Linguistic Generalizations are not Rules: Impacts on Evaluation of LMs

Leonie Weissweiler **Kyle Mahowald**
The University of Texas at Austin
{weissweiler,kyle}@utexas.edu

Adele E. Goldberg
Princeton University
adele@princeton.edu

Abstract

Linguistic evaluations of how well LMs generalize to produce or understand language often implicitly take for granted that natural languages are generated by symbolic rules. According to this perspective, grammaticality is determined by whether sentences obey such rules. Interpretation is compositionally generated by syntactic rules operating on meaningful words. Semantic parsing maps sentences into formal logic. Failures of LMs to obey strict rules are presumed 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 neatly separable, compositional rules. Rather, new utterances are produced and understood by a combination of flexible, interrelated, and context-dependent *constructions*. Considering gradient factors such as frequencies, context, and function will help us reimagine new benchmarks and analyses to probe whether and how LMs capture the rich, flexible generalizations that comprise natural languages.

1 Introduction

How well do large Language Models (LMs) generalize beyond their training data? Much work on this question has presumed that generalizations require symbolic rules for syntax and semantics that generate acceptable new forms and compositional meanings. Rules are invoked to explain that if you learn a new modifier ('blonky') and a new count noun ('gravimin'), a compositional rule could predict that 'a blonky gravimin' is a gravimin that is blonky. In what follows, we use "rule" to refer to context-free generalizations that contain variables, to be instantiated by any instance of a general type, uninfluenced by frequency, similarity, or context (Pinker, 1999). Our focus here is on the use of a strict algebraic conception of rules, which we argue, underlies certain approaches to NLP evaluation, even though the notion of a rule is used variably in linguistics today, with several frameworks incorporating functional and/or frequency-based attributes into representations (e.g., Bresnan et al., 2007; Brehm et al., 2022; O'Donnell, 2015),

Because early statistical models (e.g., *n*-gram or Markov models) seemed unable to generalize fully or capture non-local dependencies (Chomsky, 1957), early on, rules seemed to many to be the only game in town for human language. After all, if a standard bigram model hadn't seen 'blonky gravimin' before, it would be unable to form a representation of 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).

However, current LMs arose from statistical, distributional parallel models (Mikolov et al., 2013; Rumelhart et al., 1986) rather than rule-based natural language technologies. They do not rely on hard-coded rules, yet their 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., 2025). GPT-4o, for example, not only recognizes 'a blonky gravimin' as a noun phrase, it explicitly offers several naturalistic interpretations, e.g., 'A person or act that awkwardly and absurdly pretends to be serious.'

Nonetheless, an assumption that generalizations are equivalent to rules continues to motivate many evaluations of syntax, meaning, and their compositional combination: 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; Dentella et al., 2023) and rule-based compositionality (Kim and Linzen, 2020). Together,

Proceedings of the Second International Workshop on Construction Grammars and NLP, pages 61–74
September 24, 2025,

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 in the early days of LMs was taken to imply that the models did not use language the way people do and were instead merely imitating shallow surface patterns (Bender and Koller, 2020; 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 sizable proportion (39%) believed explicit discrete symbolic rules were required.

Evaluations of LMs' early challenges with algebraic or logical rules did expose certain shortcomings in their ability to reason abstractly and solve math problems (see e.g., Mahowald et al., 2024). At the same time, LM's concurrent ability to produce and respond to natural languages *naturalistically* is hard to overstate (e.g., Coil and Shwartz, 2023).

Mastering a natural language requires mastering a network of hundreds of thousands of context-dependent, gradient, flexible schemata (*constructions*, see §5), which often contain 'slots' that constrain their fillers and how those fillers are interpreted. Constrained slots allow for new combinations, flexibly adapted in context. For instance, the phrase '<time period> ago' can coerce a temporal interpretation of filler phrases that do not designate time periods (e.g., 'three rest stops ago'). Rather than rule-based compositionality, composition-by-construction allows constructions to contribute meaningfully to interpretation in ways that range from abstract to quite narrow and specific. Therefore, for LMs to use language like humans, they require interpretations that are far richer than rules can provide for thousands of collocations, conventional metaphors, idioms, and context-dependent interpretations. Even abstract grammatical patterns also regularly convey semantic and/or pragmatic information that restrict their contexts of use and interpretations. Since different languages and dialects provide speakers with different networks of constructions (ConstructionNets), cross-linguistic differences can be captured naturally.

We suggest that rule-based evaluations have been over-emphasized in the domain of natural language production and comprehension. Our goal is to emphasize the importance of recognizing context, frequencies, meaning and other gradient functional factors in modern evaluations of natural language.

We do not argue that no categorical rule exists in any language. If a categorical rule is needed, it can be treated as the limiting case, a fully abstract construction (Jackendoff, 2002). For instance, Jackendoff (2002) proposes a symbolic Verb + Particle rule for the syntax of English complex verbs. At the same time, the meanings of individual verb plus particle combinations are far from compositional by any general rule (e.g., one can *look up* a number or *look down on someone* but not *?look up on someone* nor *?look down a number*). Here we advocate for an increased focus 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 theoretical points are not new, particularly in the domain of morphology. Neural network researchers have continuously argued in favor of a single representational system and against the usefulness of rules 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; McClelland, 2015). While early work in Artificial Intelligence 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).

Our contribution is to review leading paradigms used in LM evaluations of syntax (§2), semantics (§3), and compositionality (§4). We argue that, while these paradigms have been fruitful, they inherit from a tradition that was overly focused on rules, hierarchy, compositionality, and a binary notion of grammaticality. We briefly characterize how these assumptions arose and how they are baked into evaluations. We argue that **evaluations that presume categorical and strictly compositional language ignore some of the richest elements of human language**. We review construction-based and gradient functionalist approaches to language, arguing that this tradition points to certain lacunae in existing evaluations and open up new possibilities for evaluating natural language understanding. Early work in this direction has already included more nuanced metrics, measuring gradient judgments and context-dependent interpretations (e.g., Juzek, 2024; Hu et al., 2024).

2 Syntactic Rules in LM Evaluations

Evaluations of the syntactic capabilities of LMs have frequently assumed a binary categorical notion of grammaticality, which is then used to create datasets for evaluation. Below, we discuss several such cases, attempting to make these assumptions explicit to show their limitations.

Grammaticality Judgment Tasks Human judgments on sentences are gradient rather than binary, and demonstrably 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 or linguistic theory also influences people's 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. Even sentences included in linguistic textbooks, which one might presume to have clear-cut judgments, in reality are judged gradiently by people (Juzek, 2024). Nonetheless, LMs' language skills are often evaluated on binary grammaticality judgments on sentences (Dentella et al., 2024, 2023; Warstadt et al., 2019).

The fact that human judgments are gradient can have profound consequences on evaluations. 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 measures with the same human judgments revealed a strong positive correlation (Hu et al., 2024). Using perplexity measures for models (as well as allowing humans to provide ordinal or gradient judgments) is a step in the right direction (Hu et al., 2024; Juzek, 2024).

Dependency Parsing as Evaluation Parsing text for universal dependencies (UD, de Marneffe et al., 2021) has become a well-established task for evaluating models (Zeman et al., 2017, 2018), and since Hewitt and Manning (2019) showed BERT (Devlin et al., 2019) to be somewhat skilled in UD, it has 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 and they are based on lexical items, which makes them closer to the approach advocated here rather than abstract phrase structure rules. However, UD analyses presume a universal set of grammatical relations, which is problematic, because not all languages employ the same constructs. That is, there is no universally valid way to define or test for the syntax of nouns, verbs, adjectives, subjects, or direct objects (e.g., Croft, 2001). Moreover, UD requires an asymmetric relationship between a 'head' and dependent, yet the long tail of language includes headless constructions (e.g., *the Xer, the Yer* construction) (Michaelis, 2003) and co-headed constructions (e.g., phrasal verbs, conjunctions, idioms). Therefore, UD annotations need to be determined for individual languages and need to allow for non-headed or co-headed cases to align well with formal aspects of natural languages.

3 Semantic Rules in LM Evaluation

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 was based on algebraic rules operating on categorical and broadly defined categories. Notably, many logicians did not generally assume nor endorse using formal logic to represent the meanings of natural language utterances (Carnap, 1937; Baker and Hacker, 1986), recognizing that natural languages differ from formal logic in many ways.[1] For instance, logic treats *and* and *but* as equivalent. It does not provide a natural way to capture commands or questions (Austin, 1975), nor does it naturally distinguish presuppositions from assertions (Strawson, 1967). Finally, formal logic is not intended to capture effects of context (Wittgenstein, 1953; Russin et al., 2024).

Yet the assumption that natural language semantics can be modeled by formal logic has been made in the design of certain classic LM understanding benchmarks. Below, we review some instances and discuss their connection to formal semantics.

Natural Language Inference Natural Language Inference tasks label the second of two sentences as an entailment, contradiction, or neutral, and this NLI task was originally used to train models (Wang et al., 2019; Dagan et al., 2006; Nie et al., 2020).

[1] Some logicians did advocate for using formal logic for natural language (Tarski, 1944; Montague, 1970, 1973).

Today, NLI is used as a zero-shot evaluation metric to assess natural language understanding (Zhou et al., 2024; McCoy et al., 2019). In introducing the Stanford NLI corpus, Bowman et al. (2015) state, "The semantic concepts of entailment and contradiction are central to *all* aspects of natural language meaning. ") (emphasis added, see also Katz, 1972; van Benthem, 2008). While in the same paper, Bowman et al. (2015) acknowledge that judgments depend on many factors, such as commonsense knowledge, this fact is generally overlooked in papers that use NLI as a task to evaluate LMs' general understanding.

Necessary and plausible inferences *are* a critical aspect of natural language understanding. However, they are highly dependent on the interlocutors' general communicative goals. We aim 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 but he is not DEPRESSED', listeners do not throw up their hands and shout 'contradiction!'. Instead, they may infer that the boy in question is only somewhat, and not extremely, depressed, or ask to learn more. People also assign interpretations to statements in ways that differ from what formal logic would predict (e.g., 'run fast and you've got this' or 'If it snows, it snows.' NLI tasks that rely on judging contradictions or entailments may over- or under-estimate how well LMs' understanding of natural language aligns with humans', particularly when binary judgments are required (cf. Dentella et al., 2024).

Evaluation metrics need to take humans' communicative goals into account and allow for gradient and context-dependent interpretations. An example of the type of evaluation we endorse can be found in the underappreciated NOPE testbed. Parrish et al. (2021) selected 10 distinct constructions that trigger presuppositions and curated 100+ instances of each one, based on naturally occurring examples. Each stimulus includes two preceding sentences for context. The authors then collected gradient judgments from human raters, allowing them to use their 'background knowledge about how the world works' and compared the accuracy of several models, with several controls in place. This strikes us as a highly valuable blueprint for modern evaluations of LMs.

Semantic Parsing Banarescu et al. (2013) introduced abstract graphical meaning representations (AMR) for sentential meaning that importantly includes aspects of lexical semantics. It was created to offer a repository of structured meanings to be used for evaluating understanding in LMs (Li et al., 2023; Qiu et al., 2022; Shaw et al., 2021, see also §4). Yet work on AMR concedes that it ignores so-called 'syntactic idiosyncracies.' For example, 'he described her as a genius' and 'his description of her: genius' are assigned the same AMR. Yet the former is unambiguously about her intellect, while the latter may instead be used to compliment *his cleverly tactful description*. More generally, simplifications of distinctions made in a natural language can be expected to result in lost meaning, since two utterances are rarely interchangeable in all contexts. Focusing on subtle but important differences in meaning offers an opportunity to design more challenging linguistic evaluations of LMs.

4 Compositionality

As computer coding languages became more and more widespread, rule-based syntax and semantics took root in linguistics. A Principle of Compositionality states that the semantics 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). It is intended to be a bottom-up process: syntactic rules combine words, which have determinant meanings. Fodor and Pylyshyn (1988) make this clear: "a lexical item must make approximately the same semantic contribution to each expression in which it occurs". That is, context may not influence the interpretation of words in a top-down manner; therefore downstream inferences are required to address the fact that interpretations *do* depend on context. Realizing this, like Carnap and Frege before him, Fodor and Pylyshyn (1988) acknowledge: "It's uncertain exactly how compositional natural languages actually are."

Nonetheless, compositionality is often taken as a truism, based on the standard argument for it summarized below.

> 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.

Note that one can agree with the premise without accepting the consequent. In particular, people generally agree on the interpretations of pointing gestures and novel words as well as sentences, and yet shared interpretations *must* be gleaned from non-linguistic context in the case of pointing gestures, and from a combination of linguistic and non-linguistic context in the case of novel words. Shared interpretation of sentences likewise comes in part from linguistic and non-linguistic context.

Consider the sentence, 'the Persian cat is on the mat.' If the speaker's goal is simply to help someone find the furball, there need be no commitment to the cat being a thoroughbred Persian breed nor to the cat being wholly on, rather than adjacent to, the mat. Or, comprehenders may appreciate the statement is ironic if the cat is hairless.

Cases that might seem amenable to a rule often turn out to require a good deal of item-specific memory. For instance, a compositional rule involving set-intersection may seem 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: 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 (or ought to). The meanings of familiar collocations are typically not fully determined by general compositional rules, and novel cases can be interpreted on analogy to familiar cases rather than according to some very general rule. For instance, if "flam" is interpreted to mean any kind of event or action, *a green flam* is likely to be interpreted to imply an eco-friendly or beginning-level event. Representing only the rule-compliant cases in evaluations can therefore lead to the wrong conclusions. A more comprehensive evaluation paradigm should take into account how people actually interpret familiar and novel cases.

Another issue is that rules massively overgenenerate. 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 them the problem.' Humans are sensitive to the frequencies of various types of word combinations and judge formulations unnatural if 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 syntactic and semantic evaluations. Kim and Linzen (2020)'s compositional generalization challenge (COGS) tested whether models could translate any sentence generated by a small set of syntactic rules into formal semantics. For instance, trained on representations of 'the girl,' 'the cat,' 'the hedgehog,' 'the cat loves the girl,' 'the hedgehog sees the cat,' and so on, the model was tested on how well it predicted a formal semantic representation of 'The girl loves the hedgehog.' However, note that if 'mosquitoes' is substituted for 'the cat,' different interpretations of 'love' are evoked ('Mosquitoes love the girl' vs. 'The girl loves mosquitoes'), not to mention different degrees of plausibility. The authors also anticipated generalizations from sentences like 'Jane gave the cake to John' to 'Jane gave John the cake,' and the models were found to perform poorly. Yet the two sentences differ in terms of information structure (Bresnan and Ford, 2010) and the relative frequencies and similarities of verbs witnessed in each version (Leong and Linzen, 2024; Ambridge et al., 2014; Hawkins et al., 2020). Thus, while an evaluation of this kind can capture something about how humans interpret automatically generated sentences in an experimental context, focusing on this type of task may distort our view of how well LMs handle natural language in the wild.

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 apply to natural language generally. 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'), which has a very different meaning than performing an action two times. Other cases require knowledge of specific combinations: 'to think twice' 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 tend to 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 general algebraic rules.

Another example comes from the seemingly innocuous algebraic rule: "If X is more Y than Z, then Z is less Y than Z, irrespective of the specific meanings of X, Y, and Z" (Dasgupta et al., 2020, :5). This is meant 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 matters. 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. We have so far argued that an overemphasis on symbolic abstract rules for natural languages can lead to evaluations of natural language that are not aligned with humans. Below we suggest an alternative approach to language, which we argue helps refocus evaluations on interesting new research questions.

5 The Constructionist Approach

This section briefly explains the constructionist approach to language, which conceives of a language as a vast network of interrelated *constructions*, of varying size and complexity. This differs from a perspective that treats languages as a set of sentences generated by a small set of algebraic rules. We suggest a change of perspective about the nature of language, not a mere substitution of the units on which some type of rules operate. That is, certain traditional evaluations were far too limited in requiring models to adhere to strict compositionality, when humans do not. At the same time, the constructionist approach encourages stringent evaluations by testing whether models capture the gradient and function-sensitive patterns that characterize natural languages.[2] The approach encourages us to broaden our view of language and linguistic evaluations of LMs.

(Partially-filled) Words, Common & Rare Schemata as the same type of Representations A 'construction' is any 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. As a result, the lexicon and syntax are not treated as distinct or modular systems. This allows the many parallels between them to be easily captured. It also allows a natural way to allow for the diversity found in the world's languages, in which more or less information is encoded in a single word. Formal attributes of constructions include phonology, grammatical categories, word order, discontinuous elements, specific words or morphemes, and/or intonation. Any construction may include one or more constrained open 'slots'.

A Wide Range of Functions Considered Jointly with the Forms Constructions' functions vary widely: words, collocations, and idioms convey rich, specific, contentful meaning. A plethora of other constructions are productive but constrained in a variety of semi-specific ways; argument structure constructions convey 'who did what to whom'; discourse structuring constructions indicate which parts of an utterance are at-issue or backgrounded. A range of constructions exist to ask questions, express surprise or disapproval, greetings or gossip. Construction can be associated with specific registers, genres, and/or dialects. The constructionist commitment to considering semantics jointly with syntax represents a more comprehensive understanding of their interactions, which can help develop tests that evaluate both.

Sensitivity to Similarity and Frequency Language users are sensitive to the frequencies of constructions. For instance, the passive construction is used far more frequently in Turkish than English and young Turkish speakers learn the construction far earlier than English-speaking children (Slobin, 1986). Constructions are also influenced by simi-

[2]For more comprehensive introductions to the constructionist approach, or 'Construction Grammar', see Hoffmann and Trousdale (2013) and Hoffmann (2022).

larity: Instances of a construction prime instances of the same or closely related construction (e.g., Du Bois, 2014; Pickering and Ferreira, 2008). Constructionist approaches take this to be a core aspect of language and language learning, rather than an inconvenience or afterthought. This leads to a deemphasis of definitional boundaries and an organic incorporation of fuzzy boundaries and prototypicality effects.

Productive Constructions May Include Fixed Lexical Units Syntax, semantics, and morphology are interrelated rather than assigned 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.').

Interrelated Network, Not an Unstructured Set Unlike rules, which are commonly presented as unstructured lists, constructions comprise a network of interrelated patterns. This allows for the fact that each language includes 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').

More Maximal than Minimal A ConstructionNet includes words as well as grammatical patterns, and lossy instances are included as well as generalizations across instances, as just mentioned, which provides some redundancy. There is no reason to restrict the complexity of constructions or their descriptions more than is warranted by psychological and linguistic evidence.

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 participle 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' would also require an unusual context to make sense. Finally, instances of the construction are not amenable to a general compositional rule, nor can they be translated into formal logic. Either approach 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

Outside of natural language, even in domains that are rule-like by design, rule-based interpretations are sometimes lacking, potentially due to the fact that natural language is used by people when discussing these domains. 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 rarely reversible without a different interpretation: e.g., '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, which is not symmetric.

7 New Directions for Evaluation

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. The complexity of constructions with respect to gradience in frequencies, functions, slot constraints, and prototypicality can be used to develop evaluations that demand the same complexity from LMs found in natural languages.

A caveat is required for low-resource languages, where rule-based linguistic evaluations (e.g., Jumelet et al., 2025) can be useful. More generally, evaluations should meet models where they are: if the representational complexity of an LM is restricted, restricted types of evaluations are required. But when evaluating LMs on high-resource languages, richer evaluations are appropriate. Specifically, we recommend the following.

When possible, use a variety of naturalistic sentences rather than sentences generated by a template that presupposes grammatical rules with interchangeable vocabulary items, as is done, e.g., by Multi-NLI (Williams et al., 2018). The idea that sentences can be constructed by subbing random lexical items into templates often misses lexical subtleties that are an important part of natural language. Instead, ecologically valid stimuli can be collected or adapted from natural corpora and normed for naturalness and plausibility. Since human judgments are highly context-dependent, benchmark tasks should also vary contexts systematically (see, e.g., Ross et al., 2024; Parrish et al., 2021).

Collect human assessments that allow for gradient context-sensitive interpretations that appeal to learned constructions. Evaluating LM competence on individual constructions requires assessing both acceptability judgments and interpretations from humans to draw appropriate comparisons.

We also need to be sensitive to the implications people and LMs draw from instructions and the testing context. **Do not give instructions to human-evaluators (or models) in ways that make the results a foregone conclusion.** If people are instructed to interpret 'red X' as 'X that is red for any X,' they are capable of doing so; this may reflect the instructions, not their natural intuitions. 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.

The items included in testing also influence interpretations by people and models, by providing context. For instance, if certain pairs of items are mismatched (e.g., "The cup is green." "The cup is blue.") while others are matched, people can infer which ones are intended to be contradictory. LMs, like people, are now capable of generalizing by rule when tasked to do so. For instance, Lampinen et al. (2025) found Gemini 1.5 Flash (Team et al., 2024) avoided the reversal curse, achieving 100% accuracy, when the Berglund et al. (2023) dataset was provided to the LM as context.

A variety of items, participants, and contexts ought to be valued as much as a variety of models. It has long been recognized that real words, phrases and sentences vary in an open-ended number of ways (Clark, 1973). So care must be taken to include a variety of stimuli items. Because linguistic meaning is deeply tied to local context, even seemingly similar sentences can have very different interpretations in ways that depend on context. Different subgroups of participants may perform differently so distinct dialects should be taken into account.

The most interesting questions may no longer be *whether* LMs are skilled at producing and responding to natural languages, but *how* they achieve such remarkable skills. As is familiar from the lexicon, constructions comprise an interrelated network. We can now **how relationships between constructions are picked up by LMs.** For instance, 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. Moreover, nearly every productive construction coexists with at least a few formulaic instances, and LMs offer ways to test various theoretical perspectives on the nature of those relationships. These and other newer ways of probing LMs are possible, and our toolkit will only grow.

As discussed by Weissweiler et al. (2023), investigating whether LMs distinguish subtle meaningful differences between constructions is another important direction. Recent work on this has included Weissweiler et al. (2022), who found LMs reliably discriminated 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. As mentioned earlier, 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 (ASCs) and found similarities in behavior in the model and a sorting task done by humans. 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').

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. Natural languages are not Lego sets. Instead, language involves flexible combinations of rich and varied constructions of differing sizes, complexities, 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 only begun to be explored. We believe future progress lies not in asking whether LMs obey abstract rules, but in probing what kinds of constructions they learn, how they relate them, and how those structures guide novel interpretation and production. In doing so, we may better capture what it truly means to comprehend and use language.

Limitations

While we have aimed to discuss benchmarks and evaluations in ways that reflect the historical trajectory as well as the present-day landscape, evaluations of LMs are continually developing. We feel the dominant paradigms have and continue to be based on data generated by rules and evaluated without regard for context effects, gradience, or semantic nuance, but we are keenly aware that we have likely overlooked metrics that go beyond rule-based evaluations (e.g., Parrish et al., 2021).

We recognize the growing work in multilingual evaluations, which are inherently valuable (Mueller et al., 2020; Jumelet et al., 2025; Kryvosheieva and Levy, 2025). The current perspective applies to all natural languages, but comparative work is not the focus of the current perspective, and we use English examples for the sake of easy comprehension and brevity.

Acknowledgments

We thank Najoung Kim, 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 summmer 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 Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. 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*.

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.

Laurel Brehm, Pyeong Whan Cho, Paul Smolensky, and Matthew A. Goldrick. 2022. Pips: A parallel planning model of sentence production. *Cognitive Science*, 46(2):e13079.

Joan Bresnan, Anna Cueni, Tatiana Nikitina, and R Harald Baayen. 2007. “Predicting the Dative Alternation”. In Gerlof Bouma, Irene Krämer, and Joost Zwarts, editors, *Cognitive Foundations of Interpretation*, pages 69–94. KNAW.

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.

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.

Herbert H. Clark. 1973. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 12(4):335–359.

Albert Coil and Vered Shwartz. 2023. From chocolate bunny to chocolate crocodile: Do language models understand noun compounds? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2698–2710, Toronto, Canada. Association for Computational Linguistics.

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.

Vittoria Dentella, Fritz Günther, Elliot Murphy, Gary Marcus, and Evelina Leivada. 2024. Testing ai on language comprehension tasks reveals insensitivity to underlying meaning. *Scientific Reports*, 14(1):28083.

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.

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 sentenial 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.

Robert Hawkins, Takateru Yamakoshi, Thomas Griffiths, and Adele Goldberg. 2020. Investigating representations of verb bias in neural language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4653–4663, Online. Association for Computational Linguistics.

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.

Thomas Hoffmann. 2022. *Construction Grammar*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Thomas Hoffmann and Graeme Trousdale. 2013. *The Oxford handbook of construction grammar*. Oxford University Press.

Valentin Hofmann, Leonie Weissweiler, David R. Mortensen, Hinrich Schütze, and Janet B. Pierrehumbert. 2025. Derivational morphology reveals analogical generalization in large language models. *Proceedings of the National Academy of Sciences*, 122(19):e2423232122.

Jennifer Hu, Kyle Mahowald, Gary Lupyan, Anna Ivanova, and Roger Levy. 2024. Language Models Align with Human Judgments on Key Grammatical Constructions. *Proceedings of the National Academy of Sciences*, 121(36):e2400917121.

Ray Jackendoff. 1992. *Semantic structures*, volume 18. MIT press.

Ray Jackendoff. 2002. *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford University Press.

Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2025. MultiBLiMP 1.0: A multilingual benchmark of linguistic minimal pairs. *Transactions of the Association for Computational Linguistics*. To appear.

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.

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, Arslan Chaudhry, Stephanie C. Y. Chan, Cody Wild, Diane Wan, Alex Ku, Jörg Bornschein, Razvan Pascanu, Murray Shanahan, and James L. McClelland. 2025. On the generalization of language models from in-context learning and finetuning: a controlled study. *Preprint*, arXiv:2505.00661.

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.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6):517–540.

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. 2015. Capturing gradience, continuous change, and quasi-regularity in sound, word, phrase, and meaning. *The handbook of language emergence*, pages 53–80.

R. Thomas 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.

Laura A Michaelis. 2003. Headless constructions and coercion by construction. *Mismatch: Form-function incongruity and the architecture of grammar*, pages 259–310.

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.

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.

Richard Montague. 1973. The proper treatment of quantification in ordinary english. In K. J. J. Hintikka, J. M. E. Moravcsik, and P. Suppes, editors, *Approaches to Natural Language: Proceedings of the 1970 Stanford Workshop on Grammar and Semantics*, pages 221–242. Springer Netherlands, Dordrecht.

Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. 2020. Cross-linguistic syntactic evaluation of word prediction models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5523–5539, Online. 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.

Timothy J O'Donnell. 2015. *Productivity and reuse in language: A theory of linguistic computation and storage*. MIT Press.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Alicia Parrish, Sebastian Schuster, Alex Warstadt, Omar Agha, Soo-Hwan Lee, Zhuoye Zhao, Samuel R. Bowman, and Tal Linzen. 2021. NOPE: A corpus of naturally-occurring presuppositions in English. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 349–366, Online. Association for Computational Linguistics.

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.

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.

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.

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. Compositional 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.

P. F. Strawson. 1967. Is existence never a predicate? *Crítica: Revista Hispanoamericana de Filosofía*, 1(1):5–19.

Leonard Talmy. 1975. Figure and ground in complex sentences. In *Annual meeting of the Berkeley linguistics society*, pages 419–430.

Alfred Tarski. 1944. The semantic conception of truth: and the foundations of semantics. *Philosophy and Phenomenological Research*, 4(3):341–376.

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.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, and 1118 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

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.

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.

Pia Weißenhorn, Lucia Donatelli, and Alexander Koller. 2022. Compositional 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.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

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.