Structural Generalization of Modification in Adult Learners of an Artificial Language

Najoung Kim (najoung@bu.edu)
Department of Linguistics, Boston University

Paul Smolensky (smolensky@jhu.edu)
Department of Cognitive Science, Johns Hopkins University
Microsoft Research

Abstract

Compositional generalization that requires production and comprehension of novel *structures* using observed constituent parts has been shown to be challenging for even very powerful neural network models of language. However, one of the test cases that poses the greatest difficulty—generalization of modifiers to unobserved syntactic positions—has not been empirically attested in human learners under the same exposure conditions assumed by these tests. In this work, we test adult human learners on whether they generalize or withhold the production of modification in novel syntactic positions using artificial language learning. We find that adult native speakers of English are biased towards producing modifiers in unobserved positions (therefore producing novel structures), even when they only observe modification in a single syntactic position, and even when the knowledge of their native language actively biases them against the plausibility of the target structures.

Keywords: artificial language learning, inductive bias, modification, compositional generalization, structural generalization, modifier generalization

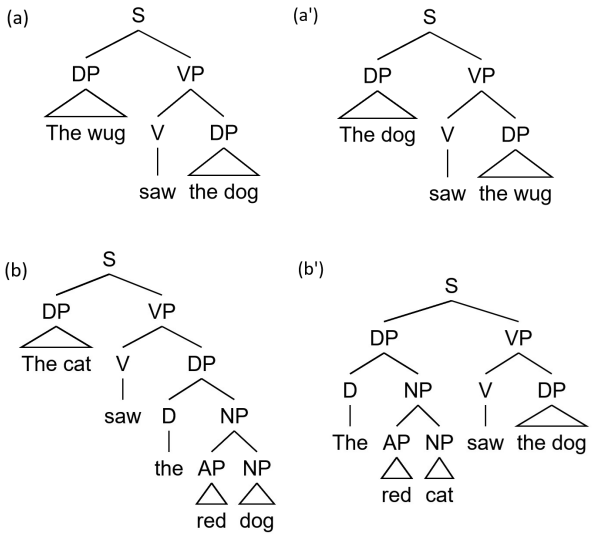

Figure 1: (a, b) Structures seen during the training phase. (a′, b′) Generalization targets. Both targets can be constructed from recombinations of parts of (a) and (b), respectively. Generalization to (a′) from (a) is a case of lexical generalization, where no new novel structures need to be constructed. On the other hand, generalization to (b′) from (b) is a case of structural generalization, where the structure of the target has not been observed during training. The tree structures shown are illustrative; they are not provided as parts of the input.

Motivation

Human linguistic capacity is often characterized by compositionality that enables generalization to novel complex utterances through composition of their constituent parts (Montague, 1970; Partee, 1984). This ability has been at the center of a longstanding debate surrounding connectionist models of the mind (Fodor & McLaughlin, 1990; Fodor & Pylyshyn, 1988; Hadley, 1994; Smolensky, 1995, *i.a.,*), with resurging interest in the evaluation of such a capacity in light of the rapid development of neural network models for language (Bastings, Baroni, Weston, Cho, & Kiela, 2018; Kim & Linzen, 2020; Lake & Baroni, 2018; Li et al., 2023). One important observation highlighted in recent work is the large performance gap between lexical and structural generalization, where lexical generalization targets an unobserved combination of a known lexical item and a known linguistic structure and structural generalization targets extrapolation to unobserved structures. For example, understanding *The wug saw the cat* based on prior observations of *The cat saw the dog* and *The dog saw the wug* (where *wug* has not been observed in the subject position) is an instance of lexical generalization because the generalization target does not require constructing a novel structure (e.g., Figure 1, (a) $\rightarrow$ (a′)). On the other hand, generalizing to an embedded structure of depth $n+1$ from prior observations of depths up to n is an instance of structural generalization because the generalization target requires constructing a novel structure. According to Li et al. (2023), neural network models—both models trained from scratch and models pretrained on a large amount of language data—perform almost perfectly on lexical generalization while failing on structural generalization.

Our work focuses on a particular type of structural generalization from Kim and Linzen (2020) (K&L henceforth). The specific test we examine is modification of noun phrases (NPs) in different syntactic positions (Figure 1, (b) $\rightarrow$ (b′)). Empirically, even very powerful models struggle on this type of generalization, and only models with augmentation in-

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

volving explicit structural cues (e.g., intermediate parsing, grammar induction) can adequately generalize (Drozdov et al., 2023; Li et al., 2023; Qiu et al., 2022). Generalization of modification to unobserved syntactic positions, in its general formulation, is intuitively expected for human learners; it is impossible to observe modification in every plausible position that an NP can occur in (e.g., subject of main verb, direct/indirect object of main verb, subject of the first embedded clause...) before concluding that modified NPs are generally licensed where unmodified NPs are licensed.

However, modifier generalization as formulated in K&L assumes a very extreme exposure condition, where a learner is expected to generalize based on exposure to modification in a *single* syntactic position: modification of the direct object NP of a transitive verb that is not embedded. With only this limited exposure, the learner is expected to generalize to modification of subject NPs. Whether this is a fair test has been questioned (Wu, Manning, & Potts, 2023), since the training data does not disambiguate between whether modifiers should be generalized or should be restricted to the observed positions. Our view, on the other hand, is that the existence of multiple plausible competing hypotheses given the training data is not problematic—this is in fact exactly the point of generalization tests that adopt the "Poverty of the stimulus" experimental paradigm (Wilson, 2006) where the goal is to tease apart learners' inductive biases in the presence of ambiguous hypotheses. Rather, the issue is that there is no attestation of the target generalization in human learners assuming such an extreme exposure condition. While there indeed is a naturally occurring frequency gap between subject and object modification in child-directed speech (K&L, Appendix B), human generalization patterns in the total *absence* of evidence in positions elsewhere (**elsewhere generalization** henceforth) has yet to be empirically tested.

To test the human generalization patterns, we conduct artificial language learning studies with adults (Brown, Smith, Samara, & Wonnacott, 2022; Culbertson, Smolensky, & Legendre, 2012; Martin, Ratitamkul, Abels, Adger, & Culbertson, 2019; Morgan, Meier, & Newport, 1987; Morgan & Newport, 1981, *i.a.,*) with gaps in the distribution of modification across training and testing stimuli. Our experiments show that human learners do indeed exhibit a bias towards extrapolating modification elsewhere, producing novel linguistic structures with modification in unobserved syntactic positions, even in the absence of observed modification in any other positions than a direct object of an unembedded transitive verb. Furthermore, this conclusion holds even when the semantics of the modification is something that is impossible to express as NP modification in English, the participants' native language.[1]

[1]We note that the current work does not test the exact type of modification tested in K&L which are PPs: we discuss the reason and implications in the Limitations section.

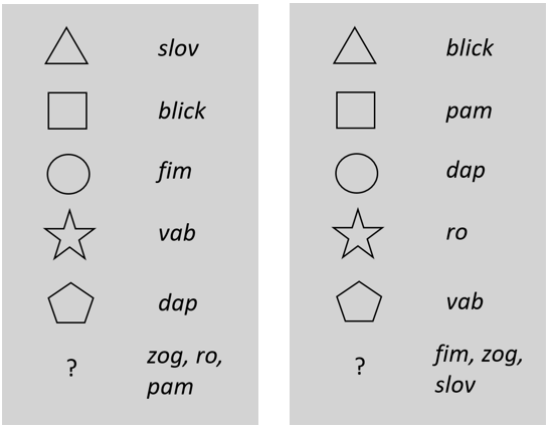

Figure 2: Examples of the on-screen lexicon. The mappings were shuffled randomly for different lists.

Methods

Overall design

We adopt the design of K&L proposed for testing neural network models, originally inspired by the poverty of the stimulus method of Wilson (2006) for testing human learning biases. Our design is the most similar to Lake, Linzen, and Baroni (2019) where meanings are part of the input accessible to participants (as opposed to acceptability judgments on surface forms only, as in the related work of McCoy, Culbertson, Smolensky, and Legendre (2021) on $n+1$ structural generalization of center embedding depth), because this setup better echoes the original tests of K&L. The experiments consist of two phases: we first expose participants to a set of stimuli containing sentences in an artificial language (the training phase), and test whether the participants produce sentences with structures unobserved during training (the testing phase). Crucially, the target sentences in the testing phase can be constructed by *recombinations* of parts of the sentences that the participants are exposed to during the training phase (Figure 1, (b) $\rightarrow$ (b$'$)).

The artificial language

The lexicon of our artificial language consists of eight nonce vocabulary items: *blick, fim, vab, slov, dap, ro, zog, pam*. Five of these are nouns referring to the shapes square, circle, star, triangle, and pentagon. The language has two verbs: one transitive verb roughly corresponding to the English verb *hit*, and one intransitive verb that roughly means *hop multiple times*. The one remaining word is a (postnominal) adjective, which took on different semantics in the two experiments to be discussed in later sections. The postnominal adjective appears hyphenated to the noun (e.g., *fim-zog*, if *zog* is the adjective and *fim* is the noun). The mapping between nonce words and the denotations were randomly determined—we used two lists per experiment (between-subjects design), where the form-meaning mappings were shuffled for each list. Our lan-

guage has VSO word order with postnominal modification to make the language substantially different from our participants' native language, English, to block the targeted generalizations being extended from their native language.

Experiment protocol

Each example shown in the training phase consisted of a sentence and a short video depicting an action involving one or two shapes. In the two-shape scenes, one of the shapes corresponded to the grammatical subject of a transitive verb, and the other, the grammatical object. The lexicon providing mappings between the shape and the shape's name in the artificial language was always displayed on screen to reduce the memorization burden.[2] The lexicon also included other non-shape-denoting lexical items, paired with a question mark (Figure 2). In the training phase, the participants were initially asked to guess what the corresponding artificial language sentence would be for a given scene. If their answer was incorrect, the correct answer was shown on screen, and the participant was asked to type the correct expression out. They could not proceed to the next example unless they provided the exact answer. In the testing phase, the participants were likewise shown a scene and asked to produce a corresponding description in the language they learned. However, unlike the training phase, there was no feedback during the test phase—their responses were simply recorded. All examples within the respective phases were presented randomly.

The test examples included three different types of targets: seen, unseen but not structurally novel, and unseen and structurally novel. Seen examples were repeated scenes from the training set—these examples were used to ensure that the participants succeeded at the learning task. We only considered learning as successful if perfect accuracy was achieved on the seen examples; the decision to use perfect accuracy as the threshold was due to the small size of of this set ($n = 5$). The unseen but not structurally novel cases refer to examples with target productions that were not shown during training but does not require constructing novel structures. For example, if the training set contained *The cat saw the red dog*, *The dog saw the red cat* would be an instance of a target production that is unseen but identical structurally to an example in the training set. This is a case of lexical generalization (Figure 1, (a) $\rightarrow$ (a$'$)). Finally, the unseen and structurally novel cases were our main target—these are production targets with structures that are not part of training. For example, if the training set contained *The cat saw the red dog*, then *The red cat saw the dog* would be an unseen production target with a novel structure (Figure 1, (b) $\rightarrow$ (b$'$)).

The examples in the training set were constructed using the templates `V N N-A`, `V N N`, and `V N`, the first two templates corresponding to transitive and the last template corresponding to intransitive constructions. The structurally novel production targets for the test phase had the form `V N-A N` or `V N-A N-A`. See Table 1 for a summary of the templates.

[2]The rate of participants who provided correct answers to examples already shown during training was very low in pilot experiments that did not include the lexicon, showing that memorization is a bottleneck to the success of the learning task.

Table 1: Templates for the examples in the training set and the target productions in the test set.

Phase	Type	Template
Training	Transitive, modified object	`V N N-A`
	Transitive, bare object	`V N N`
	Intransitive	`V N`
Testing	Seen & Unseen, not structurally novel	`V N N-A`
		`V N N`
		`V N`
	Unseen, structurally novel	`V N-A N`
		`V N-A N-A`

Participant recruitment

We recruited our participants on Prolific. Adult native speakers of English based in the US with a task approval rating of 97% or higher were included in the recruitment pool. The participants were compensated at an hourly rate of $13.50 with an estimated completion time of 20 minutes per experiment (the actual median completion time was around 16–17 mins). We compensated all participants who completed the experiment, regardless of inclusion in the final analysis.

Experiment 1: Color Modification

Stimuli

In this experiment, the postnominal adjective was a color modifier. We sampled 20 training examples (18 transitives and 2 intransitives) using the templates in Table 1, ensuring that the participants were exposed to all vocabulary items during the training phase. The transitive constructions may or may not have contained modification (`V N N-A` or `V N N`), but when they did, they were always object modification. The test set contained 14 examples: 5 seen, 5 unseen but not structurally novel, and 4 unseen and structurally novel. Two of the unseen and structurally novel test examples targeted only subject modification (`V N-A N`) and other two targeted both subject and object modifications (`V N-A N-A`). The videos corresponding to the transitive constructions showed one shape moving across the screen to hit the other shape (Figure 3). The shapes were either white or blue, and during training, the modifier was present when a shape was blue. We created two lists for this experiment, where each list contained different form-meaning mappings of the nonce vocabulary (Figure 2).

Participants

We recruited 73 participants on Prolific. 5 participants' data were excluded either due to submission errors or incomplete responses, yielding a total of 68 participants.

Results

The rate of successful learners—learners who produced the correct answer for all test examples that were repeated from

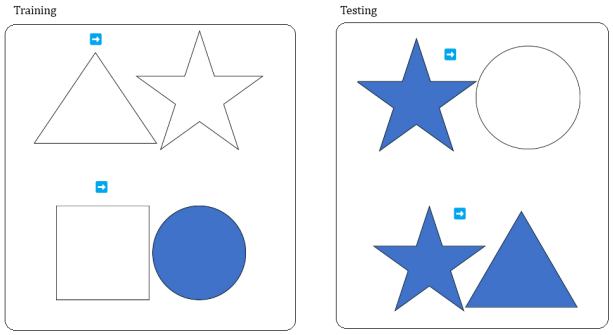

Figure 3: Example scenes from Experiment 1. The scenes were animated for the actual experiments; the blue arrows mark the shapes that are moving towards the other shape (the shapes that move are the grammatical subjects of the transitive verb). The blue arrows were not part of the actual video. In the depicted experiment, the star was never seen as the grammatical subject during the training phase, and it was never shown combined with a modifier. Additionally, there were intransitive scenes with only a single shape (depicting the shape hopping up and down) shown during both training and testing phases not included in this figure.

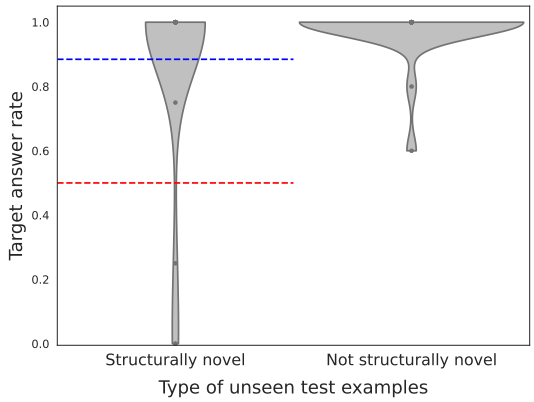

Figure 4: The distribution of target answer rates (# target/# total) in Experiment 1. The blue dotted line indicates the mean target answer rate across all participants, and the red dotted line indicates chance-level production of the target novel structures, assuming a competition between two generalization hypotheses: modification can occur in unobserved positions vs. modification cannot occur in unobserved positions.

training—was 38% (26/68). For these successful learners, we analyzed the target answer production rate for the two unseen test sets (not structurally novel and structurally novel), aggregating over all responses ignoring minor, unambiguous typos (e.g., *slove* instead of *slov*). The successful learners' target production rate was near perfect on unseen but not structurally novel test examples (98%). For the test set with structurally novel targets, the aggregate target answer production rate was significantly over chance (88%, $n = 104, z = 7.751, p < .001$, test for one proportion assuming 50% as chance level).[3] The results are visualized in Figure 4. There was no significant difference in the target production rates across two lists ($\chi^2 = 2.653, DF = 1, p = .10$, $n-1$ χ^2 test).

Furthermore, from the successful learners, an overwhelming majority (85%; 22/26) produced either the exact target answers for all of the unseen, structurally novel test cases (i.e., 100% target production rate for test cases targeting the form `V N-A N` or `V N-A N-A`; $n = 20$), or answers structurally identical to target answers barring minor typos ($n = 2$).

Experiment 2: Modification with Resultative Semantics

The color modification in Experiment 1 raises the question of transfer effect from the participants' native language, since English also uses adjectival modification to describe the property of color attached to a nominal referent (e.g., *the blue star*). Were the participants in Experiment 1 producing adjectival modification on subjects by generalizing from modifications of objects in the artificial language, or extending subject modification in English? Experiment 2 attempts to address this question and show stronger evidence for bias towards structural generalization by using semantics that cannot be expressed as adjectival modification in English. Specifically, the artificial language in Experiment 2 assigned resultative semantics to the modifier, keeping the language's syntax constant. In English, a resultative construction generally cannot be used to express the meaning that the subject is an undergoer of change as a result of the action denoted by the predicate (e.g., *The man wiped the table exhausted* or *The man wiped exhausted the table* to mean that the man became exhausted as a result of wiping the table) in transitive contexts; resultative constructions are limited to direct objects (e.g., *The man wiped the table clean*) (Levin & Rappaport Hovav, 1995).[4] Therefore, in this experiment, the knowledge of the participants' native language actively biases them *against* producing the target structure of subject modification. If we do still find similar trends to Experiment 1 despite this adversarial design, it would strongly support the subjects' inductive bias towards the elsewhere generalization for modification.

Stimuli

In this experiment, the postnominal adjective was a modifier that expressed resultative semantics; namely, cracking (or

[3] We used 50% to represent chance level, but since the task was free form elicitation rather than binary choice, the actual chance level is presumably much lower. Here we took 50% as chance assuming a narrow set of two competing hypotheses: modification can vs. cannot occur in unobserved positions.

[4] Subject resultatives are not universally implausible; for example, Korean allows for subject resultatives (Hong, 2006). Whether resultatives should be analyzed as involving adjectival modification is a topic of debate, but regardless of what kind of syntactic analysis one adopts, the current claim that subject resultatives cannot be expressed as adjectival modification in English holds.

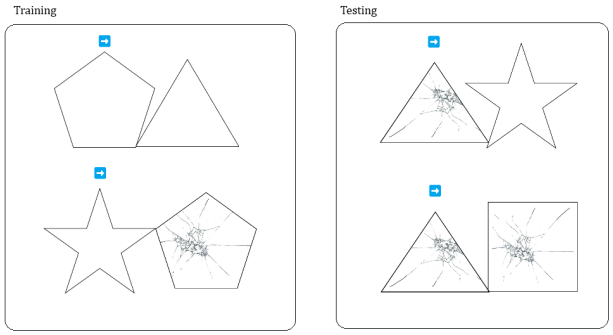

Figure 5: Example scenes from Experiment 2. The scenes were animated for the actual experiments; the blue arrows mark the shapes that are moving towards the other shape (the shapes that move are the grammatical subjects of the transitive verb). The blue arrows were not part of the actual video. The cracks on the shapes appeared only after the hitting event has happened. In the depicted experiment, the triangle was never seen as the grammatical subject during the training phase, and it was never shown combined with a modifier.

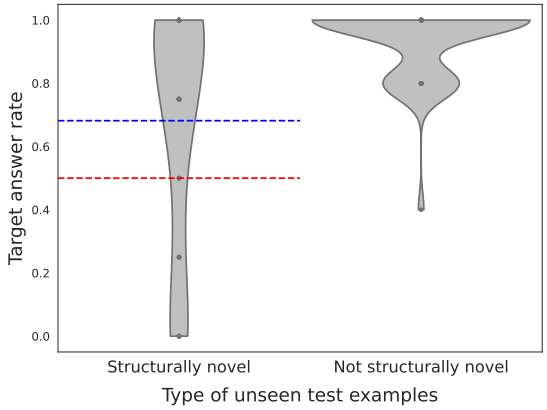

Figure 6: The distribution of target answer rates (# target/# total) in Experiment 2. The blue dotted line indicates the mean target answer rate across all participants, and the red dotted line indicates chance-level production of the target novel structures, assuming a competition between two generalization hypotheses: modification can occur in unobserved positions vs. modification cannot occur in unobserved positions.

breaking) as a result of an event such as hitting. We selected this particular meaning because cracking is a change of state that can plausibly happen to both the hitter and the hittee as a result of a hitting event. The video scenes for the transitive event depicted one shape moving across the screen to hit the other shape, and as a result, one of four things could happen: the hitter cracks, the hittee cracks, neither cracks, or the hitter and the hittee both crack (Figure 5). When the cracking happened, a shattering sound was also played. The shapes did not vary in color in this experiment. Other than this semantic change of the adjective in the artificial language and the change to the videos to reflect this semantic change, the stimuli were kept equivalent to Experiment 1.

Participants

We recruited 122 participants on Prolific.10 participants' data were excluded either due to submission errors or incomplete responses, yielding a total of 112 participants.

Results

The rate of successful learners was 49%, with 55 participants achieving full accuracy on the seen examples during training. The target answer production rate for the unseen but not structurally novel test set was again high, as expected (94%). For the test set with structurally novel targets, the target answer production rate, taking into account minor typos (e.g., *sog* instead of *zog*), was again significantly over chance (68%, $n = 220, z = 5.340, p < .001$, test for one proportion assuming 50% as chance level). The results are visualized in Figure 6. There was a significant difference in the target production rates across two lists ($\chi^2 = 6.675, DF = 1, p < .01, n-1$ χ^2 test), but the target answer production rate within each list was independently over chance (list A: $z = 5.758, p < .001$, list B: $z = 2.156, p < .05$).

We furthermore annotated the generalization patterns of each successful learner. The most frequent pattern was full structural generalization, with 47% (27/55) of the learners either producing the exact target answers for all of the unseen, structurally novel test cases ($n = 24$) or answers structurally identical to the target answers, barring lexical errors (e.g., used a different vocabulary item for a shape; $n = 3$). 38% (21/55) produced at least one target novel structure (`V N-A N` and/or `V N-A N-A`). A smaller group (13%, 7/55) generalized based on the alternative hypothesis that is compatible with the training data: modification cannot appear in other positions than where it was observed. Participants in this group consistently omitted subject modification for all test examples, never producing a novel structure. This analysis is summarized in Figure 7.[5] The rate of full structural generalizers was significantly higher than the rate of the observed structure-only generalizers (47% vs. 13%, $\chi^2 = 15, DF = 1, p < .001$), demonstrating a dominant preference towards licensing rather than withholding generalization of modification to unobserved positions.

Discussion

In both experiments, we observed a significant tendency among the learners of our artificial language to generalize modifiers to unobserved syntactic positions, producing structurally novel answers. This bias towards the elsewhere generalization is observed even when: (1) the exposure to modifi-

[5] One participant categorized as "Other" consistently used `V S O S(-A) O(-A)`, where the subject and object were duplicated whenever modification was present.

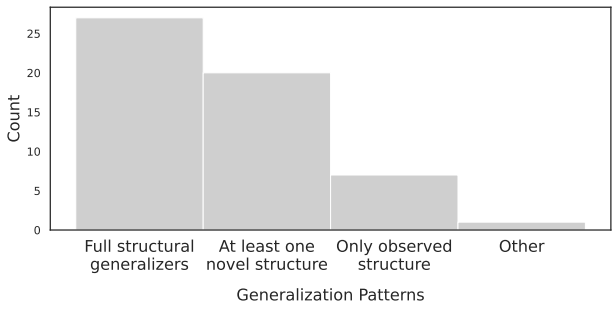

Figure 7: Successful learners (100% accuracy on seen examples during test phase) grouped by their generalization pattern on the structurally novel test set. The most common pattern is full structural generalization (produced target answers for all test cases), followed by producing at least one novel structure. Only a small group of learners consistently generalized based on the alternative hypothesis that modification is only restricted to the observed position.

cation was constrained to a single syntactic position, and (2) the semantics of the artificial language was adversarial for producing the target structures given the learners' knowledge of their native language. These results support the plausibility of the generalization condition originally proposed by K&L.

What is a "fair" generalization test?

Wu et al. (2023) question the fairness of evaluating models on the kind of generalization we tested with very restricted training, arguing that the training data should specify unambiguous generalization targets for the test to be "fair". In their words: *"It is quite reasonable for a learner to infer from this situation that [modifications] are allowed only in this position. [...] Thus, there is a case to be made that this split is not strictly speaking fair in the sense of Geiger et al. (2019): we have a generalization target in mind as analysts, but this target is not uniquely defined by the available data in a way that would invariably lead even an ideal learner to the desired conclusion."*[6] (text in square brackets ours). We argue otherwise: since our human subjects exhibited substantial bias towards the elsewhere generalization even under extremely limited exposure conditions that underdetermine the generalization hypothesis,[7] it is fair to pose this test insofar as we are interested in *human-like* compositional generalization.

Then, when are we interested in human-like compositional generalization as tests for our models? There are two main use cases for compositional generalization tests: (1) as an evaluation for cognitive models, and (2) as an evaluation for AI systems. In (1), it is clear that human generalization patterns should be the modeling target. In (2), under the assumption that how humans interpret certain linguistic inputs is how models should interpret them too, models that can match human generalization in the presence of multiple compatible hypotheses would be a more robust system. This is especially the case in low-resource settings that may not provide the ideal learning condition in which the training data clearly delineates between plausible generalization hypotheses. Furthermore, benchmarks that assume more constrained exposures can incentivize the development of models with stronger inductive biases.

Limitations and future work

Limitations of self-reports for inferring assigned structures: Our experiments included a free-form question at the end, where we asked what the participants thought the newly learned (non-shape) words meant. In Experiment 1, most responded that the nonce word intended as the adjective meant *blue*. The responses for Experiment 2 were trickier to interpret. While some participants did provide answers explicitly indicative of the targeted structure (e.g., *"ro" is an adjective indicating the act of a shape being broken presumably by being "zogged"*), many responded with single words (e.g., *broken*, *break*). However, these responses cannot be taken as indications that the modifier they learned was an exact analog of the English word mentioned. For instance, multiple words were often listed (e.g., *broken or break*, *broke/break/broken*), presumably due to the lack of an exact analog. Then, these seem to be "close enough" English words sharing the inferred semantic features of the learned word, but not necessarily matching in their syntactic categories. The possible mismatch between the English word mentioned and the actual category participants assigned to the learned word is clearly exemplified by this response: *some adjective meaning broke* (*broke* is not an adjective in the context of physical breaking). Therefore, we cannot simply take self-reports, especially analogies to single English words, as direct evidence for the structures the participants assigned to the experimental stimuli.

Different types of modification: The original modification tests used prepositional phrase (PP) modification rather than adjectival modification that we explored in this work. This change in setup was made because using PPs in our experiments led to more non-shape lexical items that the participants needed to learn, as well as compounding the complexity of the visual scene that needed to be described. This increased the difficulty of the training phase, only yielding 10% successful learner rate in our pilots even with all surface forms provided as part of the on-screen lexicon. While we believe that our current results can sufficiently speak to prior discussions about the targeted training-test gap and task fairness, we plan to develop improved protocols in future work to reduce the difficulty of the learning task for PP modification to test the exact generalizations discussed in prior work.

[6] We focus only on the criticism of fairness regarding the *conceptualization* of the generalization task here (i.e., is it fair to assume such a constrained exposure condition that underdetermines the gold target?), and not the additional criticisms raised about the specificities of logical forms affecting generalization performance, which is irrelevant to the current discussion.

[7] Of course, a finite number of observations will never uniquely determine a generalization hypothesis in the absence of constraints about the hypothesis space. Here we used the term underdetermine loosely to point to the fact that our training data does not disambiguate between licensing vs. withholding elsewhere generalization.

Acknowledgments

We thank Geraldine Legendre, Tal Linzen, Aditya Yedetore, Sebastian Schuster, and the anonymous reviewers for helpful discussions. This work was supported by NSF BCS-204122.

References

Bastings, J., Baroni, M., Weston, J., Cho, K., & Kiela, D. (2018, November). Jump to better conclusions: SCAN both left and right. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP* (pp. 47–55). Brussels, Belgium: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W18-5407 doi: 10.18653/v1/W18-5407

Brown, H., Smith, K., Samara, A., & Wonnacott, E. (2022). Semantic cues in language learning: an artificial language study with adult and child learners. *Language, Cognition and Neuroscience*, *37*(4), 509–531.

Culbertson, J., Smolensky, P., & Legendre, G. (2012). Learning biases predict a word order universal. *Cognition*, *122*(3), 306–329.

Drozdov, A., Schärli, N., Akyürek, E., Scales, N., Song, X., Chen, X., ... Zhou, D. (2023). Compositional semantic parsing with large language models. In *The eleventh international conference on learning representations.* Retrieved from https://openreview.net/forum?id=gJW8hSGBys8

Fodor, J. A., & McLaughlin, B. P. (1990). Connectionism and the problem of systematicity: Why Smolensky's solution doesn't work. *Cognition*, *35*(2), 183–204.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1-2), 3–71.

Hadley, R. F. (1994). Systematicity in connectionist language learning. *Mind & Language*, *9*(3), 247–272.

Hong, S.-M. (2006). Why English and Korean resultative constructions differ. *WECOL 2004*, 100–1011.

Kim, N., & Linzen, T. (2020, November). COGS: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 9087–9105). Online: Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.731

Lake, B. M., & Baroni, M. (2018, 10–15 Jul). Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In J. Dy & A. Krause (Eds.), *Proceedings of the 35th international conference on machine learning* (Vol. 80, pp. 2873–2882). PMLR. Retrieved from https://proceedings.mlr.press/v80/lake18a.html

Lake, B. M., Linzen, T., & Baroni, M. (2019). Human few-shot learning of compositional instructions. In *41st annual meeting of the cognitive science society* (pp. 611–617).

Levin, B., & Rappaport Hovav, M. (1995). *Unaccusativity: At the syntax-lexical semantics interface* (Vol. 26). MIT press.

Li, B., Donatelli, L., Koller, A., Linzen, T., Yao, Y., & Kim, N. (2023, December). SLOG: A structural generalization benchmark for semantic parsing. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 3213–3232). Singapore: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.emnlp-main.194 doi: 10.18653/v1/2023.emnlp-main.194

Martin, A., Ratitamkul, T., Abels, K., Adger, D., & Culbertson, J. (2019). Cross-linguistic evidence for cognitive universals in the noun phrase. *Linguistics Vanguard*, *5*(1), 20180072.

McCoy, R. T., Culbertson, J., Smolensky, P., & Legendre, G. (2021). Infinite use of finite means? evaluating the generalization of center embedding learned from an artificial grammar. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 43).

Montague, R. (1970). English as a formal language. In B. Visentini & C. Olivetti (Eds.), *Linguaggi nella societa e nella tecnica* (pp. 188–221). Edizioni di Communita.

Morgan, J. L., Meier, R. P., & Newport, E. L. (1987). Structural packaging in the input to language learning: Contributions of prosodic and morphological marking of phrases to the acquisition of language. *Cognitive psychology*, *19*(4), 498–550.

Morgan, J. L., & Newport, E. L. (1981). The role of constituent structure in the induction of an artificial language. *Journal of verbal learning and verbal behavior*, *20*(1), 67–85.

Partee, B. H. (1984). Compositionality. In F. Landman & F. Veltman (Eds.), *Varieties of formal semantics* (pp. 281–311). Dordrecht: Foris.

Qiu, L., Shaw, P., Pasupat, P., Nowak, P., Linzen, T., Sha, F., & Toutanova, K. (2022, July). Improving compositional generalization with latent structure and data augmentation. In M. Carpuat, M.-C. de Marneffe, & I. V. Meza Ruiz (Eds.), *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 4341–4362). Seattle, United States: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.naacl-main.323 doi: 10.18653/v1/2022.naacl-main.323

Smolensky, P. (1995). Constituent structure and explanation in an integrated connectionist/symbolic cognitive architecture. In C. Macdonald & G. Macdonald (Eds.), *Connectionism: Debates on psychological explanation* (Vol. 2, pp. 221–290). Blackwell.

Wilson, C. (2006). Learning phonology with substantive

bias: An experimental and computational study of velar palatalization. *Cognitive Science*, *30*(5), 945-982.

Wu, Z., Manning, C. D., & Potts, C. (2023, 12). ReCOGS: How Incidental Details of a Logical Form Overshadow an Evaluation of Semantic Interpretation. *Transactions of the Association for Computational Linguistics*, *11*, 1719-1733.