

Diverging views on the role of structured representations in linguistics and language modeling

Cas W. Coopmans^{1,2} & Andrea E. Martin^{1,2}

¹Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

²Donders Institute for Brain, Cognition, and Behaviour, Radboud University Nijmegen, The Netherlands

Abstract

Recent years have seen dramatic advances in the linguistic abilities of language models, leading some authors to argue that they can be treated as models of language. Against this background, we provide a cognitive perspective on the place of language models in the science of language, with a particular focus on the role of structured representations. Evidence from linguistics, psychology and cognitive neuroscience shows that structured representations occupy a central place in the human mind. Structured representations constrain how linguistic forms map onto meanings, thereby determining the boundaries of possible human languages. In language modeling, instead, the place of structured representations is often peripheral. They are seen as optional end point of modeling, to be induced as a consequence of the objective the model tries to optimize – that is, next-word prediction. Because structured relations and representations are not necessarily useful for predicting the next word in a sequence, it is not guaranteed that the model’s internal organization will come to encode those representations. Indeed, contemporary language models struggle to use structured representations functionally to map form onto meaning, and to determine whether they should consider their input as a possible human language or not. Building on classic insights in linguistics and (computational) cognitive science, we suggest that any use of language models as models of language requires conceptualizing the role of structured representations as computational objects.

Keywords: Language models, neural networks, next-word prediction, impossible languages, structure dependence, co-reference

1. Introduction

When cognitive and computational approaches to language are compared, it is striking that they are often not aligned in terms of their goals, scientific methods, and criteria for measuring empirical success. Many cognitively-oriented approaches emphasize the importance of structured representations in explaining linguistic phenomena, language learning, and language behavior. In contrast, contemporary computational models of language, now known as ‘language models’, are predominantly trained and tested on their ability to recognize or generate sequences. This need not be a problem – it is quite likely that some forms of structured knowledge will be acquired and represented by a computational model whose primary goal is to recognize and predict words in sequences (e.g., Manning et al., 2020). On the other hand, it is not improbable that the next-word prediction objective with which language models are trained actually prevents them from recovering structural generalizations whose effects on sequential forms are very indirect. Moreover, it might support them in making sequential generalizations that are highly predictive but not possible in human language. If so, this will limit the contribution of language models to the science of language, because humans are known to generalize in structure-dependent ways, and to avoid making structure-independent generalizations (Coopmans et al., 2022; Crain & Pietroski, 2001; Culbertson & Adger, 2014; Ferrigno et al., 2020). Indeed, misalignments such as these have led to a number of recent claims that language models are poor models of language (Dupre, 2021; Fox & Katzir, 2024; Katzir, 2023; Kodner et al., 2023; Moro et al., 2023; Vázquez Martínez et al., 2024).

We are not unsympathetic to these critical perspectives, but at the same time we believe that language models have had such an outsized impact on scientific practice in the language sciences that they must not be haphazardly dismissed. It is undeniable that language models have the potential to provide novel insights into the information that can be extracted from the statistics of human language. Moreover, language models have been applied to address issues that are central to linguistics and cognitive science: they have been used to empirically assess learnability claims (Lan et al., 2024; Wilcox et al., 2023; Yedetore et al., 2023), to generate hypotheses about how grammatical information might be compactly represented in a(n artificial) neural system (Frank, 2023; Lakretz et al., 2021), and as ‘animal models’ to test the relationship between computational mechanisms and behavior (McCloskey, 1991; Scholte, 2018). Each of these usages is informed by linguistic theory, in the sense that they target linguistic phenomena that have been discovered and described extensively in linguistics. However, in their current form, they often do not

incorporate the solutions of linguistic theory to these phenomena. In this paper, we suggest that language modeling will provide an even stronger contribution to the study of language if it is not only informed by, but also incorporates, what we already know about language – specifically, about the computational role of structured representations.

This paper is structured as follows: In Section 2, we first demonstrate the importance of structured representations in language by presenting evidence from co-reference and binding. These phenomena show that the relationship between form and meaning in natural language is mediated by hierarchically structured representations. Because of the mediating role of hierarchical syntax, the way in which forms map onto meanings is indirect and cannot be described entirely in sequential terms. The complexity of this mapping might be underestimated in the practice of language modeling because language models are primarily trained to recognize and predict sequences. Any sequential regularity in the data that helps the model achieve better performance on this objective will be learned and internalized, even though such surface regularities are not the core principles underlying the data (Figure 1). In Section 3, we discuss impossible languages, which are artificial languages that are logically possible but that are never realized in natural language, arguably because they contain grammatical rules that do not rely on structured representations. Language models seem to use the same representational infrastructure for impossible languages and natural languages. This undermines their utility as comparative system for human language, because humans naturally treat possible and impossible languages as different in kind (Figure 1). In Section 4, we argue that these mismatches between cognitive and computational approaches to language are rooted in the fact that they take a different approach to structured representations. We suggest that language modeling will provide a stronger contribution to the science of language if it makes representing structures a primary target.

2. Beyond forms: Syntax maps form onto meaning

A large body of both theoretical and empirical evidence has shown that language builds meaning in a hierarchical way: the meaning of phrases and sentences is derived not just from the order in which their words appear (their external, sequential form), but also from the position of these words in their hierarchical structure (Chomsky, 1965; Coopmans et al., 2022; Crain & Pietroski, 2001; Culbertson & Adger, 2014; Everaert et al., 2015; Hunsicker & Goldin-Meadow, 2012; Kazanina et al., 2007; Lidz et al., 2003; Momma & Ferreira, 2019; Reinhart, 1983; Rizzi, 2009). In other words, structured representations, not linear sequences, are the primary unit of linguistic interpretation. Because of this mediating role of hierarchical syntax, the relationship between forms and meanings in natural language is

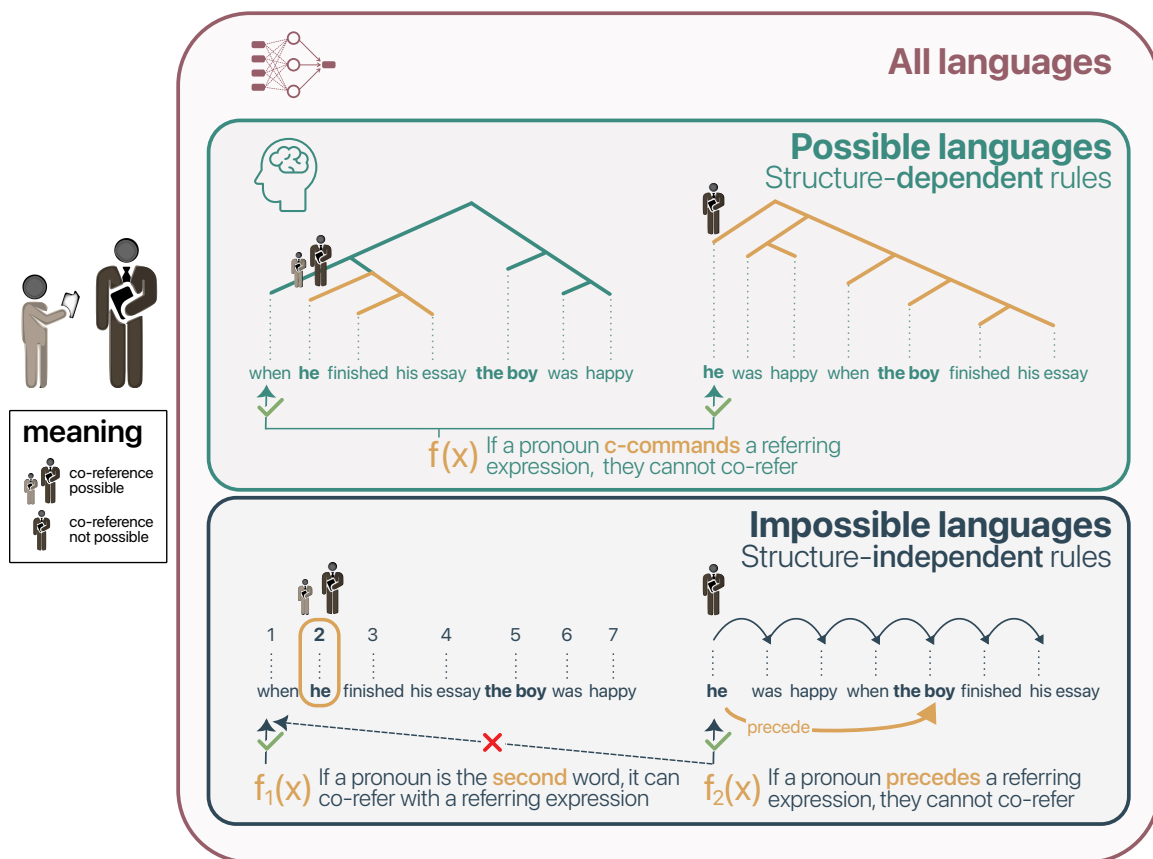


Figure 1

Only a subset of the space of all logically possible languages is considered a possible human language. Possible human languages are languages that map form onto meaning in a structure-dependent way, as is seen in structural constraints on co-reference. Impossible languages contain rules that are impossible in natural language, such as rules that do not refer to structured representations but instead rely on linear principles like precedence and counting words in a sequence. Such languages with structure-independent rules might be compatible with some of the data in the linguistic environment, but they do not capture the full set of facts (e.g., $f_2(x)$ makes the wrong prediction for "when he finished his essay, the boy was happy"). Regardless of their empirical coverage, these languages are not naturally acquired by language learners. The human brain makes a qualitative distinction between possible and impossible languages, treating only the former as natural languages. Language models instead treat all languages as similar in kind, and can therefore make structure-independent generalizations that empirically fall beyond the boundaries of possible human languages. This figure was made with icons from thenounproject.com, created by David Christensen ("neural network") and Mungang Kim ("thinking"), licensed under CC BY 3.0.

not one-to-one. Some forms do have one meaning, but others have multiple meanings, and still others have zero meanings that are grammatically licensed. An example of a form with multiple meanings is a structurally ambiguous sentence. An ungrammatical sentence instead is an example of a form without any grammaticality licensed meaning.¹ An important goal of linguistic research is explaining this relation – that is, explaining how people are able to identify both what a string of words can mean as well as what it cannot mean. In the context of this objective, accounting for the fact that certain forms are ungrammatical is only a subgoal, because ungrammaticality is a special case of a sentence having zero meanings that are grammatically licensed (Berwick et al., 2011; Pietroski, 2015; Pietroski & Hornstein, 2020). This is illustrated by the following sentences:

- (1) John kept the car in the garage.
- (2) The car was kept in the garage.
- (3) The kept car John the garage in.

These examples illustrate the ubiquitous phenomenon of constrained ambiguity. The sentence in (1) is two-way ambiguous. On one interpretation, “in the garage” refers to the place where John keeps his car – he kept it in the garage, not in the driveway. On a different reading, “in the garage” says something about “the car”, so the sequence “the car in the garage” corresponds to a syntactic unit, referring to a specific car that John kept (i.e., he did not sell it). The two meanings are derived from the fact that the string of words in (1) maps onto two hierarchical structures, each with its own interpretation. Importantly, a third, logically possible interpretation of (1) is absent: “in the garage” cannot be understood as applying to “John”, as in the paraphrase “John both kept the car and was in the garage”. In other words, the ambiguity of (1) is constrained.

The corresponding passive sentence in (2) is not ambiguous at all; it relates to just the ‘he-kept-it-in-the-garage’ interpretation. Thus, whereas there are two meanings corresponding to the form in (1), there is only one meaning for the form in (2). The reason for this is that displacement of “the car” from its canonical position, the postverbal

¹Of course, it will always be possible to creatively assign meanings to ungrammatical sentences, but any such meaning assigned is actually the meaning of another grammatical sentence. A form like “the child seems sleeping” is ungrammatical in English and has no grammatically licensed meaning. It might nevertheless be interpreted as referring to a sleeping boy, identical to the meaning of “the child seems to be sleeping”. However, the attribution of meaning to such ungrammatical forms is constrained as well, as shown by the fact that “the child seems sleeping” is not understood as “the child seems sleepy”, even though these forms are superficially very similar (Higginbotham, 1985). Such facts must be accommodated by any theory that aims to explain how we assign meanings to (un)grammatical forms.

position in the active sentence, is sensitive to the structural configuration of the sentence. The ‘he-did-not-sell-it’ reading could only be derived from (2) via a structural configuration that would violate a locality restriction on long-distance dependencies, so it is unavailable (Lightfoot, 1982). In other words, the form in (2) maps onto only one meaning because it maps onto only one structure.

The sentence in (3) is unacceptable because it does not map onto any structure. Contrasting with the grammatical sentences in (1) and (2), the word salad in (3) is typically labeled as ungrammatical. This categorical distinction is slightly misleading though, because the difference between (2) and (3) is not theoretically more important than the difference between (1) and (2). The fact that there is no meaning for which (3) would be acceptable is just as relevant as the fact that the form in (2) is not an acceptable way of expressing the ‘he-did-not-sell-it’ meaning. It is in this sense that explaining the ungrammaticality of (3) is only a subgoal; the fact that (3) has no grammatically licensed meanings (i.e., it is unacceptable on any interpretation) is just a special case of having n meanings. In other words, the relevant facts for linguistic theories concern the acceptability of a given form *under a reading*, not the acceptability or grammaticality of forms *simpliciter* (Pietroski, 2015). The overarching goal for linguistics is therefore to uncover generalizations between all possible mappings between form and meaning.

This contrasts sharply with computational approaches to language, which tend to put a lot of emphasis on forms (for comprehensive reviews, see Linzen & Baroni, 2021; Warstadt & Bowman, 2022). In short, language models are commonly trained on massive corpora (of forms), usually with the general task of predicting the next word, and sometimes with an additional training objective for fine-tuning. It is assumed that, if structured representations make the model better at next-word prediction (i.e., if they make the next word highly probable), such representations will be stored in the model’s weights and biases in a distributed and implicit way. Conversely, if structured representations do not improve the model’s predictive accuracy, they will not be learned. After fine-tuning, the model’s ability to classify grammaticality might then be evaluated by comparing the probabilities it assigns to (a critical word in) grammatical sentences than to (that same word in) minimally different ungrammatical sentences (e.g., Gulordava et al., 2018; Lan et al., 2024; Linzen et al., 2016; Marvin & Linzen, 2018; Warstadt et al., 2020; Wilcox et al., 2019; Wilcox et al., 2023). Language models are extremely successful on this task, plausibly because grammatical sentence continuations are generally more likely than ungrammatical ones. However, the types of generalizations the model comes up with are likely to be influenced by the object of optimization (Dupre, 2021; McCoy et al., 2024). That is, statistical next-

word prediction is primarily about form, so any structural relation or representation whose effects are not visible in the forms of sentences (but rather in the relationship between form and meaning) will be challenging for language models.

As an illustration of this induction problem, consider the following examples, which reveal an asymmetry in the possibility of co-reference between the pronoun “he” and the noun phrase “the boy”. Note first that all four sentences have an interpretation in which “he” and “the boy” *do not* pick out the same individual. Moreover, (4), (5) and (6) also have a reading in which “he” and “the boy” *do* pick out the same referent. The co-referential interpretation is blocked in (7); “he” and “the boy” must have disjoint reference.

(4) When **the boy** finished his essay, **he** was happy.

(5) When **he** finished his essay, **the boy** was happy.

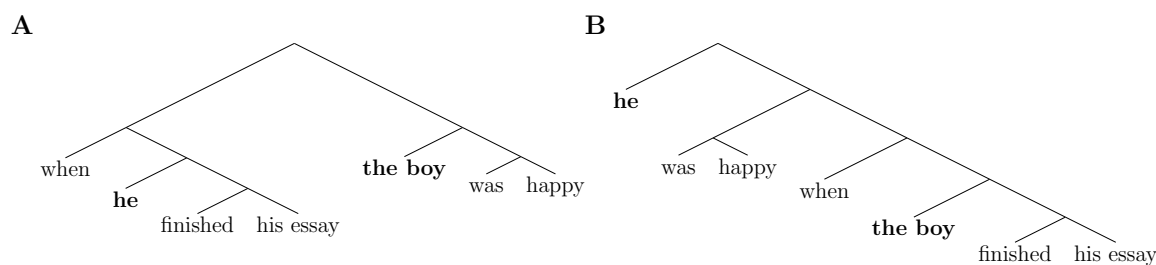
(6) **The boy** was happy when **he** finished his essay.

(7) **He** was happy when **the boy** finished his essay.

No generalization based on the linear order of the pronoun and the noun phrase is able to capture these facts. Instead, this asymmetry is captured by a generalization conventionally called Binding Principle C, a structure-dependent constraint that makes use of c-command (Chomsky, 1981; Lasnik, 1976; Reinhart, 1983). Principle C states that the interpretation of referring expressions such as “the boy” cannot be made referentially dependent on another element that c-commands them. C-command is a structural relation, formally defined as: *a node α c-commands another node β if the first branching node that dominates α also dominates β* . As can be seen in Figure 2B, which shows the structure corresponding to example (7), “the boy” is c-commanded by “he”, so a dependency relation between these two elements is ruled out by Principle C. This is not the case in example (5), because the pronoun “he” is embedded in an adverbial clause and “the boy” is not c-commanded by it, as the structure in Figure 2A shows.

As a generalization, Principle C is not limited to co-reference. It also explains the interpretation of the so-called cross-over construction in (9) (Chomsky, 1981; Freidin & Lasnik, 1981). First consider (8), which is ambiguous; a relation between “who” and “he” can, but need not, be established. In (9), however, the ambiguity is constrained: no dependency relation between “who” and “he” can be established, so this sentence has only one interpretation.

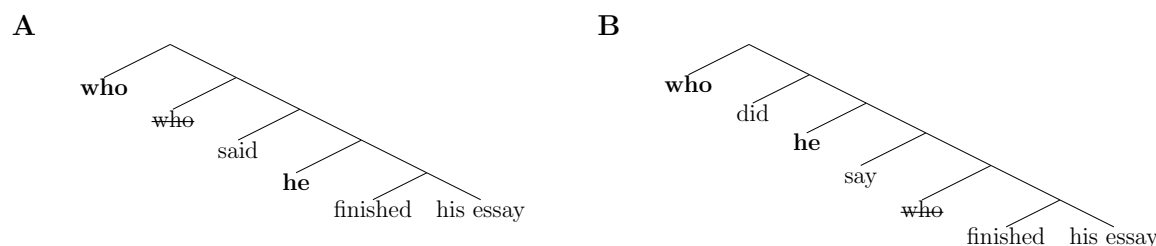
(8) **Who** said **he** finished his essay?

**Figure 2**

Two hierarchical structures in which “he” does (B) or does not (A) c-command “the boy”.

(9) **Who** did **he** say finished his essay?

Again, linear order alone cannot explain the asymmetry, because “who” precedes “he” in both cases. The difference between these examples is that “who” in (8) is the subject of “said”, which is higher in the structure than “he” (see ~~who~~ in Figure 3A), while “who” in (9) is the subject of “finished”, which occupies a position lower than “he” (see who in Figure 3B). The reason that a dependency relation between “who” and “he” in (9) is ruled out is that “he” c-commands the lower position of “who” (see Figure 3B), a violation of Principle C. This is not the case in (8), as “who” is never in a position where it is c-commanded by “he” (see Figure 3A).

**Figure 3**

Two structured representations in which “he” does (B) or does not (A) c-command the canonical position of the subject “who” (indicated as ~~who~~).

To sum up, the possibility of a referential dependency is based on a particular structural relationship between two elements. These examples therefore show that syntactic generalizations are made over structures rather than over sequential forms. If syntax relied on structure-independent principles that refer to linear order, the asymmetries would be unexpected, because the examples in (4) and (5) are identical to (6) and (7) respectively in terms of the linear order of “he” and “the boy”. The same goes for (8) and (9), which are sequentially similar – “who” precedes “he” – but structurally very different. General-

izations based on the sequential structure of sentences alone would therefore make incorrect predictions regarding the interpretation of these sentences. Because (4) licenses the same interpretations as (5), it is inductively reasonable to suppose that the interpretations of (6) and (7) are similarly linked, which they are not. Likewise, if (9) is treated as analogous to (8), then (9) should also be ambiguous, which it is not. Without considering their underlying structured representation, the constrained ambiguity of both (7) and (9) is left unaccounted for.

2.1 Forms and meanings in language models

The reason that we discuss these examples in so much detail is that they present both an analytic problem and a learning problem (Lidz, 2018). As discussed in the preceding section, the analytic problem is addressed by postulating relations over structured representations. These constrain the mapping between form and meaning, and in this way, they account for the fact that sentences (4)-(6) are ambiguous while sentence (7) is not. Thus, structured representations occupy a central role in the explanation of linguistic phenomena. The learning problem concerns the question how people or language models come to know under which conditions pronouns and noun phrases can(not) be co-referential. This problem is particularly challenging for language models, because language models learn relations and representations through the next-word prediction objective.

As noted, language models are trained by having to predict the next word in massive text corpora, after which they can be tested on their ability to distinguish minimal pairs of grammatical and ungrammatical sentences (e.g., Gulordava et al., 2018; Lan et al., 2024; Linzen et al., 2016; Marvin & Linzen, 2018; Warstadt et al., 2020; Wilcox et al., 2019; Wilcox et al., 2023). From the model’s responses to these forms, conclusions are drawn about its syntactic capacity. If the model consistently assigns a higher probability to the grammatical sentence in the minimal pair, or if it correctly ‘classifies’ the grammatical sentence as grammatical, it is assumed that the (principle underlying the) relevant grammatical construction was learned.

Besides inferential problems with such a behavioral approach to model evaluation (Guest & Martin, 2023; Kodner et al., 2023), the fact that these metrics are based on form is problematic for constructions like those in (4)-(7), because these are not grammatically distinguishable in terms of likelihood. Likewise, the training regime that is adopted might also be insufficient, because the critical difference between these constructions is a difference in meaning (only (7) is unambiguous) rather than form. To the extent that occurrence in a corpus reflects acceptability, and vice versa, the forms in (4)-(7) should be treated equally.

However, as we noted, the critical facts concern the acceptability of linguistic forms under an interpretation, not the acceptability of forms per se (Pietroski, 2015). So, to learn the relevant difference between (7) and (4)-(6) it is insufficient to only monitor which forms do and do not occur. Rather, what has to be kept track of are the relationships between forms and meanings – specifically, which meanings do and do not go together with which forms (Crain, 1991; Crain & Thornton, 1991; Pietroski & Hornstein, 2020). From the non-occurrence of certain interpretations of certain forms, the principles underlying the construction must be induced.

To determine under which conditions a structure-dependent relation like Principle C might be learned by a language model, we have to look again at how language models learn linguistic representations. Remember that the objective being optimized during training is next-word prediction. Anything regularity in the data that consistently supports that goal – i.e., any feature of the input that makes the model more accurate at predicting the next word – will in some way be internalized. We consider three scenarios that make it challenging to learn Principle C in this way. One possibility is that the model does not get better at predicting next words (e.g., referring noun phrases, proper names) when it has internalized Principle C, in which case this principle will not be learned. A second possibility is that the model does get better at predicting those words that are governed by Principle C, but that it relies on other salient features of the input, such as linear order, which is highly predictive. In that case, these non-structural features will come to represent the model’s ‘representation of co-reference’, hindering the learning of Principle C. A third possibility is that Principle C helps the model to make better predictions for constructions involving co-reference (as in (4)-(7)) but not for strong cross-over constructions involving binding (as in (8)-(9)), or vice versa. In that scenario, Principle C will be inferred for the former but not for the latter, which means that the model fails to capture an important cross-constructional generalization (Chomsky, 1986; Freidin & Lasnik, 1981).

Given these possibilities, we suspect that language models will have a hard time capturing principles of this sort, which reside at the interface between structure and meaning. Indeed, the only study that has tested a language model’s ability to learn co-referential restrictions on referring expressions showed that it was insensitive to Principle C (Mitchell et al., 2019). As another illustration of a language model’s struggle with restrictions on co-reference, we asked ChatGPT about its interpretation of our example sentences (in particular, (6) and (7)).²

²All examples use GPT-4o, which was accessed on November 4, 2024.

User:

In the sentence "The boy was happy when he finished his essay", who does "he" refer to?

ChatGPT:

"In the sentence "The boy was happy when he finished his essay," the pronoun "he" refers to "the boy." The sentence structure and context suggest that the boy is the one who finished the essay and consequently became happy. This is the most straightforward and common interpretation."

User:

Can "he" in the same sentence refer to someone else?

ChatGPT:

"In the sentence "The boy was happy when he finished his essay," the pronoun "he" most logically refers to "the boy" because this interpretation is the most straightforward and commonly expected. However, theoretically, "he" could refer to someone else if there was additional context provided by earlier sentences. For example, if the previous context introduced another male individual who finished the boy's essay for him, then "he" could refer to that person. Without such context, the default interpretation is that "he" refers to "the boy."

When probed, ChatGPT correctly reports that sentence (6) has a co-referential as well as a non-co-referential interpretation. What about sentence (7), in which people do not accept the co-referential interpretation?

User:

In the sentence "He was happy when the boy finished his essay", who does "he" refer to?

ChatGPT:

"In the sentence "He was happy when the boy finished his essay," "he" refers to a male person who is distinct from "the boy." The structure of the sentence implies that there are two different individuals: one male individual ("he") who becomes happy, and another male individual ("the boy") who finishes the essay. The identity of "he" would be understood from the broader context in which the sentence appears."

User:

Can "he" in the same sentence refer to "the boy"?

ChatGPT:

"In the sentence "He was happy when the boy finished his essay," "he" typically refers to someone other than "the boy." The structure of the sentence suggests two distinct individuals: "he" and "the boy."

However, in a broader context where it is clearly established that "he" and "the boy" are the same person, it is possible for "he" to refer to "the boy." This would rely on prior information or context that clarifies this relationship. For example, if the preceding text makes it evident that the boy is reflecting on his own actions and feelings, "he" could refer to "the boy." In general, though, without such context, the default interpretation is that they are different individuals."

ChatGPT initially reports the preferred non-co-referential interpretation of (7), but incorrectly suggests that supportive context might yield the co-referential interpretation as well. It thus appears to have no problem relating the two referential expressions, even when "he" c-commands "the boy" and the co-referential interpretation is ruled out by Binding Principle C. This pattern of behavior might at first thought seem striking because the explanation of Principle C, as well as relevant examples, can be found in linguistics textbooks and other internet resources that are likely part of the training data. However, we suggest that it is actually expected, because the form of these constructions provides no clues to the correct principle, which defines the relation between form and meaning. Because of this, we suspect that the model will not be much better at next-word prediction if it knows Principle C, due to which this principle is unlikely to be internalized.

In stark contrast to ChatGPT's structure insensitivity for these constructions, children of around three years of age have been found to adhere to Principle C in the interpretation of referring expressions, accepting co-reference in (6) but not in (7) (Crain & McKee, 1985; Eisele & Lust, 1996; Kazanina & Phillips, 2001; Lidz et al., 2021; Lukyanenko et al., 2014). At the same age, they also know how to interpret cross-over constructions, correctly rejecting the bound interpretation in (9) but not in (8) (Crain & Thornton, 1991; McDaniel & McKee, 1992). This suggests that a generalization like Principle C, which unifies the two phenomena, is on the right track as an explanatory principle of the language system. We suggest that such structure-dependent principles, which map form onto meaning, might need to become a more central goal for language modeling, because it is not guaranteed that they will be learned as a consequence of the model's training objective, i.e. optimizing

next-word prediction.

3. Limited variation: Impossible languages

Not all imaginable combinations of grammatical properties are possible in natural language. One property that does not seem to be adopted in the syntax of natural languages is structure independence: syntactic rules or operations do not refer to structure-independent properties, such as the linear position of words in a sentence (Adger, 2019; Everaert et al., 2015; Moro, 2016; Newmeyer, 2005; Rizzi, 2009). The conventional explanation for this restriction on typological variation is that structure-independent languages fall outside the boundaries of possible human languages. That is, structure dependence restricts the class of possible languages, rendering structure-independent languages *naturally* impossible. We emphasize the word *naturally* here, because ‘impossible languages’ are neither formally nor logically impossible (i.e., one can construct languages that possess these properties, such as formal languages), nor are they literally unattainable (i.e., they can be learned by other cognitive faculties). Rather, they appear to be naturally or biologically impossible in the sense that they are not acquired by the mechanisms that support naturalistic language development (Boeckx & Piattelli-Palmarini, 2005; Friederici et al., 2017; Lightfoot, 1982; Moro, 2016; Newmeyer, 2005; Singleton & Newport, 2004; Smith & Tsimpli, 1995).

Supporting evidence comes from artificial-language learning experiments, which show that people acquire ‘possible’ artificial languages differently than ‘impossible’ artificial languages. In an early study by Read and Schreiber (1982), 7-year-old children were trained to repeat word sequences that were part of a sentence produced by one of two experimenters. They had to learn implicitly which sequences to repeat by looking at what the other experimenter repeated back. In one of the conditions, the to-be-repeated sequence was a syntactic constituent (i.e., subject noun phrase), which in the training data varied in length and semantic content. Most children successfully mastered the task, consistently repeating back the right constituent. Strikingly, when the to-be-repeated sequences were non-constituents, defined in terms of their sequential properties (e.g., the first four words of the sequence), none of the children were successful. Thus, the children readily generalized structural notions like constituent, while being unable to learn structure-independent rules that refer to properties of the linear sequence.

Converging evidence comes from a series of neuropsychological studies with a subject who, while cognitively impaired according to a battery of measures, was a polyglot with a talent for learning natural languages (Smith & Tsimpli, 1995). In one of the studies, the subject had to learn an artificial language that had both ‘possible’ grammatical rules, which

were structure-dependent, as well as ‘impossible’ grammatical rules, which were structure-independent and would not be found in natural languages. For example, one impossible rule stated that the emphatic form of a sentence was created by adding the suffix *-nog* to the third orthographic word of that sentence (Smith et al., 1993). While being able to learn the structure-dependent rules, the subject failed to learn the structure-independent ones. What is interesting is that neurotypical control participants initially also failed to learn these impossible rules in the linguistic format of the experiment. However, when the problem was presented to them as a non-linguistic puzzle, they could easily solve it. The latter shows that the rules were not too difficult, nor were they impossible to learn; when the neurotypical controls could rely on “central strategies of general intelligence”, they were able to work out the problem by inducing structure-independent rules (Smith & Tsimpli, 1995, p. 154). These results suggest that people can learn impossible rules, but only when they do not rely on the mechanisms that support natural language learning.

This latter conclusion is supported by the results of fMRI experiments by Musso and colleagues, whose participants had to learn possible and impossible languages (Musso et al., 2003). Possible languages contained rules that were structure-dependent, like in natural language. Impossible languages were again artificial languages whose rules referred to the linear properties of words in a sentence. One rule, for instance, stated that the interrogative form should be created by inverting the linear sequence of the words in the corresponding declarative. Another rule stated that negative sentences should be built by putting the negation morpheme “no” after the third word of the sentence. After several training sessions, participants learned both rule types equally effectively, but they appeared to use different brain systems to do so. As performance accuracy increased, activity in the left inferior frontal gyrus (LIFG) increased as well, but only for the learning of possible languages (see Tettamanti et al., 2002, for similar results). When participants were learning impossible languages, LIFG activity actually decreased as a function of increasing performance accuracy. Subsequent studies with artificial grammars reported similar results, showing that LIFG (in particular, BA44) responds more strongly to hierarchical non-local dependencies following from a ‘possible’ phrase-structure grammar than to linear local dependencies determined by an ‘impossible’ finite-state grammar (Bahlmann et al., 2008; Chen et al., 2021; Friederici et al., 2006). The observation that people engage dissociable brain networks to learn and represent possible and impossible artificial languages, with only the former selectively activating language areas, suggests that people naturally treat possible and impossible languages as different in kind. That is, the human brain seems to make a functional distinction between possible and impossible languages, and that dis-

inction is based in part on whether their rule systems refer to structured representations or not. If language models instead learn and represent impossible languages in the same way as possible languages, that would undermine their value as comparative system for human language (Chomsky et al., 2023; Moro et al., 2023).

3.1 Impossible languages in language models

A number of studies suggest that language models do not make a qualitative distinction between possible and impossible languages (Fong & Berwick, 2008; Fong et al., 2013; Mitchell & Bowers, 2020). For instance, Mitchell and Bowers (2020) recently trained and tested an LSTM, a type of recurrent neural network, on its ability to handle number agreement in both possible and impossible structures. Possible structures were examples of natural language, like the sentence “the man whose dogs bark likes apples”. This example illustrates that number agreement in natural language is structure-dependent: the verbal morpheme -s agrees in number with “man”, to which it is hierarchically related, not with “dogs”, to which it is linearly closer. One type of impossible structure was a construction in which the linear sequence of words following a special marker was reversed, as in the sentence “the man whose dogs <marker> apples likes bark”. The correct generalization for this construction is structure-independent in two ways: the domain of application is sequence reversal, and this reversal operation is applied under the condition that the to-be-reversed elements are sequentially preceded by the marker. Mitchell and Bowers (2020) showed that the model’s performance was not affected by these structure-independent modifications: the model handled number agreement in impossible structures as effectively as in natural language constructions. More importantly, there was substantial overlap in the networks’ weights that represented agreement in possible and impossible structures, suggesting that the model does not make a fundamental distinction between the two types of structures. This contrasts with humans, who seem to employ different neurocognitive resources for possible and impossible languages (Musso et al., 2003; Smith & Tsimpli, 1995; Tettamanti et al., 2002).

One notable counter example is a recent study by Kallini et al. (2024), who assess whether a GPT-2 Transformer model can learn different classes of impossible languages. All impossible languages in their study were systematically altered versions of English, to which various structure-independent operations were applied. For instance, in their so-called *REVERSE languages, word order is perturbed: the words or tokens following a randomly inserted special marker are reversed (this is similar to the manipulation in Mitchell & Bowers, 2020). These are compared to a NoREVERSE language, which is essentially

English; the same marker is inserted but no reversal takes place. In their *HOP languages, verb inflection is perturbed using structure-independent counting rules: a special marker reflecting tense/number is placed four words or tokens after the verb. This is compared to a NoHOP language (again, essentially English), in which that same marker is placed directly after the verb, as a kind of agreement morpheme. By evaluating the model’s perplexity on each test set over the course of training, Kallini and colleagues assess how the ability to learn impossible languages is affected by the properties of these languages. What they find is that the perplexities for English are lower than those for the impossible languages, and those low perplexities are achieved in fewer training steps. Because the model learns English more efficiently than any of the impossible languages, the authors argue that it can distinguish possible from impossible languages, which in their view illustrates the utility of this language model as a model of language learning.

An important problem with this design is that all of the impossible languages were modified versions of English, so the comparison between possible and impossible languages is confounded by a difference in complexity. In fact, because the impossible languages are created by applying structure-independent perturbations to English, a language with structure-dependent syntax, the impossible languages were both structure-dependent as well as structure-independent. These are almost invariably harder to learn (or generate), so it would not be surprising if the model learns them less efficiently. As a case in point, consider the *HOP languages, where the agreement marker on the verb is moved four words or tokens to the right. To learn subject-verb agreement in this language, the model not only has to count word positions, but it first has to identify the verb that agrees with the syntactic subject, a structure-dependent relation. In the NoHOP language, only the latter is required, making this a simpler language on any definition of complexity. Given the consistent difference in complexity between the non-perturbed (natural) language and the various perturbed (impossible) languages (i.e., one rule for the possible language vs. two rules for the impossible language), the conclusion that “GPT-2 favors natural grammar rules” (Kallini et al., 2024, p. 8) is simply not licensed – one might equally interpret these results as showing that GPT-2 favors simpler languages.

To compare a model’s capacity to learn possible and impossible languages, the impossible language should not contain more rules than the possible language. It is certainly feasible to achieve this with the paradigm of Kallini et al. (2024). One could construct a *HOP language in which the agreement marker is placed n words or tokens after the subject head noun, rather than after the verb. This relation is entirely structure-independent and therefore naturally impossible, but because it is relatively easy to capture in surface

statistics, language models should be very efficient at learning this (see Coopmans et al., 2022; Frank et al., 2013; Martin & Doumas, 2017; McCoy et al., 2020; Yedetore et al., 2023). Note that this is exactly what was done in the artificial-language learning experiments by Musso et al. (2003) and Smith et al. (1993). They specifically investigated people’s ability to learn structure-independent rule types. That is, to recognize that “cookie a ate Mary” is the interrogative version of “Mary ate a cookie” (Musso et al., 2003), or that “Mary ate a-nog cookie” emphasizes Mary’s cookie eating (Smith et al., 1993), no grammatical knowledge of English is required. These impossible rules could be specified entirely in structure-independent terms, only referring to sequence position and linear order, which is why they are naturally impossible despite being computationally simple.

It might help to clarify that the argument about impossible languages with which we ended the previous section does not rest on the empirical fact that language models can learn impossible languages (as discussed, humans can do so as well, for some languages even equally effectively), but rather on the observation that they appear to use one and the same system to do so.³ When Chomsky et al. (2023) state that language models are “incapable of distinguishing the possible from the impossible”, they take “the possible” and “the impossible” as referring to distinct kinds of languages. To the extent that language models make a distinction between possible and impossible languages, that distinction is one of degree rather than kind (see Milway, 2024, for a similar critique). It is in this context that the neuroimaging data of Musso et al. (2003) and the neuropsychological data of Smith et al. (1993) are so informative – both suggest that human brains make a qualitative distinction between possible and impossible rule types. Likewise, for this discussion, the critical result of Mitchell and Bowers (2020) is not that their model could learn an impossible kind of subject-verb agreement, but that the model’s internal organization that supported this ability substantially overlapped with the part of the network that represented subject-verb agreement in natural language. This informative result therefore suggests that whatever representational infrastructure the model relied on for English, it also used for the impossible languages.

The existence of impossible languages, and the qualitative difference in learning possible and impossible languages, is a challenge for the claim that language models might be

³Note that this is not a trivial empirical result, as it could have been otherwise. In domains other than language, there is evidence for some forms of ‘emergent modularity’ in neural network systems (see Mahowald et al., 2024). The fact that such a functional or modular distinction between possible and impossible languages does not emerge in language models during the training process, even though it does arise in human brains, is problematic for the claim that these models are cognitively faithful models of language learning.

treated as linguistic theories, as some authors have recently argued (Ambridge & Blything, 2024; Baroni, 2022; Piantadosi, 2024). According to this view, the parameters and architecture of a language model are a theory of linguistic representation, which is claimed to be superior to existing linguistic theories because it can make precise quantitative predictions about the acceptability of sentences. This claim is misguided regarding the role of theories (for related critiques, but with a different focus, see Fox & Katzir, 2024; Katzir, 2023; Kodner et al., 2023). That is, the primary aim of theories is to explain key phenomena that define a field of study. In the case of language, the phenomena are the properties of the capacity for language – what it is, how it works, and why it is as it is (Cummins, 2000; van Rooij & Baggio, 2020). Thus, theories of language attempt to explain both what the nature of linguistic representations is as well as why language relies on those representations and not on others. Data and effects are secondary explananda, so being able to describe or predict the acceptability of sentences is not the main goal. Moreover, because explanations of capacities are not devised to explain effects, the predictions derived from explanatory models of a capacity will often be less accurate than the predictions of models of the data. Saying that the excellent predictive accuracy of language models makes them good theories of language is therefore conflating explanations of effects with explanations of a capacity (Cummins, 2000; Guest & Martin, 2023; van Bree, 2023; van Rooij, 2022; van Rooij & Baggio, 2020).

To be used as explanatory theories of language, language models should provide explanations for why language is the way it is. In the context of the current paper, the goal of the linguistic theory is not to show that structure-dependent generalizations can be made by computing statistics over sequential forms, but to explain why languages with structure-independent rules are not naturally and spontaneously acquired by language learners. The answer to the latter question should follow from an explanation of the capacity for language, whereas the former would be an example of capturing the data (Berwick et al., 2011).

These kinds of “why this, and not that” questions are difficult to answer for any linguistic theory, but they might be even harder to answer by relying on contemporary language models. This is because human languages would be quite different from how they actually are if they were learned in the way language models learn (see Adger, 2019; Coopmans et al., 2022; Heinz & Idsardi, 2011; Jackendoff, 1988). In that case, languages might contain such properties as sequential agreement (Adger, 2018) or probabilistic variation in the mapping between form and meaning (Hudson Kam & Newport, 2005; Singleton & Newport, 2004). Both of these patterns are compatible with many of the linguistic utterances in the environment and might therefore be learned through statistical analysis of forms.

However, neither are characteristic properties of natively acquired syntactic systems, suggesting that people (and children in particular) are in some way biased against making the corresponding inductions (Hudson Kam & Newport, 2005; Newport, 2020).

A recent study found that, over time, machine learning systems start to behave in ways increasingly unlike human behavior. This study showed that when large language models are trained on their own output for multiple ‘generations’, they eventually start to produce gibberish (Shumailov et al., 2024). The authors refer to this as ‘model collapse’, which occurs because, over generations, the model starts to forget improbable events. Events that are common in the original training data will be overrepresented in the model’s output, and low-probability events will disappear, such that the model eventually converges on a distribution very different from the original one, and with heavily reduced variance.

This important result illustrates how language models do not learn from variable input in the same way as humans, because humans obviously do not collapse under recursive training. In fact, studies of language emergence in language-deprived populations have shown the exact opposite: during development, children actually add grammatical features that are absent in the target language, thus expanding upon the resources available in the input, rather reducing them (Bickerton, 1984; Feldman et al., 1978; Gleitman & Newport, 1995; Hunsicker & Goldin-Meadow, 2012; Singleton & Newport, 2004). This is probably because humans, and children in particular, use input statistics to induce rule systems, regularizing rather than retaining probabilistic variation (Hudson Kam & Newport, 2005; Newport, 2020; Singleton & Newport, 2004). The rule-based nature of linguistic systems can explain why model collapse does not happen in humans: if people induce rules, they will always be able to produce low-probability events, as long as those events are consistent with the rule. It also allows them to impose patterns that deviate both quantitatively and qualitatively from the input they receive. This is seen most clearly in the creolization of pidgins (Bickerton, 1984; Gleitman & Newport, 1995) and in the invention of sign language systems by deaf children whose signed language input is highly impoverished (Feldman et al., 1978; Flaherty et al., 2021; Hunsicker & Goldin-Meadow, 2012; Singleton & Newport, 2004).

If language models learn and represent the regularities of possible languages in the same way as they learn and represent impossible languages, this hinders their utility as comparative model for language learning. However, such a negative conclusion should not discourage attempts at bridging the gap between language modeling and the language sciences. Quite the opposite, it means that impossible languages, like those exhibiting structure-independent rules, can occupy a more central place in future language modeling.

For instance, they might be utilized as an evaluative benchmark that goes beyond behavioral performance accuracy. By testing whether the representational resources a model assigns for processing language-like structures are different from those used for structures that do not occur in natural language, impossible languages can be used to determine if language models develop internal states that truly resemble human knowledge of language.

4. Structured representations in linguistics and language modeling

From the preceding sections, it has become apparent that structured representations occupy a different role in cognitive and computational approaches to language. In cognitive approaches, structured representations carry much of the burden of explaining what is possible and what is impossible. In Section 2, we discussed several examples of constrained ambiguity, which illustrate how structured representations constrain the mapping between form and meaning, by which they determine the boundaries of possible human languages. The neuropsychological and neuroimaging evidence discussed in Section 3 supports this conclusion by showing that people naturally distinguish possible languages, which map form onto meaning in a structure-dependent way, from impossible languages, whose rule systems do not rely on structured representations. In both cases, structured representations play a central role, quite literally: they mediate the mapping between form on the one hand and meaning on the other ($\text{form} \leftrightarrow \text{structured representations} \leftrightarrow \text{meaning}$), and they determine whether a language-like (artificial) rule system is considered a possible human language or not ($\text{input} \rightarrow \text{structured representations} \rightarrow \text{possible language}$ **X/✓**).

In studies of language modeling, structured representations typically do not play a central role. They are often seen as an optional end point – that is, as something to be induced as a consequence of the modeling objective. This is determined by probing or intervening with the internal states of the model to find causal links to performance (e.g., Lakretz et al., 2021; Manning et al., 2020) or by analyzing the model’s responses to sequences for which structure-dependent and structure-independent generalizations make different predictions (e.g., McCoy et al., 2020; Yedetore et al., 2023). Each of these endeavors is very insightful about what the model has induced during the training process. However, because model performance is the primary evaluation metric, and next-word prediction the objective being optimized, whether the model’s internal organization provides evidence for structured representations is a secondary matter. When performance is good (i.e., the model’s linguistic output is human-like), and the model’s internal states in some way encode structured representations, that is seen as relevant, because it supports the cognitive or representational adequacy of the language model. But if there is no evidence that

the model encodes structured representations, that is seen as equally informative, because it would imply that humans do not need structured representations either. We consider this ambivalent attitude towards structured representations incoherent. If it turns out that a language model can produce the right behavior on a given linguistic construction without relying on structured representations, that does not mean that humans do not use structured representations for that construction either (Guest & Martin, 2023). In Section 2, we showed that structured representations are essential in explaining linguistic phenomena, even if that is not always immediately clear from the (form of the) data, and even if structured representations by themselves might not be very helpful in optimizing a fit to the data via next-word prediction. But this is not a problem, because optimizing fit to the data is not the goal of (a theory of) structured representations, nor is it the primary explanandum (Cummins, 2000; Dupre, 2021; van Rooij, 2022; van Rooij & Baggio, 2020). Due to their strong focus on predicting probable word sequences, contemporary approaches to language modeling often view structured representations as a peripheral and optional phenomenon. As a result, seemingly successful models might fail to capture important empirical facts about language, such as those illustrated by the ubiquitous phenomenon of constrained ambiguity (Section 2).

Our suggestion for future language modeling is to reconceptualize the place of structured representations. This requires thinking about how they can play a functional role in the model’s representation of linguistic knowledge. Much inspiration, and some caution, can be found in models of structured representations in the areas of sentence parsing, analogy-making, and relational reasoning within cognitive science. For example, the large literature demonstrating the formal expressive power of structured representations across cognition (e.g., Berwick, 1985; Coopmans et al., 2023; Doumas et al., 2022; Gentner, 1983; Kemp & Tenenbaum, 2008; Pearl, 2011; Perfors et al., 2011; Shastri, 1999; Stabler, 2013; Tenenbaum et al., 2011; Yang, 2016), and the literature showing how structured representations can be realized in neural dynamics (Hummel & Holyoak, 1997, 2003) as well as learned from unstructured input (Doumas & Martin, 2018; Doumas et al., 2008). Such representations can provide a common theoretical framework for language (e.g., Martin, 2016, 2020; Martin & Doumas, 2019), and underpin generalization (e.g., Doumas et al., 2022; Lake et al., 2015), thus offering a roadmap for how language models could be enriched by embracing structure as a computational object. Recent work with neural networks has successfully adopted this approach already, by directly modeling structure and composition in language models, with benefits to performance and interpretability (e.g., Kuncoro et al., 2017; Sartran et al., 2022). What all of these approaches have in common is that they assign a computational role to

structured representations, either by specifying them or by letting them emerge in neural networks in a controlled way. In addition, they conduct the necessary tests to determine whether learned representations meet the formal and computational criteria for structure (e.g., Doumas et al., 2008, 2022; Martin & Doumas, 2019).

If successful, we suspect that language modeling will provide a stronger contribution to the study of language in each of the domains identified in Section 1, because the resulting models represent language in a way more similar to human knowledge of language, as demonstrated by formal, typological and psychological investigations into language. Beyond these computational contributions, specifying structure as a computational object will also enhance the contribution of language models to the cognitive neuroscience of language. In this field, language models are commonly used as models of, or proxies for, neural representation of language, by evaluating whether the model’s embeddings explain variance in brain activity during language processing (e.g., Caucheteux and King, 2022; Goldstein et al., 2022; Pasquiou et al., 2023; see Arana et al., 2023, for a review). Such results will be more interpretable and informative about the neurobiological representation of language if they are based on models that represent language in a structured way that is concordant with abundant research in linguistics and psychology. This, we argue, requires assigning a computational role to structured representations.

Acknowledgments

We would like to thank Bob Berwick, Charles Yang, Helen de Hoop and Peter Hagoort for helpful comments on an early draft of this paper, and Noémie te Rietmolen for making Figure 1. Andrea E. Martin was supported by a Lise Meitner Research Group “Language and Computation in Neural Systems” from the Max Planck Society, and by the Netherlands Organization for Scientific Research (NWO) VIDI grant 016.Vidi.188.029 and Aspasia grant 015.014.013.

References

- Adger, D. (2018). The autonomy of syntax. In N. Hornstein, H. Lasnik, P. Patel-Grosz, & C. Yang (Eds.), *Syntactic structures after 60 Years: The impact of the Chomskyan revolution in linguistics* (pp. 153–175). De Gruyter Mouton.
- Adger, D. (2019). *Language unlimited*. Oxford University Press.
- Ambridge, B., & Blything, L. (2024). Large language models are better than theoretical linguists at theoretical linguistics. *Theoretical Linguistics*, 50(1-2), 33–48. <https://doi.org/10.1515/tl-2024-2002>
- Arana, S., Pesnot Lerousseau, J., & Hagoort, P. (2023). Deep learning models to study sentence comprehension in the human brain. *Language, Cognition and Neuroscience*, 1–19. <https://doi.org/10.1080/23273798.2023.2198245>
- Bahlmann, J., Schubotz, R. I., & Friederici, A. D. (2008). Hierarchical artificial grammar processing engages Broca’s area. *NeuroImage*, 42(2), 525–534. <https://doi.org/10.1016/j.neuroimage.2008.04.249>
- Baroni, M. (2022). On the proper role of linguistically oriented deep net analysis in linguistic theorising. In S. Lappin & J.-P. Bernardy (Eds.), *Algebraic structures in natural language*. CRC Press.
- Berwick, R. C. (1985). *The acquisition of syntactic knowledge*. MIT Press.
- Berwick, R. C., Pietroski, P., Yankama, B., & Chomsky, N. (2011). Poverty of the stimulus revisited. *Cognitive Science*, 35(7), 1207–1242. <https://doi.org/10.1111/j.1551-6709.2011.01189.x>
- Bickerton, D. (1984). The language bioprogram hypothesis. *Behavioral and brain sciences*, 7(2), 173–188.
- Boeckx, C., & Piattelli-Palmarini, M. (2005). Language as a natural object – linguistics as a natural science. *The Linguistic Review*, 22(2-4), 447–466. <https://doi.org/10.1515/tlir.2005.22.2-4.447>
- Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1), 134. <https://doi.org/10.1038/s42003-022-03036-1>
- Chen, L., Goucha, T., Männel, C., Friederici, A. D., & Zaccarella, E. (2021). Hierarchical syntactic processing is beyond mere associating: Functional magnetic resonance imaging evidence from a novel artificial grammar. *Human Brain Mapping*, 42(10), 3253–3268. <https://doi.org/10.1002/hbm.25432>
- Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.
- Chomsky, N. (1981). *Lectures on government and binding: The Pisa lectures*. Foris.
- Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. Praeger.
- Chomsky, N., Roberts, I., & Watumull, J. (2023). Opinion | Noam Chomsky: The False Promise of ChatGPT. *The New York Times*.
- Coopmans, C. W., de Hoop, H., Kaushik, K., Hagoort, P., & Martin, A. E. (2022). Hierarchy in language interpretation: Evidence from behavioural experiments and computational modelling. *Language, Cognition and Neuroscience*, 37(4), 420–439. <https://doi.org/10.1080/23273798.2021.1980595>

- Coopmans, C. W., Kaushik, K., & Martin, A. E. (2023). Hierarchical structure in language and action: A formal comparison. *Psychological Review*, 130(4), 935–952. <https://doi.org/10.1037/rev0000429>
- Crain, S. (1991). Language acquisition in the absence of experience. *Behavioral and Brain Sciences*, 14(4), 597–612. <https://doi.org/10.1017/S0140525X00071491>
- Crain, S., & McKee, C. (1985). Acquisition of structural restrictions on anaphora. *North East Linguistics Society*, 16, 94–110.
- Crain, S., & Pietroski, P. (2001). Nature, nurture and Universal Grammar. *Linguistics and Philosophy*, 24(2), 139–186. <https://doi.org/10.1023/A:1005694100138>
- Crain, S., & Thornton, R. (1991). Recharting the course of language acquisition: Studies in elicited production. In N. A. Krasnegor, D. M. Rumbaugh, R. L. Schiefelbusch, & M. Studdert-Kennedy (Eds.), *Biological and behavioral determinants of language development* (pp. 321–337). Psychology Press.
- Culbertson, J., & Adger, D. (2014). Language learners privilege structured meaning over surface frequency. *Proceedings of the National Academy of Sciences*, 111(16), 5842–5847. <https://doi.org/10.1073/pnas.1320525111>
- Cummins, R. (2000). "How does it work?" versus "What are the laws?" Two conceptions of psychological explanation. In F. C. Keil & R. A. Wilson (Eds.), *Explanation and Cognition* (pp. 117–145). MIT Press.
- Doumas, L. A. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review*, 115(1), 1–45. <https://doi.org/10.1037/0033-295X.115.1.1>
- Doumas, L. A. A., & Martin, A. E. (2018). Learning structured representations from experience. *Psychology of Learning and Motivation*, 69, 165–203.
- Doumas, L. A. A., Puebla, G., Martin, A. E., & Hummel, J. E. (2022). A theory of relation learning and cross-domain generalization. *Psychological Review*, 129(5), 999–1041. <https://doi.org/10.1037/rev0000346>
- Dupre, G. (2021). (What) Can Deep Learning Contribute to Theoretical Linguistics? *Minds and Machines*, 31(4), 617–635. <https://doi.org/10.1007/s11023-021-09571-w>
- Eisele, J., & Lust, B. (1996). Knowledge about pronouns: A developmental study using a truth-value judgment task. *Child Development*, 67(6), 3086–3100. <https://doi.org/10.1111/j.1467-8624.1996.tb01904.x>
- Everaert, M. B. H., Huybregts, M. A. C., Chomsky, N., Berwick, R. C., & Bolhuis, J. J. (2015). Structures, not strings: Linguistics as part of the cognitive sciences. *Trends in Cognitive Sciences*, 19(12), 729–743. <https://doi.org/10.1016/j.tics.2015.09.008>
- Feldman, H., Goldin-Meadow, S., & Gleitman, L. (1978). Beyond Herodotus: The creation of language by linguistically deprived deaf children. In A. Lock (Ed.), *Action, symbol, and gesture: The emergence of language* (pp. 351–414). Academic Press.
- Ferrigno, S., Cheyette, S. J., Piantadosi, S. T., & Cantlon, J. F. (2020). Recursive sequence generation in monkeys, children, U.S. adults, and native Amazonians. *Science Advances*, 6(26), eaaz1002. <https://doi.org/10.1126/sciadv.aaz1002>
- Flaherty, M., Hunsicker, D., & Goldin-Meadow, S. (2021). Structural biases that children bring to language learning: A cross-cultural look at gestural input to homesign. *Cognition*, 211, 104608. <https://doi.org/10.1016/j.cognition.2021.104608>

- Fong, S., & Berwick, R. C. (2008). Treebank parsing and knowledge of language: A cognitive perspective. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 539–544). Cognitive Science Society.
- Fong, S., Malioutov, I., Yankama, B., & Berwick, R. C. (2013). Treebank parsing and knowledge of language. In A. Villavicencio, T. Poibeau, A. Korhonen, & A. Al-shahi (Eds.), *Cognitive aspects of computational language acquisition* (pp. 133–172). Springer Berlin Heidelberg. <https://doi.org/10.1007/978-3-642-31863-4>
- Fox, D., & Katzir, R. (2024). Large Language Models and theoretical linguistics. *Theoretical Linguistics*, 50(1-2), 71–76. <https://doi.org/10.1515/tl-2024-2005>
- Frank, M. C. (2023). Bridging the data gap between children and large language models. *Trends in Cognitive Sciences*, 27(11), 990–992. <https://doi.org/10.1016/j.tics.2023.08.007>
- Frank, R., Mathis, D., & Badecker, W. (2013). The acquisition of anaphora by simple recurrent networks. *Language Acquisition*, 20(3), 181–227. <https://doi.org/10.1080/10489223.2013.796950>
- Freidin, R., & Lasnik, H. (1981). Disjoint reference and wh-trace. *Linguistic Inquiry*, 12(1), 39–53.
- Friederici, A. D., Bahlmann, J., Heim, S., Schubotz, R. I., & Anwander, A. (2006). The brain differentiates human and non-human grammars: Functional localization and structural connectivity. *Proceedings of the National Academy of Sciences*, 103(7), 2458–2463. <https://doi.org/10.1073/pnas.0509389103>
- Friederici, A. D., Chomsky, N., Berwick, R. C., Moro, A., & Bolhuis, J. J. (2017). Language, mind and brain. *Nature Human Behaviour*, 1(10), 713–722. <https://doi.org/10.1038/s41562-017-0184-4>
- Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2), 155–170. [https://doi.org/10.1016/S0364-0213\(83\)80009-3](https://doi.org/10.1016/S0364-0213(83)80009-3)
- Gleitman, L. R., & Newport, E. L. (1995). The invention of language by children: Environmental and biological influences. In L. R. Gleitman & M. Liberman (Eds.), *An invitation to cognitive science, Vol. 1: Language* (p. 704). MIT Press.
- Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C., Fanda, L., Doyle, W., Friedman, D., ... Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3), 369–380. <https://doi.org/10.1038/s41593-022-01026-4>
- Guest, O., & Martin, A. E. (2023). On Logical Inference over Brains, Behaviour, and Artificial Neural Networks. *Computational Brain & Behavior*, 6, 213–227. <https://doi.org/10.1007/s42113-022-00166-x>
- Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1195–1205. <https://doi.org/10.18653/v1/N18-1108>
- Heinz, J., & Idsardi, W. (2011). Sentence and word complexity. *Science*, 333(6040), 295–297. <https://doi.org/10.1126/science.1210358>

- Higginbotham, J. (1985). On Semantics. *Linguistic Inquiry*, 16(4), 547–593.
- Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, 1(2), 151–195. <https://doi.org/10.1080/15475441.2005.9684215>
- Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, 104(3), 427–466. <https://doi.org/10.1037/0033-295X.104.3.427>
- Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, 110(2), 220–264. <https://doi.org/10.1037/0033-295X.110.2.220>
- Hunsicker, D., & Goldin-Meadow, S. (2012). Hierarchical structure in a self-created communication system: Building nominal constituents in homesign. *Language*, 88(4), 732–763. <https://doi.org/10.1353/lan.2012.0092>
- Jackendoff, R. (1988). Why are they saying these things about us? *Natural Language & Linguistic Theory*, 6(3), 435–442.
- Kallini, J., Papadimitriou, I., Futrell, R., Mahowald, K., & Potts, C. (2024). Mission: Impossible Language Models. <https://doi.org/10.48550/arXiv.2401.06416>
- Katzir, R. (2023). Why Large Language Models Are Poor Theories of Human Linguistic Cognition: A Reply to Piantadosi. *Biolinguistics*, 17, 1–12. <https://doi.org/10.5964/bioling.13153>
- Kazanina, N., Lau, E. F., Lieberman, M., Yoshida, M., & Phillips, C. (2007). The effect of syntactic constraints on the processing of backwards anaphora. *Journal of Memory and Language*, 56(3), 384–409. <https://doi.org/10.1016/j.jml.2006.09.003>
- Kazanina, N., & Phillips, C. (2001). Coreference in child Russian: Distinguishing syntactic and discourse constraints. *Proceedings of the 25th Boston University Conference on Language Development*.
- Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, 105(31), 10687–10692. <https://doi.org/10.1073/pnas.0802631105>
- Kodner, J., Payne, S., & Heinz, J. (2023). Why Linguistics Will Thrive in the 21st Century: A Reply to Piantadosi (2023). <https://doi.org/10.48550/arXiv.2308.03228>
- Kuncoro, A., Ballesteros, M., Kong, L., Dyer, C., Neubig, G., & Smith, N. A. (2017). What Do Recurrent Neural Network Grammars Learn About Syntax? In M. Lapata, P. Blunsom, & A. Koller (Eds.), *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers* (pp. 1249–1258).
- Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, 350(6266), 1332–1338. <https://doi.org/10.1126/science.aab3050>
- Lakretz, Y., Hupkes, D., Vergallito, A., Marelli, M., Baroni, M., & Dehaene, S. (2021). Mechanisms for handling nested dependencies in neural-network language models and humans. *Cognition*, 213, 104699. <https://doi.org/10.1016/j.cognition.2021.104699>

- Lan, N., Chemla, E., & Katzir, R. (2024). Large Language Models and the Argument from the Poverty of the Stimulus. *Linguistic Inquiry*, 1–28. https://doi.org/10.1162/ling_a_00533
- Lasnik, H. (1976). Remarks on coreference. *Linguistic Analysis*, 2, 1–21. https://doi.org/10.1007/978-94-009-2542-7_4
- Lidz, J. (2018). The explanatory power of linguistic theory. In N. Hornstein, H. Lasnik, P. Patel-Grosz, & C. Yang (Eds.), *Syntactic structures after 60 years: The impact of the Chomskyan revolution in linguistics* (pp. 225–239). De Gruyter Mouton.
- Lidz, J., Lukyanenko, C., & Sutton, M. (2021). The hunt for structure-dependent interpretation: The case of Principle C. *Cognition*, 104676. <https://doi.org/10.1016/j.cognition.2021.104676>
- Lidz, J., Waxman, S., & Freedman, J. (2003). What infants know about syntax but couldn't have learned: Experimental evidence for syntactic structure at 18 months. *Cognition*, 89(3), 295–303. [https://doi.org/10.1016/S0010-0277\(03\)00116-1](https://doi.org/10.1016/S0010-0277(03)00116-1)
- Lightfoot, D. (1982). *The language lottery: Toward a biology of grammars*. MIT Press.
- Linzen, T., & Baroni, M. (2021). Syntactic structure from deep learning. *Annual Review of Linguistics*, 7(1), 195–212. <https://doi.org/10.1146/annurev-linguistics-032020-051035>
- Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4, 521–535. https://doi.org/10.1162/tacl_a_00115
- Lukyanenko, C., Conroy, A., & Lidz, J. (2014). Is she patting Katie? Constraints on pronominal reference in 30-month-olds. *Language Learning and Development*, 10(4), 328–344. <https://doi.org/10.1080/15475441.2013.853529>
- Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6), 517–540. <https://doi.org/10.1016/j.tics.2024.01.011>
- Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48), 30046–30054. <https://doi.org/10.1073/pnas.1907367117>
- Martin, A. E. (2016). Language processing as cue integration: Grounding the psychology of language in perception and neurophysiology. *Frontiers in Psychology*, 7. <https://doi.org/10.3389/fpsyg.2016.00120>
- Martin, A. E. (2020). A compositional neural architecture for language. *Journal of Cognitive Neuroscience*, 32(8), 1407–1427. https://doi.org/10.1162/jocn_a_01552
- Martin, A. E., & Doumas, L. A. A. (2017). A mechanism for the cortical computation of hierarchical linguistic structure. *PLOS Biology*, 15(3), e2000663. <https://doi.org/10.1371/journal.pbio.2000663>
- Martin, A. E., & Doumas, L. A. A. (2019). Predicate learning in neural systems: Using oscillations to discover latent structure. *Current Opinion in Behavioral Sciences*, 29, 77–83. <https://doi.org/10.1016/j.cobeha.2019.04.008>
- Marvin, R., & Linzen, T. (2018). Targeted syntactic evaluation of language models. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1192–1202. <https://doi.org/10.18653/v1/D18-1151>

- McCloskey, M. (1991). Networks and theories: The place of connectionism in cognitive science. *Psychological Science*, 2(6), 387–395. <https://doi.org/10.1111/j.1467-9280.1991.tb00173.x>
- McCoy, R. T., Frank, R., & Linzen, T. (2020). Does syntax need to grow on trees? Sources of hierarchical inductive bias in sequence-to-sequence networks. *Transactions of the Association for Computational Linguistics*, 8(0), 125–140.
- McCoy, R. T., Yao, S., Friedman, D., Hardy, M. D., & Griffiths, T. L. (2024). Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121(41), e2322420121. <https://doi.org/10.1073/pnas.2322420121>
- McDaniel, D., & McKee, C. (1992). Which children did they show obey strong crossover? In H. Goodluck & M. Rochemont (Eds.), *Island constraints: Theory, acquisition and processing* (pp. 275–294). Springer Netherlands. https://doi.org/10.1007/978-94-017-1980-3_10
- Milway, D. (2024). On Modern Language Models, Impossible Languages, and Anti-science. In J.-L. MENDÍVIL-GIRÓ (Ed.), *Artificial Knowledge of Language: A Linguist's Perspective on its Nature, Origins and Use*.
- Mitchell, J., & Bowers, J. (2020). Priorless recurrent networks learn curiously. *Proceedings of the 28th International Conference on Computational Linguistics*, 5147–5158. <https://doi.org/10.18653/v1/2020.coling-main.451>
- Momma, S., & Ferreira, V. S. (2019). Beyond linear order: The role of argument structure in speaking. *Cognitive Psychology*, 114, 101228. <https://doi.org/10.1016/j.cogpsych.2019.101228>
- Moro, A. (2016). *Impossible languages*. MIT Press.
- Moro, A., Greco, M., & Cappa, S. F. (2023). Large languages, impossible languages and human brains. *Cortex*, 167, 82–85. <https://doi.org/10.1016/j.cortex.2023.07.003>
- Musso, M., Moro, A., Glauche, V., Rijntjes, M., Reichenbach, J., Büchel, C., & Weiller, C. (2003). Broca's area and the language instinct. *Nature Neuroscience*, 6(7), 774–781. <https://doi.org/10.1038/nn1077>
- Newmeyer, F. J. (2005). *Possible and probable languages: A generative perspective on linguistic typology*. Oxford University Press.
- Newport, E. L. (2020). Children and Adults as Language Learners: Rules, Variation, and Maturational Change. *Topics in Cognitive Science*, 12(1), 153–169. <https://doi.org/10.1111/tops.12416>
- Pasquiou, A., Lakretz, Y., Thirion, B., & Pallier, C. (2023). Information-Restricted Neural Language Models Reveal Different Brain Regions' Sensitivity to Semantics, Syntax, and Context. *Neurobiology of Language*, 4(4), 611–636. https://doi.org/10.1162/nol_a_00125
- Pearl, L. S. (2011). When Unbiased Probabilistic Learning Is Not Enough: Acquiring a Parametric System of Metrical Phonology. *Language Acquisition*, 18(2), 87–120. <https://doi.org/10.1080/10489223.2011.554261>
- Perfors, A., Tenenbaum, J. B., & Regier, T. (2011). The learnability of abstract syntactic principles. *Cognition*, 118(3), 306–338. <https://doi.org/10.1016/j.cognition.2010.11.001>

- Piantadosi, S. T. (2024). Modern language models refute Chomsky's approach to language. In E. Gibson & M. Poliak (Eds.), *From fieldwork to linguistic theory: A tribute to Dan Everett* (pp. 353–414). Language Science Press.
- Pietroski, P. (2015). Vocabulary matters. In Á. J. Gallego & D. Ott (Eds.), *MIT Working Papers in Linguistics* (pp. 199–210).
- Pietroski, P., & Hornstein, N. (2020). Universal Grammar. In A. J. Lerner, S. Cullen, & S.-J. Leslie (Eds.), *Current controversies in philosophy of cognitive science* (pp. 13–28). Routledge.
- Read, C., & Schreiber, P. (1982). Why short subjects are harder to find than long ones. In E. Wanner & L. R. Gleitman (Eds.), *Language acquisition: The state of the art* (pp. 78–101). Cambridge University Press.
- Reinhart, T. (1983). *Anaphora and semantic interpretation*. Croom Helm.
- Rizzi, L. (2009). The discovery of language invariance and variation, and its relevance for the cognitive sciences. *Behavioral and Brain Sciences*, 32(5), 467–468. <https://doi.org/10.1017/S0140525X09990574>
- Sartran, L., Barrett, S., Kuncoro, A., Stanojević, M., Blunsom, P., & Dyer, C. (2022). Transformer Grammars: Augmenting Transformer Language Models with Syntactic Inductive Biases at Scale. *Transactions of the Association for Computational Linguistics*, 10, 1423–1439. https://doi.org/10.1162/tacl_a_00526
- Scholte, H. S. (2018). Fantastic DNimals and where to find them. *NeuroImage*, 180, 112–113. <https://doi.org/10.1016/j.neuroimage.2017.12.077>
- Shastri, L. (1999). Advances in SHRUTI—A Neurally Motivated Model of Relational Knowledge Representation and Rapid Inference Using Temporal Synchrony. *Applied Intelligence*, 11(1), 79–108. <https://doi.org/10.1023/A:1008380614985>
- Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, 631(8022), 755–759. <https://doi.org/10.1038/s41586-024-07566-y>
- Singleton, J. L., & Newport, E. L. (2004). When learners surpass their models: The acquisition of American Sign Language from inconsistent input. *Cognitive Psychology*, 49(4), 370–407. <https://doi.org/10.1016/j.cogpsych.2004.05.001>
- Smith, N., & Tsimpli, I.-M. (1995). *The mind of a savant*. Blackwell Publishers.
- Smith, N., Tsimpli, I.-M., & Ouhalla, J. (1993). Learning the impossible: The acquisition of possible and impossible languages by a polyglot savant. *Lingua*, 91(4), 279–347. [https://doi.org/10.1016/0024-3841\(93\)90002-E](https://doi.org/10.1016/0024-3841(93)90002-E)
- Stabler, E. P. (2013). Two Models of Minimalist, Incremental Syntactic Analysis. *Topics in Cognitive Science*, 5(3), 611–633. <https://doi.org/10.1111/tops.12031>
- Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022), 1279–1285. <https://doi.org/10.1126/science.1192788>
- Tettamanti, M., Alkadhi, H., Moro, A., Perani, D., Kollias, S., & Weniger, D. (2002). Neural correlates for the acquisition of natural language syntax. *NeuroImage*, 17(2), 700–709. <https://doi.org/10.1006/nimg.2002.1201>
- van Bree, S. (2023). A Critical Perspective on Neural Mechanisms in Cognitive Neuroscience: Towards Unification. *Perspectives on Psychological Science*, 19(6), 993–1010. <https://doi.org/10.1177/17456916231191744>

- van Rooij, I. (2022). Psychological models and their distractors. *Nature Reviews Psychology*, 1(3), 127–128. <https://doi.org/10.1038/s44159-022-00031-5>
- van Rooij, I., & Baggio, G. (2020). Theory Development Requires an Epistemological Sea Change. *Psychological Inquiry*, 31(4), 321–325. <https://doi.org/10.1080/1047840X.2020.1853477>
- Vázquez Martínez, H. J., Heuser, A., Yang, C., & Kodner, J. (2024). Evaluating the Existence Proof: LLMs as Cognitive Models of Language Acquisition. In J.-L. Mendívil-Giró (Ed.), *Artificial Knowledge of Language: A Linguist’s Perspective on its Nature, Origins and Use*.
- Warstadt, A., & Bowman, S. R. (2022). What Artificial Neural Networks Can Tell Us about Human Language Acquisition. In S. Lappin & J.-P. Bernardy (Eds.), *Algebraic Structures in Natural Language*. CRC Press.
- Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., & Bowman, S. R. (2020). BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, 8, 377–392. https://doi.org/10.1162/tac1_a_00321
- Wilcox, E., Levy, R., & Futrell, R. (2019). What Syntactic Structures block Dependencies in RNN Language Models? *41st Annual Meeting of the Cognitive Science Society*, 1199–1205.
- Wilcox, E. G., Futrell, R., & Levy, R. (2023). Using Computational Models to Test Syntactic Learnability. *Linguistic Inquiry*, 55(4), 805–848. https://doi.org/10.1162/ling_a_00491
- Yang, C. (2016). *The Price of Linguistic Productivity: How Children Learn to Break the Rules of Language*. MIT Press.
- Yedetore, A., Linzen, T., Frank, R., & McCoy, R. T. (2023). How poor is the stimulus? Evaluating hierarchical generalization in neural networks trained on child-directed speech. <https://doi.org/10.48550/arXiv.2301.11462>