]

This code **is** a theory of the acquisition of (amongst other things) verb argument structure; it's even – like traditional linguistic theories – written in a language, albeit an artificial programming language, rather than a natural language like English. We know *exactly* what the model is doing. The objection that "we don't know what it's

organization". And while I'm always grateful for a citation, papers like Ambridge (2020a) are the problem, not the solution. As a verbal (non-computational) theory, Ambridge (2020a) is just a broad sketch of what a successful theory would look like. As I had begun to realise by Ambridge (2020b; *Abstractions made of exemplars or "You're all right and I've changed my mind". Response to commentators*) a large language model is an actual instantiation of an exemplar model of language acquisition, rather than just a vague outline of one.

3 In most cases, this code is not available to the public (e.g., Lisenfeld et al. 2023; see https://platform.openai.com/docs/model-index-for-researchers for details of which davinci/GPT models are available). But the point here is the existence of the code, not its availability, which is a separate issue. That said, there are at least three fully open-source large language models available: https://www.eleuther.ai, https://bigscience.huggingface.co and https://together.ai/ (many more make available the trained models, but not the code and/or training data).

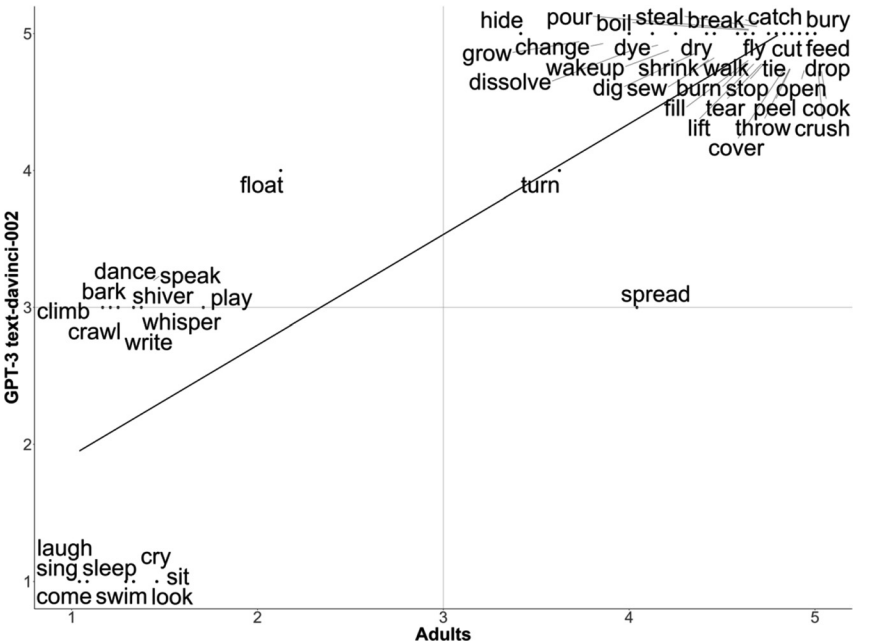

Figure 3: Raw acceptability ratings (5-point scale) for transitive-causatives for text-davinci-002 (Y axis) and adults from Ambridge et al. (2020) (X axis).

doing" boils down to "You can't give me an intuitive description". But as Piantadosi (2023) notes, there is no requirement for a scientific theory to be intuitively comprehensible. Indeed, for highly complex phenomena like language (or – to borrow Piantadosi's examples – stock market behaviour or why an oxygen molecule hits a particular place on an eyeball), it is inevitable that any successful theory will be too complex for us to grasp intuitively.

That said, to the extent that we *can* intuitively summarise, in very general terms, what a large language model is doing, it's doing something quite similar to what exemplar-, input-, and construction-based approaches to language have long assumed (e.g., Ambridge 2020a, 2020b; Ambridge and Lieven 2015; Goldberg 1995, 2006, 2019; see also Footnote 2): In the service of predicting the next word (or a masked word) in each input utterance, a large language model is (a) "storing" or in some other sense "representing" each input utterance, and (b) "abstracting" across those utterances in a way that makes it better at its prediction task. The notion that human listeners are also constantly engaged in prediction is as close as you'll get to an established fact of psycholinguistics (see Ryskin and Nieuwland 2023, for a review). And the fact that clusters of lexical items/strings lead to similar predictions

regarding likely upcoming lexical items/strings is exactly what enables the abstractions that we see in both (a) the hidden layers of large language models and (b) traditional linguistic theories (where they go by names like "valence frames" or "constructions"). The reason that the computational models do better is that the verbal theories – precisely *because* they strive for intuitive interpretability – must vastly, *vastly* oversimplify the picture, collapsing together instances of "intransitive verbs" or "transitive causative constructions" that differ probabilistically in hundreds of fine-grained ways that have consequences for (amongst other things) the acceptability of particular strings.

OK, so large language models do an excellent job of simulating human data, but are they really "theories"? After all, as an anonymous reviewer noted, human English speakers are able to give reliable judgments about the relative acceptability of individual sentences, but we would not call those speakers "theories". Is it therefore, as this reviewer suggested, a category error to call large language models "theories"?

No. The difference is that LLMs were built by humans, while human speakers were built by – take your pick – natural selection or God. "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. There is nothing analogous for flesh-and-blood human speakers.

What about semantics? Famously, verbs that can and cannot appear in the (English) causative, dative/ditranstive, figure/ground locative, verbal *un*-prefixation and passive constructions are not an arbitrary set, but are clustered according to their semantics (e.g., Ambridge et al. 2008, 2014, Bidgood et al. 2014; Levin 1993, 2015; Schäfer 2009), and there are semantic similarities in these patterns across even unrelated languages (e.g., Ambridge et al. 2020, 2022; Haspelmath 1993; Levshina 2016; Malchukov and Comrie 2015; Shibatani and Pardeshi 2002). In order to investigate whether large language models can, like human speakers, use the semantic properties of a novel verb to determine its argument structure privileges we conducted an informal "wug" test modelled after Ambridge et al. (2008). In this previous study, native English speaking adults and children were introduced to novel verbs meaning *to laugh/fall/disappear* in a particular way and, in an acceptability judgment task, rated transitive causative uses (e.g., **The magician NOVEL-DISAPPEARED the rabbit*) as unacceptable.

In the present investigation, we presented three novel verbs to GPT 3 davinci-002, alongside definitions of 'laughing/falling/disappearing in a particular way' (see https://osf.io/ctvqy/ for prompts), and interrogated the model for its "surprisal" at

transitive versus periphrastic causative uses of each verb (which we interpreted as a measure of relative acceptability). Unfortunately, we were not able to use the GPT 3 text-davinci-002 model that we used for the familiar-verb analyses reported above, because it has since been withdrawn from public access.[4] The GPT 3 model that we were able to use as a replacement (the confusingly-named davinci-002, as opposed to *text*-davinci-002) is less sophisticated at understanding complex tasks, and so does not straightforwardly yield acceptability ratings. Instead, we obtained (normalised) "surprisal" ratings for transitive- versus periphrastic-causative uses of each novel verb in a forced-choice task (see the project OSF site for details). Nevertheless, all three novel verbs were clearly, and correctly, interpreted as preferring the periphrastic- over transitive-causative, with normalised 0-to-100 ratings as follows: novel *disappear*, *M* = 75.47 conviction for the periphrastic causative (i.e., 24.53 conviction for the transitive causative), novel *fall*, *M* = 75.64, novel *laugh*, *M* = 75.49. That is, when asked to choose between **The magician NOVEL-DISAPPEARED the rabbit* and *The magician made the rabbit NOVEL-DISAPPEAR*, davinci-002 chooses the latter with a conviction of around 75 % (and similar for the novel *falling* and *disappearing* verbs).

How is the model achieving this effect? Essentially, it is operating on the basis of similarity in the distribution of the novel *falling/laughing/disappearing* verbs in our prompts and familiar *falling/laughing/disappearing* verbs in its previous training. Some would no doubt argue that this is nothing like what humans are doing, but the reality is not so clear cut. As a former colleague of ours, Franklin Chang, liked to ask, "How did you learn the meaning of the verb *to abdicate*?". Was it through hearing *abdicate* paired with real-world observations of abdication events? Of course not: No British monarch has abdicated in our lifetime (much to [finally-] King Charles' chagrin). You learned the meaning of *abdicate* solely from its distribution in your linguistic input. Or how, famously asked Landau and Gleitman (1985), do children with complete loss of sight learn the meaning of the verb *to see*? Their answer was, essentially, through its linguistic distribution; and the same must be true – for all learners – for verbs (and words of other types) that lack a straightforward visual correlate like *seem, need, miss* (or *none, the, any*) etc.; that is, for *every* word that is not either a concrete noun or an action verb.

4 https://platform.openai.com/docs/deprecations. Text-davinci-002 was ideal because it is the most sophisticated variant of GPT 3 that has not received Reinforcement Learning from Human Feedback (albeit it has used a simple form of supervised fine-tuning focused on understanding task instructions as evidenced for our Likert response scale, see https://jmcdonnell.substack.com/p/openai-comes-clean-about-gpt-35), so its demonstration of implicit knowledge about verb argument structure can be attributed almost entirely to its training via prediction of the next word in a given context using billions of corpora sequences. The GPT 3 model that we were able to use as a replacement (the confusingly-named davinci-002, as opposed to *text*-davinci-002) also avoids RLHF but has had no instruction following training at all, so does not straightforwardly yield acceptability ratings.

Neither is it uncontroversially the case that large language models solely track distributions. Yes, they are trained "only" on masked or next-word prediction. But, argue many experts (see Lappin 2023, for discussion), to become *so* good at masked/next-word prediction that you can compose a Shakespearean-style sonnet summarizing the plot of Ted Lasso (as the first author recently asked a large language model to do), you need to create – in your patterns of hidden-unit activation – something that approximates, in some sense "a representation of the real word". To take a trivial example, the words *cat, kitten* and *meow* will show somewhat overlapping patterns of hidden-unit activation, and the overlap between *cat* and *kitten* will mirror to some extent the overlap between *dog* and *puppy*. In just the same way, crosslinguistic similarities between verbs' argument structure privileges arise because English *laugh* and French *rire* pattern similarly in terms of their intra-linguistic distributions (e.g., occurring frequently with *make/faire*).

We accept, of course, that human language learning requires some non-trivial amount of real-world "grounding" (as it is called in the computational literature), and that large language models would presumably be improved by the addition of real-world visual and/or haptic information; but it is not the be-all and end-all. And, again, what's the alternative? Show us the traditional theoretical-linguistics account that can go directly from visual/haptic representations of the real world to quantitative predictions regarding verbs' argument structure privileges. There's nothing even close. Large language models aren't *perfect* in their representation of (syntactically-relevant) semantics, but they're the best we have by a long, long way.

Before the (as you might have guessed, bullish) conclusion, a few caveats: First, as we said right at the start, we are focussing – for obvious reasons – on the only domain that we happen to know quite a bit about: verbs' argument structure privileges. It may be that there are other areas of theoretical linguistics in which traditional accounts do far better than LLMs. But there's no reason to think that the present domain is a singular exception. 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). What has taken the place of phonological features? State of the art computational models that operate on audio input and generate "internal representation that corresponds to phonetic and phonological features" (Beguš 2023: 138). Second, as we have already seen, large language models lack both communicative intentions and, to some extent (though see Lappin 2023), real-world semantics (though, we hasten to add, it is not as if traditional linguistic theories have anything even approaching fully-specified theories of these things). Third, and more prosaically, LLMs require (or at least are currently trained on) much more input data than children receive (e.g., Frank 2023b; Katzir 2023). As Piantadosi (2023) notes, it may be that they don't in

fact *need* as much input as they typically get, and/or that much of the input they need is required for approximating semantic or real-world knowledge, rather than language *per se*.

Or maybe not. Maybe LLMs really *can't* learn language when given a better, more realistic multi-modal approximation of children's input. Like all good theories of language (or anything else), LLM are testable, and maybe they will fail that test.

But for now, at least in the domain of verb argument structure acquisition and representation, LLMs are the leading theory by a country mile. Traditional linguistic theories don't come close, because they insist on intuitively interpretable, but empirically inadequate, (over)generalizations. Intuitive interpretability is a weakness, not a strength; a fatal flaw, not a feature. Only by ditching intuitive interpretability can we hope to arrive at theories that capture language in all its complexity.

Theoretical linguistics is dead. Long live theoretical linguistics.

References

Adger, David. 2003. *Core syntax: A minimalist approach*. Oxford: Oxford University Press.

Aguado-Orea, Javier, Nuria Otero & Ben Ambridge. 2016. Statistics and Semantics in the acquisition of Spanish word order: Testing two accounts of the retreat from locative overgeneralization errors. *Linguistics Vanguard* 2. 1–22.

Ambridge, Ben. 2013. How do children restrict their linguistic generalizations? An (un-)grammaticality judgment study. *Cognitive Science* 37. 508–543.

Ambridge, Ben. 2020a. Against stored abstractions: A radical exemplar model of language acquisition. *First Language* 40. 509–559.

Ambridge, Ben. 2020b. Abstractions made of exemplars or "You're all right and I've changed my mind". Response to commentators. *First Language* 40. 640–659.

Ambridge, Ben & Chloe Ambridge. 2020. The retreat from transitive-causative overgeneralization errors: A review and diary study. In Caroline F. Rowland, Anna L. Theakston, Ben Ambridge & Katherine E. Twomey (eds.), *Current Perspectives on Child Language Acquisition: How children use their environment to learn*, 113–130. Amsterdam, John Benjamins.

Ambridge, Ben & Ryan P. Blything. 2016. A connectionist model of the retreat from verb argument structure overgeneralization. *Journal of Child Language* 43. 1245–1276.

Ambridge, Ben & Silke Brandt. 2013. "Lisa filled water into the cup": The roles of entrenchment, preemption and verb semantics in German speakers' L2 acquisition of English locatives. *Zeitschrift fur Anglistik und Amerikanistik* 61. 245–263.

Ambridge, Ben & Elena V.M. Lieven. 2015. A constructivist account of child language acquisition. In Brian MacWhinney & William O'Grady (eds.), *Handbook of language emergence*, 478–510. Hoboken, NJ: Wiley Blackwell.

Ambridge, Ben, Inbal Arnon & Dani Bekman. 2023. He was run-over by a bus. Passive, but not pseudo-passive, sentences are rated as more acceptable when the subject is highly affected. New data from Hebrew, and a meta-analytic synthesis across English, Balinese, Hebrew, Indonesian and Mandarin. *Glossa: Psycholinguistics*. 2(1).

Ambridge, Ben, Julian M. Pine, Caroline F. Rowland & Christopher R. Young. 2008. The effect of verb semantic class and verb frequency (entrenchment) on children's and adults' graded judgements of argument-structure overgeneralization errors. *Cognition* 106. 87–129.

Ambridge, Ben, Julian M. Pine, Caroline F. Rowland, Rebecca L. Jones & Victoria Clark. 2009. A semantics-based approach to the 'no negative-evidence' problem. *Cognitive Science* 33. 1301–1316.

Ambridge, Ben, Julian M. Pine & Caroline F. Rowland. 2011. Children use verb semantics to retreat from overgeneralization errors: A novel verb grammaticality judgment study. *Cognitive Linguistics* 22. 303–323.

Ambridge, Ben, Julian M. Pine, Caroline F. Rowland & Franklin Chang. 2012a. The roles of verb semantics, entrenchment and morphophonology in the retreat from dative argument structure overgeneralization errors. *Language* 88. 45–81.

Ambridge, Ben, Julian M. Pine & Caroline F. Rowland. 2012b. Semantics versus statistics in the retreat from locative overgeneralization errors. *Cognition* 123. 260–279.

Ambridge, Ben, Julian M. Pine, Caroline F. Rowland, Daniel Freudenthal & Franklin Chang. 2014. Avoiding dative overgeneralization errors: Semantics, statistics or both? *Language, Cognition and Neuroscience* 29. 218–243.

Ambridge, Ben, Amy Bidgood, Katie Twomey, Julian M. Pine, Caroline F. Rowland & Daniel Freudenthal. 2015. Preemption versus Entrenchment: Towards a construction-general solution to the problem of the retreat from verb argument structure overgeneralization. *PLoS One* 10(4). e0123723.

Ambridge, Ben, Amy Bidgood, Julian M. Pine, Caroline F. Rowland & Daniel Freudenthal. 2016. Is passive syntax semantically constrained? Evidence from adult grammaticality judgment and comprehension studies. *Cognitive Science* 40. 1435–1459.

Ambridge, Ben, Libby Barak, Elizabeth Wonnacott, Coin Bannard & Giovanni Sala. 2018. Effects of both preemption and entrenchment in the retreat from verb overgeneralization errors: Four reanalyses, an extended replication, and a meta-analytic synthesis. *Collabra: Psychology* 4(1). 23.

Ambridge, Ben, Ramya Maitreyee, Tatsumi Tatsumi, Laura Doherty, Shira Zicherman, Pedro Mateo-Pedro, Colin Bannard, Soumitra Samanta, Stewart McCauley, Inbal Arnon, Dani Bekman, Amir Efrati, Ruth Berman, Bhuvana Narasimhan, Dipti Misra Sharma, Rukmini Bhaya Nair, Kumiko Fukumura, Seth Campbell, Clifton Pye, Sindy F.C. Pixabaj, Mario M. Peliz & Margarita J. Mendoza. 2020. The Crosslinguistic acquisition of causative sentence structure: Computational modeling and grammaticality judgments from adult and child speakers of English, Japanese, Hindi, Hebrew and Kiche. *Cognition* 202. 104301.

Ambridge, Ben, Amy Bidgood & Kate Thomas. 2021. Disentangling syntactic, semantic and pragmatic impairments in ASD: Elicited production of passives. *Journal of Child Language* 48. 184–201.

Ambridge, Ben, Laura Doherty, Ramya Maitreyee, Tomoko Tatsumi, Shira Zicherman, Pedro Mateo-Pedro, Ayuno Kawakami, Amy Bidgood, Clifton Pye, Bhuvana Narasimhan, Inbal Arnon, Dani Bekman, Amir Efrati, Sindy F.C. Pixabaj, Mario M. Peliz, Margarita J. Mendoza, Soumitra Samanta, Sean Campbell, Stewart McCauley, Ruth Berman, Dipti Misra Sharma, Rukmini Bhaya Nair & Kumiko Fukumura. 2022. Testing a computational model of causative overgeneralizations: Child judgment and production data from English, Hebrew, Hindi, Japanese and Kiche. *Open Research Europe* 1(1). https://doi.org/10.12688/openreseurope.13008.1.

Aryawibawa, I. Nyoman & Ben Ambridge. 2018. Is syntax semantically constrained? Evidence from a grammaticality judgment study of Indonesian. *Cognitive Science* 42. 3135–3148.

Aryawibawa, I. Nyoman, Yana Qomariana, Ketut Artawa & Ben Ambridge. 2021. Direct versus indirect causation as a semantic linguistic universal: Using a computational model of English, Hebrew, Hindi, Japanese and Kiche Mayan to predict grammaticality judgments in Balinese. *Cognitive Science* 45(4). e12974.

Beguš, Gašper. 2023. Modeling unsupervised phonetic and phonological learning in Generative Adversarial Phonology. *SCiL* 6. 138–148.
Bidgood, Amy, Ben Ambridge, Julian M. Pine & Caroline F. Rowland. 2014. The retreat from locative overgeneralisation errors: A novel verb grammaticality judgment study. *PLoS One* 9(5). e97634.
Bidgood, Amy, Caroline F. Rowland, Julian M. Pine & Ben Ambridge. 2020. Syntactic representations are both abstract and semantically constrained: Evidence from children's and adults' comprehension and production/priming of the English passive. *Cognitive Science* 44(9). ee12892.
Bidgood, Amy, Julia M. Pine, Caroline F. Rowland, Giovanni Sala, Daniel T. Freudenthal & Ben Ambridge. 2021. Verb argument structure overgeneralisations for the English intransitive and transitive constructions: Grammaticality judgments, production priming and a meta-analytic synthesis. *Language and Cognition* 13. 397–437.
Blything, Ryan P., Ben Ambridge & Elena V.M. Lieven. 2014. Children use statistics and semantics in the retreat from overgeneralization. *PLoS One* 9(10). e110009.
Boyd, Jeremy K. & Adele E. Goldberg. 2011. Learning what not to say: The role of statistical preemption and categorization in a-adjective production. *Language* 87. 55–83.
Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.
Chomsky, Noam. 1981. *Lectures on government and binding*. Dordrecht: Foris.
Darmasetiyawan, I. Made Sena & Ben Ambridge. 2022. Syntactic representations contain semantic information: Evidence from Balinese passives. *Collabra: Psychology* 8(1). 33133.
Diessel, Holger. 2023. *The constructicon: Taxonomies and networks*. Cambridge: Cambridge University Press.
Frank, Michael C. 2023a. Large language models as models of human cognition. Available at: https://psyarxiv.com/wxt69.
Frank, Michael C. 2023b. Bridging the data gap between children and large language models. Available at: https://psyarxiv.com/qzbgx/.
Goldberg, Adele E. 1995. *Constructions: A construction grammar approach to argument structure*. Chicago, IL: University of Chicago Press.
Goldberg, Adele E. 2006. *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.
Goldberg, Adele E. 2011. Corpus evidence of the viability of statistical preemption. *Cognitive Linguistics* 22. 131–153.
Goldberg, Adele E. 2019. *Explain me this: Creativity, competition, and the partial productivity of constructions*. Princeton, NJ: Princeton University Press.
Haspelmath, Martin. 1993. More on the typology of inchoative/causative verb alternations In Bernard Comrie & Maria Polinsky (eds.), *Causatives and transitivity*, 87–120. Amsterdam: John Benjamins.
Katzir, Roni. 2023. Why large language models are poor theories of human linguistic cognition: A reply to Piantadosi. *Biolinguistics* 17. https://doi.org/10.5964/bioling.13153.
Kodner, Jordan, Sarah Payne & Jeffrey Heinz. 2023. Why linguistics will thrive in the 21st century: A reply to Piantadosi (2023). Available at: http://arxiv.org/abs/2308.03228v1.
Landau, Barbara & Lila R. Gleitman. 1985. *Language and experience: Evidence from the blind*. Cambridge, MA: Harvard University Press.
Lappin, Shalom. 2023. Assessing the strengths and weaknesses of large language models. Available at: https://qmro.qmul.ac.uk/xmlui/handle/123456789/90593.
Levin, Beth. 1993. *English verb classes and alternations: A preliminary investigation*. Chicago, IL: University of Chicago Press.

Levin, Beth. 2015. Semantics and pragmatics of Argument Alternations. *Annual Review of Linguistics* 1. 63–83.

Levshina, Natalia. 2016. Why we need a token-based typology: A case study of analytic and lexical causatives in fifteen European languages. *Folia Linguistica* 50. 507–542.

Liesenfeld, Andreas, Alianda Lopez & Mark Dingemanse. 2023. Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators. In *Proceedings of the 5th International Conference on Conversational User Interfaces (CUI '23)*, 1–6. New York, NY: Association for Computing Machinery. Article 47.

Liu, Li & Ben Ambridge. 2021. Balancing information-structure and semantic constraints on construction choice: Building a computational model of passive and passive-like constructions in Mandarin Chinese. *Cognitive Linguistics* 32. 349–388.

Mahowald, Kyle. 2023. A discerning several thousand judgments: GPT-3 rates the article + adjective + numeral + noun construction. *EACL* 17. 265–273.

Malchukov, Andrej & Bernard Comrie (eds.). 2015. *Valency classes in the world's languages*, vol. 2. Berlin: De Gruyter Mouton.

Milway, Daniel. 2023. A response to Piantadosi (2023). Available at: https://lingbuzz.net/lingbuzz/007264.

Müller, Stefan. 2023. *Grammatical Theory: From transformational grammar to constraint-based approaches*. Berlin: Language Science Press.

Müller, Stefan & Stephen Wechsler. 2014a. Lexical approaches to argument structure. *Theoretical Linguistics* 40. 1–76.

Müller, Stefan & Stephen Wechsler. 2014b. Two sides of the same slim Boojum: Further arguments for a lexical approach to argument structure. *Theoretical Linguistics* 40. 187–224.

Perek, Florent & Adele E. Goldberg. 2017. Linguistic generalization on the basis of function and constraints on the basis of statistical preemption. *Cognition* 168. 276–293.

Piantadosi, Steven. 2023. Modern language models refute Chomsky's approach to language. Available at: https://lingbuzz.net/lingbuzz/007180.

Pinker, Steven. 1989. *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA: MIT Press.

Reiss, Charles. 2023. Research methods in Armchair linguistics. Available at: https://lingbuzz.net/lingbuzz/007568.

Robenalt, Claris & Adele E. Goldberg. 2015. Judgment evidence for statistical preemption: It is relatively better to vanish than to disappear a rabbit, but a lifeguard can equally well backstroke or swim children to shore. *Cognitive Linguistics* 26. 467–503.

Ryskin, Rachel & Mante S. Nieuwland. 2023. Prediction during language comprehension: What is next? *Trends in Cognitive Sciences* 27. 1032–1052.

Schäfer, Florian. 2009. The causative alternation. *Language and Linguistics Compass* 3. 641–681.

Shibatani, Masayoshi & Prashant Pardeshi. 2002. The causative continuum. In Masayoshi Shibatani (ed.), *The grammar of causation and interpersonal manipulation*, 85–126. Amsterdam: John Benjamins.

Stefanowitsch, Anatol. 2008. Negative evidence and preemption: A constructional approach to ungrammaticality. *Cognitive Linguistics* 19. 513–531.

Twomey, Katie, Franklin Chang & Ben Ambridge. 2014. Do as I say, not as I do: A lexical distributional account of English locative verb class acquisition. *Cognitive Psychology* 73. 41–71.

Twomey, Katie, Franklin Chang & Ben Ambridge. 2016. Lexical distributional cues, but not situational cues, are readily used to learn abstract locative verb-structure associations. *Cognition* 153. 124–139.