Distributional learning over meaningful words facilitates semantic inferences about previously unknown words

Abigail Laver (alaver@sas.upenn.edu)
Department of Psychology, University of Pennsylvania, 425 S. University Ave, Philadelphia, PA 19104 USA

Heesu Yun (yun.he@northeastern.edu)
Department of Psychology, Northeastern University, 105-107 Forsyth St. #125, Boston, MA 02115 USA

Albert Kim (albert.kim@colorado.edu)
Institute of Cognitive Science and Department of Psychology & Neuroscience, University of Colorado, Boulder, 1905 Colorado Ave., Boulder, CO 80309 USA

John Trueswell (trueswel@psych.upenn.edu)
Department of Psychology, University of Pennsylvania, 425 S. University Ave, Philadelphia, PA 19104 USA

Abstract

Prior research suggests that a small vocabulary of meaningful words (a semantic seed) aids distributional learning. In two experiments, we show that adults who are exposed to a complex artificial language are better at inferring the meaning of previously unknown pseudowords when they were taught a semantic seed prior to distributional exposure. We further show that the benefit of a semantic seed is driven primarily by using seed words to discover the relationship between distributional and semantic classes. These results have implications for how syntactic bootstrapping begins.

Keywords: language acquisition; distributional learning; semantic seed; artificial language learning

Introduction

Understanding how humans learn language, whether a first or a second language, is a central question in cognitive science. One key mechanism underpinning language learning is the extraction of distributional regularities from the input. Infants and young children are adept at detecting statistical patterns in speech, even prior to acquiring word meanings. They segment speech by identifying transitional probabilities between syllables, enabling them to parse word boundaries (Saffran, Aslin, & Newport., 1996; Aslin, Saffran, & Newport, 1998), and infer syntactic structure from limited exposure (e.g., Newport, 2016; 2020). These early distributional learning mechanisms lay the groundwork for acquiring more complex linguistic representations.

At some point, distributional learning interfaces with meaningful words. Some computational models predict that augmenting distributional learning systems with access to meaning (e.g., by augmenting lexical representations with semantic features like concreteness) alters the dynamics of distributional learning to enable the learning of richer grammatical categories (Cartwright & Brent, 1997; Fitz & Chang, 2017; Brusini et al., 2021). Word meanings could potentially scaffold subsequent language learning by highlighting the semantic and syntactic structures of the language ('semantic seeding', see e.g., Babineau, Shi, & Christophe, 2020; Babineau et al., 2021; Barbir et al., 2023; also Gleitman et al., 2005; Christodoulopoulos, Roth, & Fisher, 2016; Fisher, Jin, & Scott, 2020; Mintz, 2003). Because form-meaning mappings emerge as early as six months of age (Bergelson & Swingley, 2012), this scaffolding could shape very early language development.

Some prior work has begun to examine these issues. Ouyang, Boroditsky, and Frank (2017) found that adult English speakers were better at learning grammar and inferring meanings of pseudowords when they occurred within a language that also contained real English words. The English words served an anchor-like function for distributional learning. However, the grammar was simple with only 18 possible sentences. Moreover, inferences about novel word meanings were made based on direct co-occurrence with English words; the study did not test if participants could make semantic similarity inferences by observing similar distributional history across, rather than within, sentences.

Barbir et al. (2023) exposed French 20-month-old infants to novel determiners (ko/ka) that marked animacy (animate vs. inanimate). Following exposure, infants could use these determiners to guess meanings for novel nouns (e.g., nouns marked by ko were likely to be animate). However, these novel function words were in a distributional class infants already had familiarity with. Infants may have been able to apply some existing knowledge of French determiners (e.g., that determiners precede nouns) to these novel exemplars.

In the current study, we examined how a semantic seed shapes distributional learning of a novel grammar in adults, a relevant issue for second language acquisition. Our study builds on prior work by exposing participants to a complex, novel artificial language composed only of pseudowords. We ask (a) whether adults can use the distributional facts of an artificial language to identify syntax-to-semantics relationships in a novel language and (b) if a semantic seed is beneficial or necessary. In Exp. 2, we also investigate (c) the mechanism by which semantic seeds influence distributional learning. In particular, we probe whether learners focus distributional learning on exposure sentences that contain seed words and whether learners use semantic seeds to discover semantic features of distributional categories.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

Experiment 1

We sought to determine how distributional language learning is affected by the presence of a small number of meaningful words in the lexicon before distributional exposure begins and also how such distributional learning would affect future learning about the meanings of words in the language.

Methods

Participants Forty-three undergraduates from the University of Pennsylvania participated for course credit. Participants were required to be native English speakers, but could be multilingual. One participant was excluded for not being a native speaker. Two participants were excluded due to smartphone use during the study.

Grammar The grammar had three sentence types: AB, AAB, and AAAB. Unbeknownst to participants, the language was designed to mimic a Subject-Object-Verb (SOV) language with A mimicking nouns and B verbs. The AB, AAB, and AAAB structures corresponded to intransitive, transitive, and ditransitive sentences, respectively (Fig. 1).

The A-Class had two subcategories, A_1 and A_2, which were distributed in a probabilistic pattern as shown in Fig. 1. This pattern was designed to mimic the distributional properties of animate and inanimate nouns, respectively. Specifically, in SOV languages, grammatical Subjects and Indirect Objects tend to be Animate and Direct Objects tend to be Inanimate, such that animate nouns typically precede inanimate nouns.

The B-Class, designed to mimic verbs, was subdivided into B_1, B_2 and B_3. B_1 words only occurred in the AB structure, B_2 words only in the AAB structure, and B_3 words only in the AAAB structure. Thus, B_1 words mimicked the distributional properties of verbs with one-participant meanings (e.g., *jump*), B_2 two-participant meanings (e.g., *lift*), and B_3 three-participant meanings (e.g., *give*).

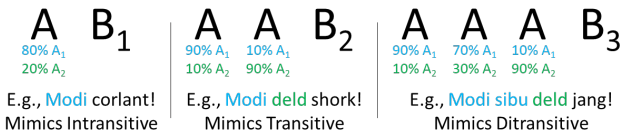

Figure 1: Grammar. Designed to mimic SOV language. Three sentence types: Intransitive (AB_1), Transitive (AAB_2) and Ditransitive ($AAAB_3$). A_1 and A_2 distributions designed to mimic animate and inanimate noun distributions.

Vocabulary The vocabulary consisted of 35 pseudowords (18 monosyllabic, 17 disyllabic) with English-like spelling and pronunciation, randomly assigned to distributional categories such that twenty-three were A-Class (13 A_1; 10 A_2) and 12 were B-Class (3 B_1; 6 B_2; 3 B_3).

Exposure Sentences 300 pseudoword sentences (75 AB, 150 AAB, 75 AAAB) were generated from the probabilistic grammar. Vocabulary words were randomly selected and placed into sentences with the constraint that all words appeared with approximately equal frequency.

Procedure All participants experienced three phases: (1) Seed Phase, (2) Distributional Exposure, and (3) Cross-Situational Word Learning (CSWL).

During the **Seed Phase**, participants were randomly assigned to the Semantic Seed or Non-Semantic Seed condition. Semantic Seed participants were taught the meanings of three A_1, three A_2, and three B pseudowords from the language. The meanings belonged to semantic categories that corresponded to distributional categories: A_1 words had animate meanings (*cat*, *doctor*, *mom*), A_2 words inanimate meanings (*apple*, *book*, *house*), and B words event meanings (*jump*, *lift*, *give*, assigned to B_1, B_2, and B_3, respectively). During teaching, participants heard and also read a definition ("Modi means doctor."). This was followed by three repetitions of hearing and reading the pseudoword along with its printed English translation and three example images depicting its meaning (e.g., clipart images of doctors). Following teaching of all 9 pseudowords, participants were tested on the meanings in three test trials per pseudoword. On each test trial, participants heard and read printed on-screen questions like, "What does modi mean?" and selected its meaning from among four images. Here, and in other phases of the study, auditory samples of the language were generated by the Amazon Polly English text-to-speech tool.

Participants in the Non-Semantic Seed condition were familiarized with the same pseudowords as Seed participants, but were not taught meanings. Familiarization trials consisted of hearing and reading a sentence introducing the pseudoword ("One of the words is modi."), followed by three repetitions of hearing and reading the pseudoword. Participants were tested three times on each pseudoword with a four-way forced choice among printed pseudowords in response to spoken and printed on-screen questions like, "Which one is modi?".

Figure 2: Structure of dialogues in Exp. 1 & 2. Sentences were not numbered for participants.

Next, all participants completed the **Distributional Exposure Phase** where they encountered the 300 sentences described above, which contained both Seed and unfamiliar pseudowords. Sentences appeared in four-sentence dialogues between animated cartoon characters. All dialogues had the structure depicted in Fig. 2 where character 1 produces the

first sentence, character 2 repeats it, character 1 produces another sentence, and character 2 repeats it. One pseudoword always appeared in all four sentences. This resulted in 150 dialogues (presented in a fixed random order). Periodic attention checks (every 24 trials) asked participants to listen to a sentence and indicate whether it had appeared in the preceding dialogue. At two points, participants were reminded of forms and meanings of the nine Seed pseudowords (Semantic Seed condition) or just their forms (Non-Semantic Seed condition).

Finally, participants completed the **Cross Situational Word Learning (CSWL) Phase**, in which participants learned the meanings of 12 pseudowords that appeared in the Distributional Exposure Phase but were not Seeds. On each trial, participants heard and saw the prompt "What does [pseudoword] mean?", while viewing three clipart images (Fig. 3). Images were labeled in English. One image came from each semantic category (animate, inanimate, and event). Participants clicked one image to select the meaning of the presented pseudoword. Each of 12 pseudowords occurred once in each of six blocks (total of 72 trials). On each occurrence of a pseudoword, the same target image was always present, accompanied by two distractor images that were always different across trials. The consistent co-occurrence between each word and one meaning across different trials enabled learning of the word-referent mappings, over the course of trials.

The 12 pseudowords were drawn from each of the distributional classes A_1, A_2, and B (four each). Target meanings came from three semantic categories (animate: *dad*, *dog*, *firefighter*, *teacher*; inanimate: *chair*, *pear*, *pencil*, *shoe*; event: *sneezing*, *kicking*, *pushing*, *sending*, assigned to B_1, B_2, B_2, and B_3, respectively).

Block 1 Block 3 Block 6

Figure 3: Three of six example CSWL trials for "regli." A consistent referent (*dog*) appeared on all six trials. Participants received no feedback on responses.

Control Participants and Procedure An additional 45 participants from the same pool served as Controls. Two were excluded for learning English after the age of five years and three were excluded for technical reasons. Like Test participants, half of the Controls were assigned to the Semantic Seed condition and half to the Non-Semantic Seed condition. However, during the CSWL phase, Controls attempted to learn word meanings that did NOT align with the distributional categories (A_1, A_2, and B). For example, the four pseudowords paired with animate referents came from all three distributional classes, rather than just A_1. Thus, the distributional category had no bearing on the referential consistencies encountered in the CSWL phase.

Predictions We predicted that participants would learn pseudowords' meanings, with accuracy increasing over CSWL blocks (1-6) given that the word-referent mappings were consistently presented.

The primary question addressed by the CSWL task was whether participants' learning, and possibly even their initial hypotheses, about word meanings would be affected by the words' distributional histories during the preceding Exposure Phase. In particular, we asked whether participants would expect words with similar distributional histories to have similar meanings (A_1 = animates; A_2 = inanimates; B = events), based on knowledge acquired in the Distributional Exposure Phase. If so, Test participants should outperform Control participants, because only the former learned word-referent mappings that aligned with the words' distributional histories.

Notably, the advantage for Test participants may be enhanced for, or even limited to, participants in the Semantic Seed condition. The meaningful Seed words may augment distributional category learning by serving as exemplars for semantic categories that correspond to the distributional categories (e.g., if *modi=doctor* and is in category A_1, then perhaps other A_1 words are animate).

Finally, if participants learn *semantic* generalizations about pseudowords during the Distributional Exposure Phase, then this might bias their hypotheses about meanings in CSWL Block 1. We predicted that Test participants in the Semantic Seed condition will be above chance on CSWL Block 1.

Analysis We analyzed accuracy in the CSWL task using logistic mixed effect models on trial-level data with CSWL Block (1-6), Seed (Semantic Seed vs Non-Semantic Seed), and Test status (Test vs Control) as fixed effects and participant and item (pseudoword) as random effects. We also analyzed the trial-level data from CSWL Block 1, using a logistic mixed effect model with Seed and Test status as fixed effects and participant and item (pseudoword) as random effects. We also ran separate t-tests on CSWL Block 1 for each condition to test if participants' initial semantic biases aligned with the grammar.

Results & Discussion

Fig. 4 presents average accuracy of selecting the target meaning for each test pseudoword as a function of CSWL Block. Accuracy increased across blocks for participants in all conditions, consistent with learning. This was reflected in an effect of Exposure Block ($\lambda^2(1) = 27.09$, $p < .001$; $\beta = 1.04$, $z = 8.36$, $p < .001$).

Performance was higher in the Semantic Seed than the Non-Semantic Seed condition, reflected in an effect of Seed condition ($\lambda^2(1) = 4.50$, $p = .034$; $\beta = 0.24$, $z = 2.15$, $p = .031$). The benefit of Semantic Seeding occurred only for Test participants and not Control participants, reflected in an interaction between Seed condition and Test status ($\lambda^2(1) =$

7.69, p = .005; β = 0.29, z = 2.84, p = .005). There was no main effect of Test status ($\lambda^2(1)$ = 2.41, p = .120; β = 0.20, z = 1.57, p = .116).

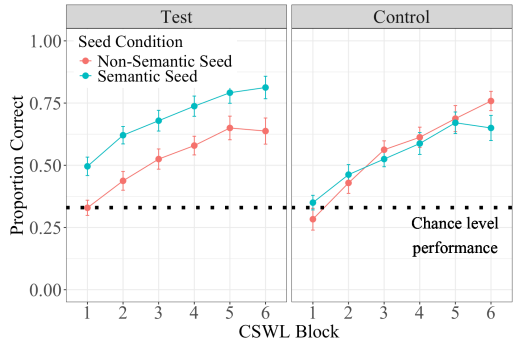

Figure 4: Exp. 1 CSWL results. Error bars represent + or -1 SE. Dashed line represents chance (33%).

We analyzed CSWL Block 1 to test for effects of prior distributional learning on initial hypotheses. Seed condition and Test status did not interact. However, performance was higher for Semantic than Non-Semantic Seed participants ($\lambda^2(1) = 8.70$, $p = .003$; $\beta = 0.30$, $z = 3.30$, $p < .001$) and higher for Test than Control participants ($\lambda^2(1) = 5.72$, p = .017; β = 0.24, z = 2.71, p = .007), suggesting an advantage for Test participants in the Semantic Seed condition.

One-tailed t-tests assessed whether participants were above chance (33%) on CSWL Block 1. Only Test participants in the Semantic Seed condition were above chance (M = 0.50, SD = 0.50; $t(239) = 5.02$, $p < .001$); participants in the other three conditions were not above chance (all p's > .300).[1]

Conclusions We found that participants could use the distributional history of otherwise meaningless pseudowords to inform hypotheses about word-referent mappings during CSWL. This effect was limited to Semantically Seeded Test participants, who had learned the meaning of a small set of pseudowords (which were not part of the CSWL test) prior to Distributional Exposure. This effect of Seeding on Test participants was present on CSWL Block 1, suggesting that they discovered the relationship between the distributional and semantic classes during Distributional Exposure, before the CSWL phase began.

Experiment 2

Experiment 2 investigates the learning mechanism behind the semantic seed effect—how knowledge about a small set of word meanings can affect hypotheses about the meanings of unknown words, through distributional learning. Our proposed explanation is that knowing a few words in each distributional class leads to semantic generalization (e.g., if some A_1 words are animate, others likely are too). However, an alternative mechanism is that during Distributional Exposure, encountering a sentence with a meaningful word (a Semantic Seed) may heighten learning as participants infer the sentence's meaning (e.g., if modi means doctor, what might the sentence "modi dax mipen" mean and what might *dax* and *mipen* mean?). This could create semantic expectations leveraged in the CSWL task, in principle without requiring generalizations about distributional class.

Experiment 2 tests these possibilities by comparing two types of pseudowords. Half appeared in sentences with Seed words during Distributional Exposure (Seed-Co-Occurring Test words), while the rest appeared only in sentences without them (Non-Seed-Co-Occurring Test words). Both types were assigned to distributional classes and matched in frequency and distributional evidence. If learning occurs via semantic generalization, both should inherit class meaning and show similar benefits. However, if co-occurrence with meaningful words drives learning, Seed-Co-Occurring words should show stronger semantic expectations.

Since seeding effects appeared in Block 1 of the CSWL task in Experiment 1, we administered only Block 1 here, now called the Referent Selection task.

Finally, we added a grammar test, where participants rated the familiarity of novel grammatical and ungrammatical sentences. Greater familiarity for grammatical sentences would indicate learning of distributional constraints. We examined whether grammar learning from distributional evidence is enhanced if one already knows the meaning of a small set of words, by comparing grammar test performance on a Semantic Seeded language to a Non-Semantic Seeded language. We also examined whether grammar learning is disproportionately influenced by sentences that contain Seeds, by testing whether test sentences composed of Seed-Co-Occurring Test words show stronger grammaticality effects than those composed of Non-Seed-Co-Occurring Test words, despite their matched distributional properties.

Methods

Participants Sixty-seven undergraduates from the University of Pennsylvania subject pool participated for course credit. Participants were required to be native English speakers, but could be multilingual. Three participants were excluded for technical reasons.

Grammar The grammar was the same as in Exp. 2.

Vocabulary The vocabulary extended Exp. 1's vocabulary by 13 pseudowords (6 monosyllabic, 7 disyllabic), totaling 48. The new pseudowords were randomly placed into distributional categories; nine were A-Class (5 A_1; 4 A_2) and four were B-Class (1 B_1; 2 B_2; 1 B_3).

Twelve pseudowords were chosen as Seed words (see below). Another twelve were chosen as Non-Seed words,

[1] One might expect that performance of Controls would be below chance since the target meanings of most pseudowords (9 out of 12) mismatched their Distributional history. A separate analysis of these 9 words found that Control participants in the Semantic Seed condition were not below chance, while Control participants in the Non-Semantic Seed condition were below chance.

which were neither pre-trained during the Seed Phase nor tested at the end of the study. Another nine pseudowords were selected as Seed-Co-Occurring Test words. Nine additional words were selected as Non-Seed-Co-Occurring Test words. The remaining six words were neither pre-trained nor tested.

Exposure Sentences 160 pseudoword sentences (40 AB, 80 AAB, 40 AAAB) were generated based on the probabilistic grammar. Vocabulary words were randomly selected and placed into sentences with the first constraint being that each sentence contained at least one Seed word and no Non-Seed or Non-Seed-Co-Occurring words. The second constraint was that Seed and Seed-Co-Occurring words appeared equally often. These were the Seed-Co-Occurring sentences. We substituted each Seed and Seed-Co-Occurring word for a corresponding Non-Seed and Non-Seed-Co-Occurring word, respectively, to generate an additional 160 sentences. These were the Non-Seed-Co-Occurring sentences.

Procedure The experiment had four phases: (1) Seed Phase, (2) Distributional Exposure, (3) Referent Selection Phase, and (4) Grammar Test.

The **Seed Phase** and **Distributional Exposure Phases** were the same as in Exp. 1, except that we included one additional Seed pseudoword per distributional class, and used the 320 exposure sentences described above.

Next participants completed the **Referent Selection Phase**, which was a single block of the CSWL task from Exp 1. We added six additional pseudowords for a total of 18 (nine Seed-Co-Occurring and nine Non-Seed-Co-Occurring). Distributional classes A_1, A_2, and B were each represented by six pseudowords. The six additional meanings were (animate: *girl*, *monkey*; inanimate: *banana*, *table*; event: *crying*, *throwing*, assigned to B_1, and B_3, respectively). There were twelve distractor meanings per semantic class.

Finally, in a **Grammar Test,** participants rated the familiarity of 48 sentences on a 5-point scale (1=*Definitely NOT familiar*; 5=*Definitely FAMILIAR*). Half the sentences violated the language's category order rules by placing a B-class word in a non-final position (Ungrammatical). All other sentences were Grammatical. In both Grammatical and Ungrammatical sentences, the B word was of the correct subtype (B_1 in AB, B_2 in AAB, and B_3 in AAAB). AB, AAB, and AAAB subtypes were equally frequent. 24 sentences consisted solely of Seed-Co-Occurring words, and 24 consisted solely of Non-Seed-Co-Occurring words.

Predictions We predicted that, as in Experiment 1, Semantic Seed participants should perform above chance in the Referent Selection task, while Non-Semantic Seed participants should be at chance. If learners use semantic seeds to interpret sentences and constrain the meanings of co-occurring but unknown words within those sentences, then the benefit of seed words should be greater for Seed-Co-Occurring words than for Non-Seed-Co-Occurring words. Alternatively, if learners use semantic seeds to discover semantic features of distributional categories, then the benefit of a semantic seed should extend even to Non-Seed-Co-Occurring words; for Semantic Seed participants, Referent Selection accuracy should be above chance for both Seed- and Non-Seed-Co-Occurring pseudowords.

We predicted that participants would rate Grammatical sentences as more familiar than Ungrammatical ones in the Grammar Test. We also predicted that discrimination of grammatical and ungrammatical sentences should be greater for Semantic Seed participants than Non-Semantic Seed participants, because seeds would aid grammar learning. If direct co-occurrence with seed words aids grammar learning, then we should observe enhanced sensitivity to grammaticality for sentences consisting of Seed Co-Occurring words instead of Non-Seed-Co-occurring words.

We also expected that sensitivity to grammaticality would be predictive of performance in the Referent Selection task for Semantic Seed, but not Non-Semantic Seed, participants.

Analysis We analyzed the data from the Referent Selection Phase using logistic mixed effect models on trial-level data with Seed (Semantic Seed vs Non-Semantic Seed), and Seed Co-Occurrence (Seed- vs Non-Seed-Co-Occurring) as fixed effects and participant and item (pseudoword) as random effects. We ran separate t-tests for each condition to test if participants' semantic biases aligned with the grammar.

We analyzed the data from the Grammar Test using cumulative link mixed effects models on trial-level data with Seed, Seed Co-Occurrence, and Grammaticality (Grammatical vs Ungrammatical) as fixed effects and participant and item (sentence) as random effects.

Results & Discussion

Referent Selection Fig. 5 presents accuracy in the Referent Selection task. Semantic Seed participants were more accurate than Non-Semantic Seed participants ($\lambda^2(1) = 9.54$, $p = .002$; $\beta = 0.24$, $z = 3.29$, $p = .001$), replicating Exp. 1. Seed Co-Occurrence did not affect accuracy. Semantic Seed participants were above chance for Seed-Co-Occurring words ($M = 0.49$, $SD = 0.50$, $t(287) = 5.41$, $p < .001$) and Non-Seed-Co-Occurring words ($M = 0.44$, $SD = 0.50$, $t(287) = 5.41$, $p < .001$). Non-Semantic Seed participants were not above chance.

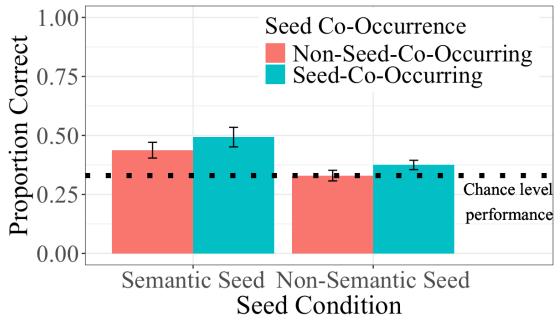

Figure 5: Exp. 2 Referent Selection results. Error bars represent + or -1 SE. Dashed line represents chance (33%).

These results suggest that the benefit of Seed words is primarily the result of learning the relationship between the distributional and semantic classes through the Seed words and generalizing this to words with similar distributional histories. We did not find evidence that pseudowords gained more semantic knowledge during the Exposure Phase when they occurred in the same sentence as Seed words.

Grammar Test Novel grammatical sentences were rated as more familiar than novel ungrammatical sentences ($M_{gram.}$ = 3.45; $M_{ungram.}$ = 2.60; $\lambda^2(1) = 39.26$, $p < .001$; $\beta = 0.76$, $z = 7.34$, $p < .001$), indicating that participants learned the distributional patterns of the language. There were no other effects or interactions (all p's > .140), indicating that participants learned the rules of the grammar regardless of Seed condition and Seed-Co-Occurrence condition and regardless of whether sentences were comprised of Seed-Co-Occurring or Non-Seed-Co-Occurring test words (see Fig. 6).

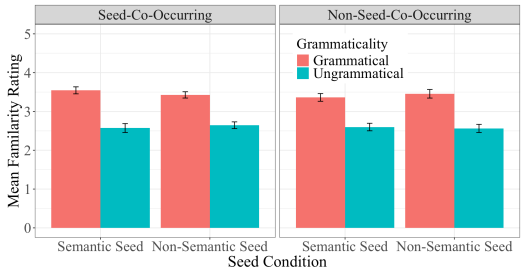

Figure 6: Average rating of Grammar Test sentences. 5 = Definitely Familiar; 1 = Definitely NOT familiar. Error bars represent + or -1 SE.

Participants who better learned the distributional properties of the language trend toward showing greater effects of Semantic Seeding in the Referential Task. For each participant we computed a Grammar Rating difference score (Grammatical - Ungrammatical) and used it to predict Referent Selection performance, in separate models of those in the Semantic Seed and Non-Semantic Seed conditions. In the Semantic Seed condition, the effect of Grammar Test difference score was positive, though not significant ($\lambda^2(1) = 3.60$, $p = .068$; $\beta = 0.13$, $t = 1.90$, $p = .068$), while it had no relation in the Non-Semantic Seed Condition ($\lambda^2(1) = 0.17$, $p = .685$; $\beta = 0.02$, $t = 0.41$, $p = .685$).

Conclusions The evidence from the Referent Selection Phase replicates Exp. 1. Participants used the distributional history of otherwise meaningless pseudowords to make inferences about their meaning, but only when they had previously learned the meaning of a small set of pseudowords (which were not part of the Referent Selection Phase). Moreover, the results suggest that the benefit of a Semantic Seed is primarily the result of using Seed words to discover the relationship between the distributional and semantic classes.

The Grammar Test results suggest that all participants were able to learn the distributional patterns of the language. Having a semantic seed did not aid grammar learning, nor was grammar learning improved for Test pseudowords that appeared in the same sentences as Seed words.

Sensitivity to grammaticality in the Grammar Test trends towards being predictive of performance in the Referent Selection Phase for Semantic Seed participants only. While further work is needed to verify this relationship, it is consistent with the idea that learning the relationship between the semantic and distributional classes is driving the Semantic Seed effect.

General Discussion

In two experiments, we showed that adults can use the distributional history of previously unknown words to form hypotheses about their meanings, but only when they already have a small vocabulary of meaningful words (a semantic seed). These biases are present immediately following distributional exposure before additional evidence about the meanings of the unknown words is provided.

Experiment 2 provided insight into the learning mechanism underlying this effect. Seeding the language with a small set of meaningful words led participants to generate semantic expectations for other words in the same distributional class, even for those words that had never directly co-occurred with seed words (i.e., those that appeared only in non-seeded sentences during exposure). This suggests that introducing a few meaningful words into a distributional class prompts learners to generalize meaning across all words in that class. It further suggests that these semantic biases were not a result of learners selectively interpreting only the words that appeared in potentially meaningful exposure sentences (i.e., the seeded sentences).

The results of our grammar test in Exp. 2 provide further evidence that learners were not focusing their learning exclusively to exposure sentences that contained seed words. Participants were sensitive to grammatical violations of test sentences regardless of whether these sentences were composed of words that had appeared only in semantically seeded exposure sentences or were composed of words that had appeared only in non-seeded exposure sentences. This indicates that learners acquired the distributional properties of all words in the language, not just those that co-occurred with seed words.

In future studies, we will modify the methods used here to investigate how semantic seeds shape distributional learning in children. This will extend prior work with children (e.g., Barbir et al., 2023) by examining this question in the context of a fully novel grammatical system that differs from the grammar of the children's native language.

This work has important implications for syntactic bootstrapping theories, which state that language learners can use the syntactic contexts in which words appear to learn their meanings (Gleitman, 1990; Landau & Gleitman, 1985). Learning a few words from word-to-world mapping (e.g., Gleitman & Trueswell, 2020) may provide language learners with a semantic seed that they can use to learn the relationship between distributional and semantic classes, allowing them to begin syntactic bootstrapping.

References

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*(4), 321-324.

Babineau, M., De Carvalho, A., Trueswell, J., & Christophe, A. (2021). Familiar words can serve as a semantic seed for syntactic bootstrapping. *Developmental Science*, *24*(1), e13010.

Babineau, M., Shi, R., & Christophe, A. (2020). 14-month-olds exploit verbs' syntactic contexts to build expectations about novel words. *Infancy*, *25*(5), 719-733.

Barbir, M., Babineau, M. J., Fiévet, A.-C., & Christophe, A. (2023). Rapid infant learning of syntactic–semantic links. *Proceedings of the National Academy of Sciences*, *120*(1), e2209153119.

Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, *109*(9), 3253–3258.

Brusini, P., Seminck, O., Amsili, P., & Christophe, A. (2021). The acquisition of noun and verb categories by bootstrapping from a few known words: A computational model. *Frontiers in Psychology*, *12*, 661479.

Cartwright, T. A., & Brent, M. R. (1997). Syntactic categorization in early language acquisition: Formalizing the role of distributional analysis. *Cognition*, *63*(2), 121-170.

Christodoulopoulos, C., Roth, D., & Fisher, C. (2016). An incremental model of syntactic bootstrapping. *Proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning*, 38–43.

Fisher, C., Jin, K., & Scott, R. M. (2020). The developmental origins of syntactic bootstrapping. *Topics in Cognitive Science*, *12*(1), 48–77.

Fitz, H., & Chang, F. (2017). Meaningful questions: The acquisition of auxiliary inversion in a connectionist model of sentence production. *Cognition*, *166*, 225-250.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, *1*(1), 3-55.

Gleitman, L. R., Cassidy, K., Nappa, R., Papafragou, A., & Trueswell, J. C. (2005). Hard words. *Language Learning and Development*, *1*(1), 23-64.

Gleitman, L. R., & Trueswell, J. C. (2020). Easy words: Reference resolution in a malevolent referent world. *Topics in Cognitive Science*, *12*(1), 22–47.

Landau, B. & Gleitman, L.R. (1985). *Language and experience: Evidence from the blind child*. Cambridge, MA: Harvard University Press.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*(1), 91–117.

Newport, E. L. (2016). Statistical language learning: Computational, maturational, and linguistic constraints. *Language and Cognition*, *8*(3), 447–461.

Newport, E. L. (2020). Children and adults as language learners: Rules, variation, and maturational change. *Topics in Cognitive Science*, *12*(1), 153–169.

Ouyang, L., Boroditsky, L., & Frank, M. C. (2017). Semantic coherence facilitates distributional learning. *Cognitive Science*, *41*(S4), 855–884.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.