

Linguistic Generalizations are not Rules: Impacts on Evaluation of LMs

Leonie Weissweiler

The University of Texas at Austin
{weissweiler, kyle}@utexas.edu

Kyle Mahowald

Adele E. Goldberg

Princeton University
adele@princeton.edu

Abstract

Linguistic evaluations of how well LMs generalize to produce or understand novel text often implicitly take for granted that natural languages are generated by symbolic rules. Grammaticality is thought to be determined by whether or not sentences obey such rules. Interpretation is believed to be compositionally generated by syntactic rules operating on meaningful words. Semantic parsing is intended to map sentences into formal logic. Failures of LMs to obey strict rules have been taken to reveal that LMs do not produce or understand language like humans. Here we suggest that LMs’ failures to obey symbolic rules may be a feature rather than a bug, because natural languages are not based on rules. New utterances are produced and understood by a combination of flexible interrelated and context-dependent schemata or *constructions*. We encourage researchers to reimagine appropriate benchmarks and analyses that acknowledge the rich flexible generalizations that comprise natural languages.

1 Introduction

How well do large Language Models (LMs) generalize beyond their training data? The majority of work intended to address this question has presumed that symbolic rules for syntax and semantics are required to generalize: producing acceptable new forms and compositional meanings. If you learn a new color term (‘simony’) and a new count noun (‘blurk’), you know how to combine them and have a strong intuition about what ‘a simony blurk’ must be. Symbolic rules are crucial for generalizations in math, logic, formal syntax, and programming languages. They are valid *in general* and contain variables that can be instantiated by any instance of a general type (e.g., numbers in math; propositions in logic; grammatical categories in phrase structure rules).

Because earlier statistical models (e.g., n-gram or Markov models) seemed unable to generalize fully or capture non-local dependencies (Chomsky, 1957), rules seemed to many to be the only game in town for human language, too. After all, if a standard bigram model hadn’t seen ‘simony blurk’ before, it would be unable to interpret it. Influential thinkers argued that neural networks, which did not involve rules, would never be appropriate models of human cognition for this reason (Fodor and Pylyshyn, 1988; Pinker and Prince, 1988; Marcus, 1998; Fodor and Lepore, 2002; Marcus, 2001; Calvo and Symons, 2014).

Yet today’s LMs arose from statistical, distributional parallel models (Mikolov et al., 2013; Rumelhart et al., 1986) rather than rule-based natural language technologies. Though they do not rely on hard-coded rules, LMs ability to produce coherent, naturalistic language and respond appropriately is unparalleled by purely symbolic systems (Piantadosi, 2024; Goldberg, 2024; Weissweiler et al., 2023; Hofmann et al., 2024).

Despite the game-changing performance of LMs, they have inherited some of the skepticism that was directed at their forbearers (Marcus, 2001; Dentella et al., 2023; Leivada et al., 2024). And NLP researchers continue to test whether the new models learn syntactic, semantic or compositional rules: e.g., Natural Language Inference (Bowman et al., 2015), Semantic Parsing (Palmer et al., 2005; Reddy et al., 2017), tests of binary grammatical acceptability (Warstadt et al., 2019) and rule-based compositionality (Kim and Linzen, 2020). Together, such tasks made up more than half of the GLUE benchmark (Wang et al., 2018), created to evaluate language models on their skill at being “general, flexible, and robust.”

Lackluster performance on rule-based tasks, particularly in the early days of LMs, was taken to imply that although LMs may appear to be mastering natural language, they are merely imitating

Constructions: Learned Pairings of Form and Function	Rules
(Partially-filled) words, common and rare (partially-filled) schemata	Common abstract patterns
Wide range of functions	Only abstract functions
Combination of open slots and/or fixed lexical units	Include only variables
Context-sensitive (and plentiful)	Context-free (and few)
Inter-related within a complex network	Unstructured list
Sensitive to similarity and frequency	Insensitive to similarity or frequency
Slots constrained in open-ended range of ways	Variables constrained by gram. category

Table 1: Differences between constructions and rules

shallow surface patterns (Lake and Baroni, 2018; Kim and Linzen, 2020; Weißenhorn et al., 2022; Bolhuis et al., 2023). In a survey of 79 NLP researchers, McCurdy et al. (2024) reported that 87% believed LMs were not sufficiently compositional and a sizeable proportion (39%) believed explicit discrete symbolic rules were required.

Evaluations of LMs’ ability to follow algebraic or logical rules did expose certain shortcomings in their ability to reason abstractly. However, rules are not sufficient for mastering natural language, and they are only necessary in limited cases, if at all. Therefore, we suggest that rule-based evaluations of LMs’ skill with *natural language* have been over-emphasized.

Rules are not sufficient for generalization because humans depart from rule-based generalizations in a multitude of cases. For LMs to use language like humans, richer interpretations are required for thousands of collocations, conventional metaphors, idioms, and context-dependent interpretations. To the extent that apparent failures of rule-based compositionality in LMs reflect human-like behavior (Hu et al., 2024b; Lampinen et al., 2024), we further suggest that algebraic **rules are not necessary for generalization** for natural language (e.g., Hofmann et al., 2024; McClelland and Plaut, 1999).

We propose that natural language requires mastering a network of hundreds of thousands of context-dependent, gradient, flexible schemata or *constructions*, which may contain ‘slots’ that constrain their fillers and how those fillers are interpreted, as LMs do (Tseng et al., 2022). Constrained slots allow constructions to be combined in new ways, flexibly adapting to context. For instance, the construction ‘<time period> ago’ can coerce a temporal interpretation of filler phrases that do not designate time periods (e.g., ‘three rest stops ago’). Differences between rules and constructions are indicated in Table 1. Once languages are recognized to include a vast network of restricted types

of constructions (and slots), which are sensitive to similarity, frequency and context, it is unproblematic to allow rule-like constructions as a limiting case. We suggest that researchers should move past evaluating LMs on how well they obey rigid rules and focus more on the *extent to which* and *how* LMs manage to produce and comprehend human-like natural languages in all their context specificity and complexity.

Many of our points are not new. While early AI relied on algebraic rules (Minsky and Papert, 1969; Lenat, 1995), many researchers soon realized that rules were too brittle to scale up beyond highly restricted domains such as artificial block worlds (Winograd, 1980). Neural network researchers have continuously argued against the usefulness of rules, primarily in the domain of words and inflectional morphology (e.g., Rumelhart et al., 1986; Rogers and McClelland, 2004; Elman, 2009; Christiansen and Chater, 1999; MacDonald et al., 1994).

Our contribution is to review leading paradigms used in LM evaluation for syntax (§2), semantics (§3), and compositionality (§4). We explain how rules are implicitly assumed in each case, briefly describing how the assumptions arose, and why we feel they are problematic. We propose constructions as an alternative theoretical basis (§5), encouraging the field to evaluate LMs on the extent to which LMs learn and represent the complex network of constructions that comprises each language and *how* they generalize.

2 Formal Syntax in LM Evaluation

Syntax as Rules The notion that natural languages are generated by syntactic rules such as phrase structure rules, movement rules, or the operation ‘Merge’ has been assumed by most versions of generative grammar since Chomsky (1957). Syntactic rules are intended to operate on broad and clearly defined ‘grammatical categories’ (e.g., Nouns, Verbs, Adjectives), which are understood

to combine in rule-like fashion to create larger units (e.g., Noun Phrases, Verb Phrases, Adjective Phrases). The Lexicon, or system of words, was kept separate and distinct, as words, but not rules, were recognized to be influenced by frequency, similarity, meaning, or context (Pinker, 1999).

An alternative to the rule-based approach in linguistics is the constructionist approach. The latter recognizes that grammatical patterns *are* sensitive to frequencies, context, and can convey meaning, information structure and other functions—an approach we discuss in detail in Section 5.

Grammaticality Tasks LMs’ syntactic knowledge is regularly evaluated by classification tasks that require models to distinguish grammatical from ungrammatical sentences. CoLA (Warstadt et al., 2019), which includes example sentences from linguistics textbooks, is commonly used to evaluate such binary classifications. While it might be natural to assume that textbook examples represent extreme ends of a grammaticality spectrum, this is not the case. Juzek (2024) collected human acceptability judgments for part of the dataset and reported that humans assign systematically gradient judgments.

Human judgments on sentences depend on frequency, plausibility, complexity, memory demands, potential alternatives, and context (Grodner and Gibson, 2005; Schütze and Sprouse, 2013; Robenalt and Goldberg, 2015; Gibson and Hickok, 1993; Fang et al., 2023). The amount of exposure to written language and even training in linguistics also influences judgments. For instance, Dąbrowska (2010) found that laypeople’s judgments on sentences containing long-distance dependencies were more sensitive to lexical content than linguists’ judgments were.

The recognition that human judgments are gradient can have profound consequences. For instance, Dentella et al. (2023) compared humans and LMs against predetermined binary acceptability labels, reporting that LMs’ performance correlated poorly. However, comparing gradient perplexity-based judgments with the human judgments collected by Dentella et al. (2023) revealed a strong positive correlation (Hu et al., 2024a).

Dependency Evaluation The task of parsing text for universal dependencies (UD, de Marneffe et al., 2021) was a well-established task before transformer-based LMs (Zeman et al., 2017, 2018). After Hewitt and Manning (2019) showed BERT

(Devlin et al., 2019) to be somewhat skilled in UD, UD became the default operationalization of syntax in the NLP world (Amini et al., 2023; Kryvosheieva and Levy, 2025; Müller-Eberstein et al., 2022) and in discussions of inductive biases (Lindemann et al., 2024; Glavaš and Vulić, 2021). UD annotations are partially determined by semantics which draws them closer to the approach advocated here; but UD analyses presume a universal set of grammatical relations, which is problematic (e.g., Croft, 2001). Moreover, annotation is inconsistent for the long tail of language phenomena, including head-less constructions (e.g., *the Xer, the Yer* construction) or idioms. Therefore, evaluating LM accuracy on UD annotations can give misleading results.

The Chomsky Hierarchy The assumption that natural language syntax is based on formal rules is connected to the claim that it is located somewhere above regular language on the Chomsky hierarchy (Chomsky, 1956). But transformers cannot handle context-free grammars in general (Someya et al., 2024; Strobl et al., 2024), which would seem to undermine their ability to model human language *in principle*. This predicts that LMs’ successes must only be apparent. But the constructionist approach recognizes that languages are context-dependent, rather than being generated by strict rules. Therefore, under the constructionist approach, the limitations of transformer models as formal models are not limitations *qua* models of human language.

3 Formal Semantics in LM Evaluation

Formal Semantics Formal logic was developed as a branch of mathematics, used to prove mathematical and philosophical theorems, and identify provability gaps (Frege, 1918; Russell, 1905; Gödel, 1931). It is based on algebraic rules operating on clearcut and broadly defined categories (e.g., propositions). Notably, logicians did not generally assume nor endorse using this formalism to represent the meanings of natural language utterances (Carnap, 1937; Baker and Hacker, 1986). Natural language differs from formal logic in many ways. Formal logic fails to capture the different meanings of *and* and *but*; or *all*, *every* and *each*. It does not capture ambiguity or context effects (Wittgenstein, 1953; Russin et al., 2024). It does not provide a natural way to capture anything other than propositions (e.g., commands, questions, wishes, Austin, 1975), nor does it naturally distinguish presuppositions and assertions (Strawson, 1967).

Semantic Parsing Semantic parsing evaluations arose as an extension of syntactic parsing. They require models to map sentences into rule-based symbolic representations (Banarescu et al., 2013) to evaluate semantic understanding in LMs (Li et al., 2023; Qiu et al., 2022; Shaw et al., 2021, see also §4). At times, semantic parsing is applied explicitly to restricted domains designed to obey rules, but such domains are necessarily limited. For instance, Piantadosi et al. (2016) trained a model on representations of ‘the girl,’ ‘the cat,’ ‘the hedgehog,’ ‘the cat loves the girl,’ and ‘the hedgehog sees the cat,’ and so on to test whether the model predicted a formal semantic representation for ‘The girl loves the hedgehog.’ However, note that if ‘mosquitoes’ were substituted for ‘the cat,’ different interpretations of ‘love’ would be evoked (‘Mosquitoes love the girl’ vs. ‘The girl loves mosquitoes’), not to mention markedly different degrees of plausibility.

Natural Language Inference Natural Language Inference tasks label the second of two sentences an entailment, contradiction, or neutral. While entailment and contradiction are key concepts in mathematical proofs, their importance in language understanding has been overstated as in the quote from Bowman et al. (2015) (emphasis added, see also Katz, 1972; van Benthem, 2008):

The semantic concepts of entailment and contradiction are central to *all* aspects of natural language meaning.

NLI tasks were originally used to train models (Wang et al., 2019; Dagan et al., 2006; Nie et al., 2020). In the age of LMs, they are used as a zero-shot evaluation metric to assess natural language understanding (Zhou et al., 2024; McCoy et al., 2019). But this may lead us to underestimate LMs. The recognition of necessary and plausible inferences is an important aspect of natural language understanding, but the NLI task is oversimplified: it fails to account for the communicative goal or context-dependent interpretations.

Humans’ goal is to make sense of others’ messages, so we assume others are trying to be relevant and helpful and do our best to assign coherent meanings to all utterances (Grice, 1975). For example, outside of logic classes or heated arguments, people rarely conclude that two statements made by the same person are contradictory. If someone utters: ‘The boy is depressed; The boy is not depressed,’ listeners do not throw up their hands and

shout “contradiction.” Instead, they may infer that clinicians disagree about whether the boy is depressed, or understand that the boy is sad at the moment but not truly depressed. Humans also assign distinct context-dependent interpretations to apparent tautologies such as ‘Either it’s alive or it’s not’ and ‘If it snows, it snows.’ Therefore, NLI tasks that rely on judging contradictions or entailments may over- or under-estimate how well LMs understand natural language the way people do.

4 Compositionality in LM Evaluation

Compositionality in Linguistics As computer coding languages became more and more widespread, rule-based semantics and syntax took root in linguistics. A Principle of Compositionality combined the two traditions. It states that the meaning of a sentence is determined by the meanings of the words and the syntactic rules used to combine them (Montague, 1970; Partee, 1984; Dowty, 1979; Jackendoff, 1992; Fodor and Lepore, 2002). This is a bottom-up process: syntactic rules combine words, which have determinant meanings. Context was not supposed to influence the interpretation of words in a top-down manner. Instead, downstream inferences were invoked to address the obvious fact that interpretation does depend on context. As Fodor and Pylyshyn (1988) state, “a lexical item must make approximately the same semantic contribution to each expression in which it occurs”. Yet they, like Carnap and Frege, acknowledge: “It’s uncertain exactly how compositional natural languages actually are” (Fodor and Pylyshyn, 1988).

The standard argument in favor of compositionality is outlined in White et al. (2024):

1. People tend to agree on the interpretation of new sentences. \Rightarrow There must be some set of rules that determine the meaning of new sentences.
2. Sentences are generated by a relatively small set of syntactic rules that combine words (Chomsky, 1957). \Rightarrow Meaning is determined by meaningful words and the syntactic rules used to combine them.

While people more or less agree on the meanings of new sentences in context, this does not entail that meaning is determined by rule-based algorithms operating on familiar words (contra [1]). People also generally agree on the interpretations of pointing gestures and novel words, and yet in each of these

cases, the shared interpretations must be gleaned from non-linguistic context (in the case of pointing gestures), or from a combination of linguistic and non-linguistic context.

Similarly, the meanings of familiar collocations, compounds and idioms are not determined by general rules. For instance, a compositional rule involving set-intersection may be appealing for '<color term> noun' combinations in the domain of artificial block worlds (e.g., a green cube is something that is both a cube and green). However, violations of such rules abound e.g., green tea is more yellow than green and Cambridge blue is actually green. Even more common are instances that evoke richer meanings than predicted by any algebraic rule: e.g., a green light implies that forward motion or progress is permitted, and a green card provides a path toward citizenship in the US.

Critically, people tend to mostly agree on the interpretations of utterances in context-dependent ways. Consider 'the Persian cat is on the rug.' If the goal of the speaker is only to find the furball, there need be no commitment to the cat being a thoroughbred Persian breed. Likewise, the speaker need not be committed to the cat being wholly on, rather than adjacent to, the rug. Or, comprehenders may appreciate the statement as ironic, if the cat is hairless. And the rug need not have been created or used as a rug for people to share the same intended interpretation of the sentence.

The second premise in the standard argument is also problematic. Rules massively over-generalize, which is why they were never used for production. That is, rules predict all manner of odd locutions (Pawley and Syder, 1983; Sag et al., 2002): e.g., 'Meeting you is pleasing to me'; 'The tall winds hit the afraid boy'; 'Explain him the problem.' Humans are sensitive to the frequencies of various types of word combinations and judge formulations unnatural whenever there exists a more conventional way to express the intended message in context (e.g., Goldberg, 2019).

Evaluating LMs for Compositionality Compositionality benchmarks combine elements from one or another evaluation paradigm already described. Kim and Linzen (2020)'s compositional generalization challenge (COGS) tested whether models could formalize into formal semantics, any sentence generated by a small set of syntactic rules. They anticipated generalizations from sentences like 'Jane gave the cake to John' to 'Jane gave

John the cake.' The models were found to perform poorly. Speakers' choice between these two constructions is highly sensitive to information structure, dialect (Bresnan and Ford, 2010), and the relative frequencies and similarities of verbs witnessing in each version (Goldberg, 2019; Leong and Linzen, 2024; Ambridge et al., 2014). Insofar as natural language is not amenable to logical representations, failures of LMs to map to such representations may be consistent with human interpretation.

Other compositionality benchmarks adopt NLI tasks, which commonly presume interpretation is determined by rules. For example, in the context of robotic agents interpreting instructions, Lake and Baroni (2018, p.1) state:

Humans can understand and produce new utterances effortlessly, thanks to their compositional skills. Once a person learns the meaning of a new verb 'dax', he or she can immediately understand the meaning of 'dax twice'...

The robotic agents struggled to interpret the rule-based command, though it was appropriate in the narrow domain tested. Notably, the rule does not extend to a broader swath of language. For instance, unbounded actions are not countable, so if 'twice' appears at all, it is likely followed by a comparative phrase (e.g., 'work twice as hard') and a very different meaning than performing an action two times. Other cases require knowledge of specific combinations: 'to think twice,' which means 'to hesitate' and 'going twice' tends to evoke the context of an auction. Familiar phrases with meanings not fully captured by compositional rules are common: By one estimate, we learn tens of thousands of them (Jackendoff, 2002). Importantly, we largely agree on their interpretations, even though each means something more or different than predicted simply by the words and their syntactic combination. In this way, phrasal combinations regularly involve subregularities or item-specific interpretations not predicted by a general algebraic rule.

Another example comes from the seemingly innocuous algebraic rule stated below:

If X is more Y than Z, then Z is less Y than X, irrespective of the specific meanings of X, Y, and Z. (Dasgupta et al., 2020, p.5)

The rule is intended to capture that 'Anne is more cheerful than Bob' should both contradict 'Anne

is less cheerful than Bob’, and entail ‘Bob is less cheerful than Anne.’ NLI models that failed to draw these inferences were considered lacking. Yet natural language rarely relies on free variables. The content of X, Y, and Z matter. No one would infer that because Anne_x is more cheerful_y than careful_z, that ‘Careful_z is less cheerful_y than Anne_x.’ Perhaps more importantly, if a speaker uttered ‘Anne is higher than Bob and Bob is higher than Anne,’ listeners would likely infer either that Bob climbed above Anne in the time it took to utter the first clause or that Bob has been smoking.

5 Constructions

We have argued against the idea that natural language is generated or interpreted by symbolic rules, but we agree that speakers are generally able to agree on the meanings of new sentences well enough for communication to be successful. This section briefly explains how constructions offer an alternative with respect to each of the differences cited in Table 1. Language is generated by flexibly combining constructions, which comprise a rich and complex ConstructionNet for each language. These include words, but are far broader than the traditional lexicon, encompassing schemata larger than individual words as well (Table 2).

(Partially-filled) Words, Common and Rare Schemata We use the term ‘construction’ to refer to a learned association between a formal pattern and a range of related functions. This simple definition treats words, idioms, rare *and* common grammatical patterns as constructions, and recognizes that each case may include open ‘slots’ (Table 2). Formal attributes may include phonology, lexical content, grammatical categories, word order, discontinuous elements, and/or intonation.

Wide Range of Functions Constructions’ functions vary widely: Constructions may convey rich, specific contentful meaning in the case of words, collocations, idioms. A plethora of other constructions are productive but constrained in a wide variety of semi-specific ways; argument structure constructions convey ‘who did what to whom’; discourse structuring constructions indicate which subparts of a sentence are at-issue or backgrounded. Constructions exist to ask questions, express surprise or disapproval, for example. Any level of construction can be associated with specific registers, genres, and/or dialects.

Types of Construction and Examples

Words: pregame; running; nevertheless; ago

Partially filled words (morphemes): pre-N_{event}; V-ing

Collocations, fixed idioms: high winds; jump the shark

Partially-filled productive constructions: X is the new Y; It’s Adj of <agent> VP_{to}

Partially-filled argument structure constructions: give <recipient> a call

Argument structure constructions: Intransitive; Caused-motion; Double Object; Resultative

Discourse-structuring constructions: get-passive; information questions; it-clefts; relative clauses

Table 2: Example constructions at varying levels of complexity and abstraction

Sensitive to Similarity and Frequency Language users are sensitive to the frequencies of constructions. For instance, the passive construction is far more frequent in Turkish than English and young Turkish speakers learn the construction far earlier than English-speaking children (Slobin, 1986). Constructions are also influenced by similarity: Instances of a construction prime instances of the same or closely related construction (e.g., Du Bois, 2014; Pickering and Ferreira, 2008).

Productive Constructions May Include Fixed Lexical Units As shown in Table 2, syntax, semantics and morphology are interrelated rather than assigned to distinct levels. This is useful because even productive hierarchical constructions often include particular words and semantic constraints. For example, an English construction that implies real or metaphorical motion allows a wide range of verbs but requires the particular noun ‘way’ (‘He charmed his way into the meeting’).

Plentiful and Context-Sensitive The broad definition of constructions as pairings of form *and* function, including words *and* grammatical patterns, is another difference between constructions and rules. Constructions are far richer and more plentiful than the class of rules is commonly envisioned to be. Constructions also do far more work than rules since they capture frequency information and contextual constraints, while rules are presumed to be context-free and uninfluenced by frequency.

Interrelated System, Not Unstructured List Unlike rules, which are commonly presented as unstructured lists, constructions comprise a network of interrelated statistical patterns. This allows

for the fact that languages have families of related constructions. It also allows for the simple fact that productive constructions simultaneously co-exist with specific conventional instances. For instance, the English ‘double object’ construction is productive, and speakers are also familiar with dozens of conventional instances (e.g., ‘give <someone> the time of day’, ‘throw <someone> a bone’).

Construction Slots Are Constrained The open ‘slots’ of constructions are constrained in a wide variety of ways. For instance, the English double-object construction can appear with a wide range of verbs, but prefers simple verbs to those that sound Latinate (e.g., ‘She told them something’ vs. ‘She proclaimed them something’). The English comparative suffix ‘-er’ (e.g., ‘calmer’, ‘quicker’) is available for most single-syllable adjectives that allow a gradient interpretation, but it is not used with past participles adjectives (? ‘benter’).

An Example Consider ‘X is the new Y’. It is productive and can be used to create new utterances e.g., ‘Semiconductor chips are the new oil.’ As is typical of productive constructions, the generalization co-exists with several familiar instances (e.g., ‘50 is the new 40’; ‘Orange is the new black’). The construction is not an algebraic rule. Its slots, indicated by X and Y, are not variables that range freely over fixed syntactic categories. Instead, ‘X’ must be construed (playfully) as currently functioning in the culture as ‘Y’ used to. Therefore not all combinations of slot fillers make sense: (e.g., ? ‘Orange is the new oil’). Adding a parallelism constraint between X and Y is insufficient since ‘103 is the new 101’ also makes little sense. Finally, instances of the construction are not amenable to translations into formal logic, which would presumably treat ‘Orange is the new black’ as equivalent to ‘Black is the old orange,’ which does not conventionally evoke the same meaning.

6 Implications Beyond Natural Language

The current observations help make sense of LM behavior outside the domain of pure language. Even in domains that are rule-like by design, certain types of non-compositional behavior exist, likely due to their interface with natural language. For instance, LMs have been found unreliable at drawing the following inference, which the authors dubbed the *reversal curse*: “if ‘A is B’ [...] is true, then ‘B is A’ follows by the symmetry property of

the identity relation” (Berglund et al., 2023, p. 2).

Why are LMs prone to the reversal curse? Although the quote above is stated in natural language, it does not apply to natural language sentences, which are actually rarely reversible. For example, ‘A mental illness is the same as a physical illness’ means something very different than ‘A physical illness is the same as a mental illness’ (see also Tversky, 1977; Talmy, 1975). Even simple conjunctions are not generally reversible in natural language. For instance, ‘night & day’ and ‘day & night’ are both acceptable, but their interpretations differ: the former conveys a stark contrast (e.g., ‘as different as night and day’), the latter suggests a relentless activity or process (e.g., ‘he worried day and night’). In summary, it is perhaps reasonable to expect truly symmetric knowledge to be reversible. But LMs are trained on natural language and natural language utterances are not symmetric.

Human reasoning depends on the context in which performance is tested (Klauer et al., 2000; Wason, 1968; Tversky and Kahneman, 1974) and how instructions are formulated (Evans et al., 1994). Lampinen et al. (2024) find that LMs and humans are influenced by semantic context in similar ways (see also McCoy et al., 2023). Even math is not fully rule-compositional when equations are intended for communication: for instance, $2 + 2 = 4$ means something different than $4 = 2 + 2$ (Mirin and Dawkins, 2022), a preference picked up on by LMs (Boguraev et al., 2024).

7 New Directions

Natural languages involve complex and context-sensitive systems of constructions, which vary from being wholly fixed to highly abstract and productive. Constructions are combined when a unit, potentially itself composed of constructions, fills a slot in another construction. Viewing language as a system of constructions rather than words and rules may fundamentally change how the successes and failures of models are construed, and new goals and questions come into focus.

Balancing Constructions and Rules Coding languages are compositional by design. They are unambiguous with variables filled by any instance of a clearly defined and general type. Accordingly, increasing the proportion of code in pretraining improves performance on tasks that rely on rule-based compositionality such as logic and math (e.g., Kim et al., 2024; Madaan et al., 2022). Yet Petty

et al. (2024) report that adding code to pretraining data hurts performance on naturalistic language processing including tasks involving the English passive (Mueller et al., 2022) and BigBench’s Implicatures and CommonMorpheme tasks (Srivastava et al., 2023). Sprague et al. (2024) report that while Chain-of-Thought prompting had been believed helpful across-the-board, performance only improves on problems that require algorithmic reasoning. Thus, adding code to pre-training or using CoT prompts benefit tasks that are designed to be rule-compositional but may be detrimental to natural language tasks.

Better Datasets Rather than using abstract rules to generate stimuli for natural language benchmarks, ecologically valid stimuli may be more usefully collected or adapted from natural corpora and then normed for naturalness and plausibility. Since human judgments are highly context-dependent, benchmark tasks should ideally also vary contexts systematically (see, e.g., Ross et al., 2024).

It is also important to avoid inadvertently training human coders to give the type of responses only suitable in logic or coding classes. If people are instructed to interpret ‘red X’ as ‘X that is red for any X,’ they can do so. Yet in natural contexts, people understand that red grapefruits are closer to pink, red hair is more orange, a red book may be about communism, and crossing a red line may have consequences. Finally, a variety of items, participants, and contexts ought to be valued as much as a variety of models.

Probing for Constructions Recognizing that LMs are already extremely skilled at producing and responding to several natural languages allows for a shift in research agendas. The interesting question may now be not *if*, but *how* LMs achieve such remarkable skills. We can now also ask how well LMs capture appropriately nuanced interpretations and relationships among constructions. New ways of probing LMs make this possible and this toolkit will only grow.

As is familiar from the lexicon, constructions are related to one another because human memory is highly associative. Misra and Mahowald (2024) have demonstrated that even when all instances of a rare non-compositional construction are ablated from training data, non-trivial learning of the construction remains, enabled by the presence of related constructions in training. Another type of relationship among constructions are

the relationships between conventional instances and productive generalizations. Nearly every productive construction co-exists with at least a few formulaic instances, and LMs offer ways of testing relationships among instances that give rise to productive generalizations. Other recent work includes Weissweiler et al. (2022), who found LMs reliably discriminate instances of the English Comparative Correlative from superficially similar expressions. Tayyar Madabushi et al. (2020) tested a dataset of automatically induced constructions and reported that BERT (Devlin et al., 2019) could determine whether two sentences contained instances of the same construction. However, Zhou et al. (2024) found LMs failed to distinguish entailment differences between the causal excess construction (e.g., ‘so heavy that it fell’) and two structurally similar constructions (‘so happy that she won’; ‘so certain that it rained’). Similarly addressing challenging construction semantics, Weissweiler et al. (2024) showed LMs to struggle with the meaning of the caused-motion construction.

Tseng et al. (2022) showed that LMs gradiently predict appropriate slot fillers. Li et al. (2022) probed RoBERTa’s implicit semantic representations of four argument structure constructions and found similarities in behavior in the model and a sorting task done by humans. Potts (2023) found that despite its rarity, LMs acquire the Preposing in PP construction (Huddleston and Pullum, 2002).

8 Conclusion

Generalization is a key component of human language—and a big part of why LMs are successful at processing language. But we have argued that evaluations of the linguistic abilities of LMs are too often based on an assumption that generalization requires algebraic rules operating on words. But natural languages are not Lego sets. We suggest instead that language involves flexible combinations of rich and varied constructions of varying sizes, complexity, and degrees of abstraction, which differ from algebraic rules in many ways. By designing new evaluations that accurately reflect the complexities of language, we can avoid under- or overestimating language models. The extent to which LMs produce and interpret combinations of constructions has been limited to date. LMs offer fertile ground for new types of evaluations and new analyses that offer deeper understanding of the remarkable skills required for natural language.

Limitations

The claims here are based on existing evaluations of LMs. LMs are rapidly improving in a variety of ways. While we have aimed to discuss benchmarks and evaluations in ways that reflect the historical trajectory as well as the present-day landscape, it is always possible that newer models could behave differently than our characterizations.

Another potential limitation is that evaluating LMs is a moving target. While we give a number of suggestions for evaluation, we also recognize that the kinds of linguistic evaluations needed may move away from tests of grammaticality (which seems largely mastered) and towards more general kinds of language understanding. The rapid pace of LM technology makes future-proofing such designs potentially difficult.

Finally, while we focus mostly on constructionist approaches, there are related usage-based approaches from the functionalist tradition that would likely make similar predictions. Future work can further flesh out these directions.

Acknowledgments

We thank Kanishka Misra and Will Merrill for helpful discussions and feedback. We are grateful to audiences at the NSF-sponsored New Horizons in Language Science workshop and the Analytical approaches to understanding neural networks summer school sponsored by Simon's Foundation for helpful feedback. Leonie Weissweiler was supported by a postdoctoral fellowship of the German Academic Exchange Service (DAAD).

References

- Ben Ambridge, Julian M. Pine, Caroline F. Rowland, Daniel Freudenthal, and Franklin Chang. 2014. [Avoiding dative overgeneralisation errors: semantics, statistics or both?](#) *Language, Cognition and Neuroscience*, 29(2):218–243.
- Afra Amini, Tiago Pimentel, Clara Meister, and Ryan Cotterell. 2023. [Naturalistic causal probing for morpho-syntax](#). *Transactions of the Association for Computational Linguistics*, 11:384–403.
- J.L. Austin. 1975. *How To Do Things With Words: The William James Lectures delivered at Harvard University in 1955*. Oxford University Press.
- Gordon P. Baker and P. M. S. Hacker. 1986. *Language, sense and nonsense: a critical investigation into modern theories of language*, reprinted edition. Blackwell, Oxford.
- Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. [Abstract Meaning Representation for sembanking](#). In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Dis-course*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.
- Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2023. The reversal curse: LLMs trained on "a is b" fail to learn "b is a". *arXiv preprint arXiv:2309.12288*.
- Sasha Boguraev, Ben Lipkin, Leonie Weissweiler, and Kyle Mahowald. 2024. [Models can and should embrace the communicative nature of human-generated math](#). *Preprint*, arXiv:2409.17005.
- Johan J. Bolhuis, Stephen Crain, and Ian Roberts. 2023. [Language and learning: the cognitive revolution at 60-odd](#). *Biological Reviews*, 98(3):931–941.
- Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. [A large annotated corpus for learning natural language inference](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.
- Joan Bresnan and Marilyn Ford. 2010. Predicting syntax: Processing dative constructions in american and australian varieties of english. *Language*, 86(1):168–213.
- Paco Calvo and John Symons. 2014. *The Architecture of Cognition: Rethinking Fodor and Pylyshyn's Systematicity Challenge*. MIT Press.
- Rudolf Carnap. 1937. *Logical Syntax of Language*, 1st edition. Routledge.
- N. Chomsky. 1956. [Three models for the description of language](#). *IRE Transactions on Information Theory*, 2(3):113–124.
- Noam Chomsky. 1957. *Syntactic Structures*. Mouton.
- Morten H Christiansen and Nick Chater. 1999. [Toward a connectionist model of recursion in human linguistic performance](#). *Cognitive Science*, 23(2):157–205.
- William Croft. 2001. *Radical construction grammar: Syntactic theory in typological perspective*. Oxford University Press, USA.
- Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

- Ishita Dasgupta, Demi Guo, Samuel J. Gershman, and Noah D. Goodman. 2020. [Analyzing machine-learned representations: A natural language case study](#). *Cognitive Science*, 44(12):e12925.
- Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. [Universal Dependencies](#). *Computational Linguistics*, 47(2):255–308.
- Vittoria Dentella, Fritz Günther, and Evelina Leivada. 2023. [Systematic testing of three language models reveals low language accuracy, absence of response stability, and a yes-response bias](#). *Proceedings of the National Academy of Sciences*, 120(51):e2309583120.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- David R. Dowty. 1979. *The Semantics of Aspectual Classes of Verbs in English*, pages 37–132. Springer Netherlands, Dordrecht.
- John W. Du Bois. 2014. [Towards a dialogic syntax](#). *Cognitive Linguistics*, 25(3):359–410.
- Ewa Dąbrowska. 2010. [Naive v. expert intuitions: An empirical study of acceptability judgments](#). *The Linguistic Review*, 27(1):1–23.
- Jeffrey L. Elman. 2009. [On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon](#). *Cognitive Science*, 33(4):547–582.
- J.St.B.T. Evans, S. E. Newstead, J.L. Allen, and P. Pollard. 1994. [Debiasing by instruction: The case of belief bias](#). *European Journal of Cognitive Psychology*, 6(3):263–285.
- Cyn X Fang, Edward Gibson, and Moshe Poliak. 2023. [Individual difference in sentence preferences vs. sentence completion abilities](#).
- Jerry Fodor and Ernie Lepore. 2002. [Why compositionality won't go away: Reflections on horwich's 'deflationary' theory](#). *Meaning and representations*, ed. Emma Borg. Oxford: Blackwell.
- Jerry Fodor and Zenon W Pylyshyn. 1988. [Connectionism and cognitive architecture: A critical analysis](#). *Cognition*, 28(1-2):3–71.
- Gottlob Frege. 1918. [Der gedanke. eine logische untersuchung](#). *Beiträge zur Philosophie des deutschen Idealismus*.
- Edward Gibson and Gregory Hickok. 1993. [Sentence processing with empty categories](#). *Language and Cognitive Processes*, 8(2):147–161.
- Goran Glavaš and Ivan Vulić. 2021. [Is supervised syntactic parsing beneficial for language understanding tasks? an empirical investigation](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3090–3104, Online. Association for Computational Linguistics.
- Adele E Goldberg. 2019. *Explain me this: Creativity, competition, and the partial productivity of constructions*. Princeton University Press.
- Adele E. Goldberg. 2024. [Usage-based constructionist approaches and large language models](#). *Constructions and Frames*, 16(2):220–254.
- HP Grice. 1975. [Logic and conversation](#). *Syntax and semantics*, 3.
- Daniel Grodner and Edward Gibson. 2005. [Consequences of the serial nature of linguistic input for sentential complexity](#). *Cognitive Science*, 29(2):261–290.
- Kurt Gödel. 1931. [Über formal unentscheidbare sätze der principia mathematica und verwandter systeme i](#). *Monatshefte für Mathematik und Physik*, 38:173–198.
- John Hewitt and Christopher D. Manning. 2019. [A structural probe for finding syntax in word representations](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.
- Valentin Hofmann, Leonie Weissweiler, David Mortensen, Hinrich Schütze, and Janet Pierrehumbert. 2024. [Derivational morphology reveals analogical generalization in large language models](#). *Preprint*, arXiv:2411.07990.
- Jennifer Hu, Kyle Mahowald, Gary Lupyan, Anna Ivanova, and Roger Levy. 2024a. [Language Models Align with Human Judgments on Key Grammatical Constructions](#). *Proceedings of the National Academy of Sciences*, 121(36):e2400917121.
- Xinyu Hu, Mingqi Gao, Sen Hu, Yang Zhang, Yicheng Chen, Teng Xu, and Xiaojun Wan. 2024b. [Are LLM-based evaluators confusing NLG quality criteria?](#) In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9530–9570, Bangkok, Thailand. Association for Computational Linguistics.
- Rodney Huddleston and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press.
- Ray Jackendoff. 1992. *Semantic structures*, volume 18. MIT press.

- Ray Jackendoff. 2002. *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford University Press.
- Tom S Juzek. 2024. [The syntactic acceptability dataset \(preview\): A resource for machine learning and linguistic analysis of English](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16113–16120, Torino, Italia. ELRA and ICCL.
- Jerrold J. Katz. 1972. *Semantic Theory*. Harper & Row, New York.
- Najoung Kim and Tal Linzen. 2020. [COGS: A compositional generalization challenge based on semantic interpretation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9087–9105, Online. Association for Computational Linguistics.
- Najoung Kim, Sebastian Schuster, and Shubham Toshniwal. 2024. [Code pretraining improves entity tracking abilities of language models](#). *Preprint*, arXiv:2405.21068.
- Karl Christoph Klauer, Jochen Musch, and Birgit Naumer. 2000. On belief bias in syllogistic reasoning. *Psychological review*, 107(4):852.
- Daria Kryvosheieva and Roger Levy. 2025. [Controlled evaluation of syntactic knowledge in multilingual language models](#). In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 402–413, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Brenden Lake and Marco Baroni. 2018. [Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks](#). In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2873–2882. PMLR.
- Andrew K Lampinen, Ishita Dasgupta, Stephanie C Y Chan, Hannah R Sheahan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. 2024. [Language models, like humans, show content effects on reasoning tasks](#). *PNAS Nexus*, 3(7):pgae233.
- Evelina Leivada, Fritz Günther, and Vittoria Dentella. 2024. [Reply to hu et al.: Applying different evaluation standards to humans vs. large language models overestimates ai performance](#). *Proceedings of the National Academy of Sciences*, 121(36):e2406752121.
- Douglas B. Lenat. 1995. [Cyc: a large-scale investment in knowledge infrastructure](#). *Commun. ACM*, 38(11):33–38.
- Cara Su-Yi Leong and Tal Linzen. 2024. [Testing learning hypotheses using neural networks by manipulating learning data](#). *Preprint*, arXiv:2407.04593.
- Bai Li, Zining Zhu, Guillaume Thomas, Frank Rudzicz, and Yang Xu. 2022. [Neural reality of argument structure constructions](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7410–7423, Dublin, Ireland. Association for Computational Linguistics.
- Bingzhi Li, Lucia Donatelli, Alexander Koller, Tal Linzen, Yuekun Yao, and Najoung Kim. 2023. [SLOG: A structural generalization benchmark for semantic parsing](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3213–3232, Singapore. Association for Computational Linguistics.
- Matthias Lindemann, Alexander Koller, and Ivan Titov. 2024. [Strengthening structural inductive biases by pre-training to perform syntactic transformations](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11558–11573, Miami, Florida, USA. Association for Computational Linguistics.
- Maryellen C. MacDonald, Neal J. Pearlmutter, and Mark S. Seidenberg. 1994. [The lexical nature of syntactic ambiguity resolution](#). *Psychological Review*, 101(4):676–703.
- Aman Madaan, Shuyan Zhou, Uri Alon, Yiming Yang, and Graham Neubig. 2022. [Language models of code are few-shot commonsense learners](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1384–1403, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Gary Marcus. 1998. [Rethinking eliminative connectionism](#). *Cognitive Psychology*, 37(3):243–282.
- Gary Marcus. 2001. *The algebraic mind: Integrating connectionism and cognitive science*. MIT Press.
- James L McClelland and David C Plaut. 1999. Does generalization in infant learning implicate abstract algebra-like rules? *Trends in Cognitive Sciences*, 3(5):166–168.
- R. Thomas McCoy, Shunyu Yao, Dan Friedman, Matthew Hardy, and Thomas L. Griffiths. 2023. [Embers of autoregression: Understanding large language models through the problem they are trained to solve](#). *Preprint*, arXiv:2309.13638.
- Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. [Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.
- Kate McCurdy, Paul Soulos, Paul Smolensky, Roland Fernandez, and Jianfeng Gao. 2024. [Toward compositional behavior in neural models: A survey of current views](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language*

- Processing*, pages 9323–9339, Miami, Florida, USA. Association for Computational Linguistics.
- Tomáš Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. [Efficient estimation of word representations in vector space](#). *Preprint*, arXiv:1301.3781.
- Marvin Minsky and Seymour Papert. 1969. An introduction to computational geometry. *Cambridge tiass.*, HIT 479.480:104.
- Alison Mirin and Paul Christian Dawkins. 2022. Do mathematicians interpret equations asymmetrically? *The Journal of Mathematical Behavior*, 66:100959.
- Kanishka Misra and Kyle Mahowald. 2024. [Language models learn rare phenomena from less rare phenomena: The case of the missing AANNs](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 913–929, Miami, Florida, USA. Association for Computational Linguistics.
- Richard Montague. 1970. [Universal grammar](#). *Theoria*, 36(3):373–398.
- Aaron Mueller, Robert Frank, Tal Linzen, Luheng Wang, and Sebastian Schuster. 2022. [Coloring the blank slate: Pre-training imparts a hierarchical inductive bias to sequence-to-sequence models](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1352–1368, Dublin, Ireland. Association for Computational Linguistics.
- Max Müller-Eberstein, Rob van der Goot, and Barbara Plank. 2022. [Probing for labeled dependency trees](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7711–7726, Dublin, Ireland. Association for Computational Linguistics.
- Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. [Adversarial NLI: A new benchmark for natural language understanding](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.
- Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. [The proposition bank: An annotated corpus of semantic roles](#). *Computational Linguistics*, 31(1):71–106.
- Barbara H. Partee. 1984. [Nominal and temporal anaphora](#). *Linguistics and Philosophy*, 7(3):243–286.
- Andrew Pawley and Frances Hodgetts Syder. 1983. [Natural selection in syntax: Notes on adaptive variation and change in vernacular and literary grammar](#). *Journal of Pragmatics*, 7(5):551–579.
- Jackson Petty, Sjoerd van Steenkiste, and Tal Linzen. 2024. [How does code pretraining affect language model task performance?](#) *Preprint*, arXiv:2409.04556.
- Steven T. Piantadosi. 2024. Modern language models refute chomsky’s approach to language. In Edward Gibson and Moshe Poliak, editors, *From fieldwork to linguistic theory: A tribute to Dan Everett*. Language Science Press.
- Steven T Piantadosi, Joshua B Tenenbaum, and Noah D Goodman. 2016. The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological review*, 123(4):392.
- Martin J. Pickering and Victor S. Ferreira. 2008. [Structural priming: A critical review](#). *Psychological Bulletin*, 134(3):427–459.
- Steven Pinker. 1999. *Words and Rules: The Ingredients of Language*. Basic Books, New York.
- Steven Pinker and Alan Prince. 1988. [On Language and Connectionism: Analysis of a Parallel Distributed Processing Model of Language Acquisition](#). *Cognition*, 28(1-2):73–193.
- Christopher Potts. 2023. Characterizing english preposing in pp constructions. *Lingbuzz Preprint*. <https://lingbuzz.net/lingbuzz/007495>.
- Linlu Qiu, Peter Shaw, Panupong Pasupat, Pawel Nowak, Tal Linzen, Fei Sha, and Kristina Toutanova. 2022. [Improving compositional generalization with latent structure and data augmentation](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4341–4362, Seattle, United States. Association for Computational Linguistics.
- Siva Reddy, Oscar Täckström, Slav Petrov, Mark Steedman, and Mirella Lapata. 2017. [Universal semantic parsing](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 89–101, Copenhagen, Denmark. Association for Computational Linguistics.
- Clarice Robenalt and 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(3):467–503.
- Timothy T. Rogers and James L. McClelland. 2004. *Semantic Cognition: A Parallel Distributed Processing Approach*. MIT Press.
- Hayley Ross, Kathryn Davidson, and Najoung Kim. 2024. [Is artificial intelligence still intelligence? LLMs generalize to novel adjective-noun pairs, but don’t mimic the full human distribution](#). In *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP*, pages 131–153, Miami, Florida, USA. Association for Computational Linguistics.
- David E. Rumelhart, James L. McClelland, and PDP Research Group. 1986. *Parallel Distributed Processing, Volume 1: Explorations in the Microstructure of Cognition: Foundations*. The MIT Press.

- Bertrand Russell. 1905. [On denoting](#). *Mind*, 14(56):479–493.
- Jacob Russin, Sam Whitman McGrath, Danielle J. Williams, and Lotem Elber-Dorozko. 2024. [From frege to chatgpt: Compositionality in language, cognition, and deep neural networks](#). *Preprint*, arXiv:2405.15164.
- Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing*, pages 1–15, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Carson T Schütze and Jon Sprouse. 2013. Judgment data. *Research methods in linguistics*, pages 27–50.
- Peter Shaw, Ming-Wei Chang, Panupong Pasupat, and Kristina Toutanova. 2021. [Compositionality generalization and natural language variation: Can a semantic parsing approach handle both?](#) In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 922–938, Online. Association for Computational Linguistics.
- Dan I Slobin. 1986. The acquisition and use of relative clauses in turkic and indo-european languages. *Studies in Turkish linguistics*, 8:273.
- Taiga Someya, Ryo Yoshida, and Yohei Oseki. 2024. [Targeted syntactic evaluation on the Chomsky hierarchy](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15595–15605, Torino, Italia. ELRA and ICCL.
- Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Many Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2024. [To CoT or not to CoT? chain-of-thought helps mainly on math and symbolic reasoning](#). *Preprint*, arXiv:2409.12183.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, and 431 others. 2023. Beyond the imitation game: quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023(5):1–95.
- P. F. Strawson. 1967. [Is existence never a predicate?](#) *Crítica: Revista Hispanoamericana de Filosofía*, 1(1):5–19.
- Lena Strobl, William Merrill, Gail Weiss, David Chiang, and Dana Angluin. 2024. [What formal languages can transformers express? a survey](#). *Transactions of the Association for Computational Linguistics*, 12:543–561.
- Leonard Talmy. 1975. Figure and ground in complex sentences. In *Annual meeting of the Berkeley linguistics society*, pages 419–430.
- Harish Tayyar Madabushi, Laurence Romain, Dagmar Divjak, and Petar Milin. 2020. [CxGBERT: BERT meets construction grammar](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4020–4032, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Yu-Hsiang Tseng, Cing-Fang Shih, Pin-Er Chen, Hsin-Yu Chou, Mao-Chang Ku, and Shu-Kai Hsieh. 2022. [CxLM: A construction and context-aware language model](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6361–6369, Marseille, France. European Language Resources Association.
- Amos Tversky. 1977. [Features of similarity](#). *Psychological Review*, 84(4):327–352.
- Amos Tversky and Daniel Kahneman. 1974. [Judgment under uncertainty: Heuristics and biases](#). *Science*, 185(4157):1124–1131.
- Johan van Benthem. 2008. A brief history of natural logic. In M. Chakraborty, B. Löwe, M. Nath Mitra, and S. Sarukki, editors, *Logic, Navya-Nyaya and Applications: Homage to Bimal Matilal*. College Publications.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. [Superglue: A stickier benchmark for general-purpose language understanding systems](#). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. [GLUE: A multi-task benchmark and analysis platform for natural language understanding](#). In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.
- Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. [Neural network acceptability judgments](#). *Transactions of the Association for Computational Linguistics*, 7:625–641.
- P. C. Wason. 1968. [Reasoning about a rule](#). *Quarterly Journal of Experimental Psychology*, 20(3):273–281.
- Pia Weißenhorn, Lucia Donatelli, and Alexander Koller. 2022. [Compositionality generalization with a broad-coverage semantic parser](#). In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 44–54, Seattle, Washington. Association for Computational Linguistics.

- Leonie Weissweiler, Taiqi He, Naoki Otani, David R. Mortensen, Lori Levin, and Hinrich Schütze. 2023. [Construction grammar provides unique insight into neural language models](#). In *Proceedings of the First International Workshop on Construction Grammars and NLP (CxGs+NLP, GURT/SyntaxFest 2023)*, pages 85–95, Washington, D.C. Association for Computational Linguistics.
- Leonie Weissweiler, Valentin Hofmann, Abdullatif Köksal, and Hinrich Schütze. 2022. [The better your syntax, the better your semantics? probing pretrained language models for the English comparative correlative](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10859–10882, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Leonie Weissweiler, Abdullatif Köksal, and Hinrich Schütze. 2024. [Hybrid human-LLM corpus construction and LLM evaluation for rare linguistic phenomena](#). *Preprint*, arXiv:2403.06965.
- Michelle J. White, Frenette Southwood, and Sefela Londiwe Yalala. 2024. [Early child language resources and corpora developed in nine African languages by the SADiLaR child language development node](#). In *Proceedings of the Fifth Workshop on Resources for African Indigenous Languages @ LREC-COLING 2024*, pages 86–93, Torino, Italia. ELRA and ICCL.
- Terry Winograd. 1980. [What does it mean to understand language?](#) *Cognitive Science*, 4(3):209–241.
- Ludwig Wittgenstein. 1953. *Philosophische Untersuchungen*. Suhrkamp Verlag, Frankfurt am Main.
- Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. [CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies](#). In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.
- Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, and 43 others. 2017. [CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies](#). In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.
- Shijia Zhou, Leonie Weissweiler, Taiqi He, Hinrich Schütze, David R. Mortensen, and Lori Levin. 2024. [Constructions are so difficult that Even large language models get them right for the wrong reasons](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language*
- Resources and Evaluation (LREC-COLING 2024)*, pages 3804–3811, Torino, Italia. ELRA and ICCL.