Semantic coherence facilitates distributional learning of lexical categories

Long Ouyang, Lera Boroditsky, Michael C. Frank

Department of Psychology, Stanford University

Word count: 6,251

Author Note

An earlier version of this paper appeared in the Proceedings of the 34th Annual Meeting of the Cognitive Science Society. Address correspondence to:

Long Ouyang

Jordan Hall, Building 01-420

450 Serra Mall

Stanford, CA, 94305

Email: `longouyang@post.harvard.edu`

Abstract

Computational models have shown that purely statistical knowledge about words' linguistic contexts is sufficient to learn many properties of words, including grammatical category and meaning. For example, a learner might infer that "postman" and "mailman" belong to the same category because they have quantitatively similar patterns of association with *other* words (e.g., both "postman" and "mailman" tend to occur with words like "deliver", "truck", "package"). Are people able to leverage distributional statistics in this way in the service of learning lexical categories? Results from prior artificial language learning experiments suggest that the answer may be no. However, experiments in this paradigm expose participants to entirely novel words, whereas real language learners encounter input that contains some known words that are semantically organized. In three experiments, we show that (1) the presence of familiar semantic reference points facilitates distributional learning of categories and (2) this effect crucially depends both on the presence of known words and the adherence of these known words to some semantic organization.

Keywords: distributional learning; word learning; semantic coherence

Semantic coherence facilitates distributional learning of lexical categories

"*You shall know a word by the company it keeps.*" Firth (1957, p.11)

How do people learn language? Research indicates that learners use many information sources, including physical, social, conceptual, and linguistic cues (Baldwin, 1993; E. V. Clark, 1988; Gleitman, 1990; Hollich, Hirsh-Pasek, & Golinkoff, 2000; Markman, 1991). One abundant source is distributional information about language itself—people can use statistical knowledge about language input during learning. For example, learners can group sounds together into word forms based on their statistical co-occurrence (Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996) and pair word forms with their referents based on consistent associations (L. Smith & Yu, 2008; Yu & Ballard, 2007). In the current paper, we explore a more sophisticated type of distributional learning: using evidence about the linguistic contexts that a word occurs in as a cue toward categorization (Braine, 1987; Maratsos & Chalkley, 1980; Redington, Chater, & Finch, 1998; K. H. Smith, 1966).

As an example of distributional learning, one might infer that "postman" and "mailman" belong to the same category based solely on the fact that they both tend to occur with words like "deliver", "package", and "truck." To give an intuition in a different domain, we might judge whether two people are similar based on their patterns of association with other people. If we know that Alice associates with accountants and lawyers and that Bob associates with professors and college sophomores, we might judge that they are dissimilar. For words, this kind of learning is driven not by *direct* co-occurrence between two words of interest but rather by the similarity in their linguistic contexts—their patterns of co-occurrence with *other* words. As Firth (1957) put it, one learns a word "by the company it keeps" (p.11). We will refer to learning from such learning as *distributional learning*.

Distributional learning is a general mechanism thought to be important throughout language learning, including acquisition of word meaning (Landauer & Dumais, 1997) and grammatical category (Redington et al., 1998). Nevertheless, a number of experiments, conducted mainly on acquisition of grammatical category, suggest that human learners' capacities are limited in this

regard (Braine, 1987; Brooks, Braine, Catalano, Brody, & Sudhalter, 1993; Frank & Gibson, 2011; Frigo & McDonald, 1998; Gerken, Wilson, & Lewis, 2005; Kempe & Brooks, 2001).

Our study addresses this mismatch by specifying conditions under which people succeed in distributional learning. In particular, we investigate the effect of existing linguistic knowledge. Prior experiments typically exposed learners to artificial languages composed of entirely novel words. Our experiments suggest that *semantic coherence*, the presence of known words adhering to some semantic organization, facilitates distributional learning. Put another way, a small semantic hook allows people to better leverage distributional information for learning lexical categories. Phrased in terms of the Alice–Bob example, we stand a better chance of learning about Alice from her associates if those associates play roles that are already meaningful to us. Knowing that Alice associates with professors and college sophomores gives us a clearer picture of her than, say, knowing that she associates with taphonomists and bryologists.[1] Learning a word from its company is easier if this company is already familiar.

We begin by briefly reviewing some of the computational evidence for distributional learning. We then introduce the specific language structure, MNPQ, that we explore in this paper. Next, we discuss research that uses the MNPQ structure to study distributional learning of grammatical categories. In Experiment 1, we present evidence that semantic coherence can facilitate distributional learning of word categories and that learners make use of acquired categories when performing an online referent assignment task. In Experiments 2 and 3, we further explore this effect by isolating two components of semantic coherence—familiarity of words and presence of an overarching semantic organization —and find that neither alone facilitates learning. We conclude by discussing the limitations of purely artificial language learning, possibilities for future empirical and computational work, and potential mechanisms for the effects we observe.

---

[1]Though, we can infer much more about Alice if we know that taphonomy is the study of decay and fossilization and that bryology is the study of certain plants called bryophytes, which include mosses, hornworts, and liverworts.

Computational evidence for distributional learning

Initial proposals about distributional learning came from philosophical and linguistic research on the nature of meaning. In the philosophical literature, Wittgenstein (1953/1997) objected to the idea that words have precise, formal definitions and argued that meaning derives from patterns of usage. To illustrate, he considered the word "game" – a term used to describe activies as varied as chess, poker, basketball, the Olympics and so forth – and observed that these activities lack a shared essence that we could distill to a definition. To use a spatial metaphor, the set of things called games does not appear to be a contiguous region that we could draw a boundary for, but rather a "complicated network of similarities overlapping and crisscrossing" (§66). Wittgenstein argued that understanding meaning requires describing usage.

In linguistics, Firth (1957) similarly argued for a theory of meaning based on patterns of "habitual collocation." For example (p.12), he asserts that part of the meaning of the "cow" is its co-occurrence with "milk", as in "They are milking the cows" or "Cows give milk." "Tigress" and "lioness" do not co-occur with "milk" as often and thus must differ somewhat in meaning. Firth stressed the utility of *pure* co-occurrence independent of extralinguistic or even grammatical aspects. He even outlined a prescient kind of cluster analysis quite similar to modern-day statistical approaches:

> "In the study of selected words, compounds and phrases in a restricted language for which there are restricted texts, an exhaustive collection of collocation will suggest a small number of groups of collocations for each word studied. The next step is the choice of definitions for meanings suggested by the groups" (p. 13)

Firth's contemporary Harris advanced a quantitative version of this notion called the *distributional hypothesis*, which proposes that words are semantically similar to the degree that they participate in the same contexts (Harris, 1951). Harris argued that, even in cases where word meaning was determined by extralinguistic influences, such influences would have distributional correlates. Thus, meaning could be divined by the quantitative analysis of purely linguistic information.
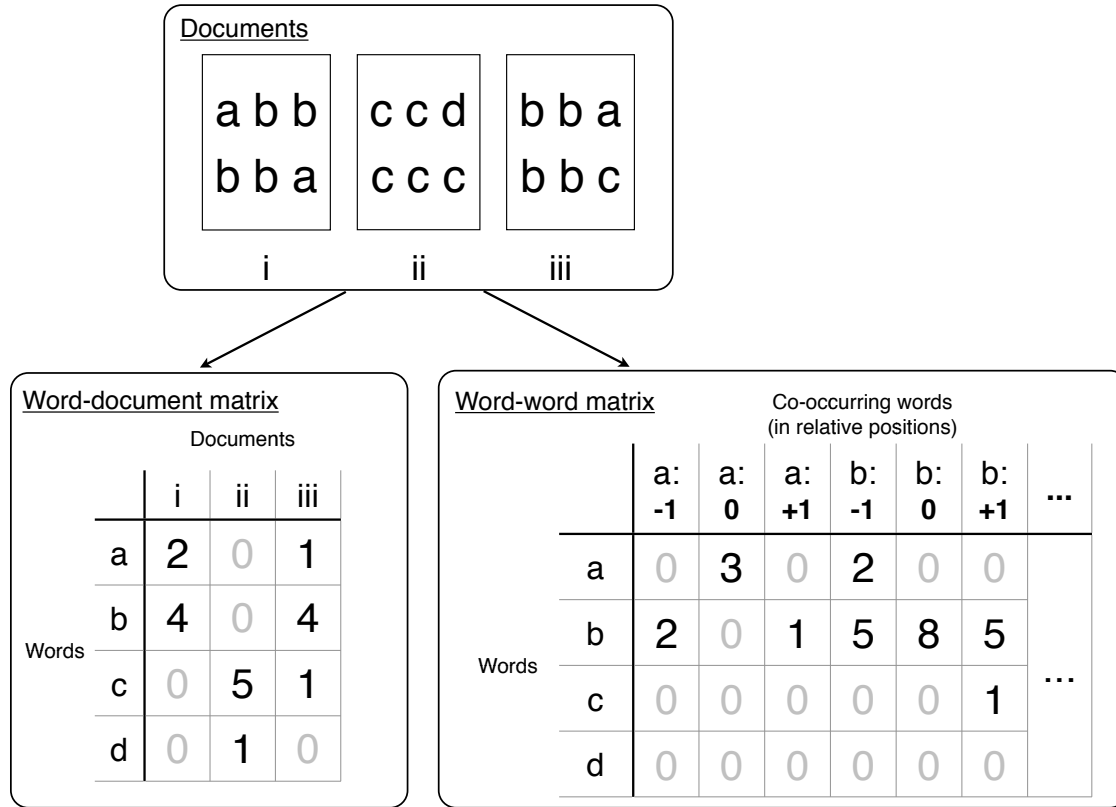
Documents

	a b b	c c d	b b a
	b b a	c c c	b b c
	i	ii	iii

Word-document matrix

Documents

Words		i	ii	iii
	a	2	0	1
	b	4	0	4
	c	0	5	1
	d	0	1	0

Word-word matrix Co-occurring words (in relative positions)

Words		a: -1	a: 0	a: +1	b: -1	b: 0	b: +1	...
	a	0	3	0	2	0	0	
	b	2	0	1	5	8	5	...
	c	0	0	0	0	0	1	
	d	0	0	0	0	0	0	

Figure 1. Word-document and word-word matrices. Word-document matrices measure how often words occur in particular documents. Word-word matrices measure how often pairs of words occur in certain relative positions.

These proposals about the distributional theory of meaning stimulated early empirical work with humans, which typically supported the distributional hypothesis using small samples of human judgments and corpora (Berry-Rogghe, 1973; H. Clark, 1968; Geffroy, Lafon, Seidel, & Tournier, 1973; Rubenstein & Goodenough, 1965; Stefflre & Reich, 1971; Szalay & Bryson, 1974). For example, Rubenstein and Goodenough (1965) compared synonym judgments for pairs of concrete nouns generated by one group of subjects with co-occurrence statistics in a corpus generated by a separate group of subjects. They found a positive relationship between synonymy and degree of linguistic context match.

In the 1980s, computer scientists devised techniques that paved the way for large scale investigations of distributional learning. Motivated by practical issues in the field of information

retrieval, they considered the relationships between words and documents. A typical problem was to retrieve documents relevant to a user query with certain search terms (i.e., a search engine). One solution to this problem is to represent documents as points in a high dimensional space whose dimensions are frequencies for different words (Salton & McGill, 1983). This approach lends itself quite naturally to a matrix representation with words labeling rows, documents labeling columns, and cells encoding how often a particular word occurs in a particular document (Figure 1). While we can interpret such matrices as representing documents in terms of their component words, we can also interpret them representing words in terms of their frequency of use across different documents (a useful form for distributional learning). To make an analogy with our Alice–Bob example, such *word-document* matrices license inferences about Alice and Bob based on *where* they appear (rather than *who* they appear with).

Note, however, that this representation discards information about the relative position of words within documents. The *word-word* approach retains this information (Church & Hanks, 1990; Schütze, 1992), labeling both rows and columns with words; a row might represent the meaning of word A and the count in a particular column might indicate frequency that word A occurred 2 words before word X (Figure 1). This corresponds to the initial example of inferences about Alice and Bob based on *who* they appear with.

Computational models operating on such co-occurrence representations can acquire both semantic and syntactic properties of words. In the domain of semantics, Landauer and Dumais (1997) built a word-document matrix from a corpus of encyclopedia articles, applied a dimensionality reduction technique (singular value decomposition), and computed the similarity between words using a cosine measure. This Latent Semantic Analysis model (LSA) was able to successfully pass a TOEFL synonym test. In the domain of syntax, Redington et al. (1998) developed a model that performs hierarchical clustering on a word-word matrix to that successfully acquires grammatical categories like noun, verb, and adjective. In fact, these results suggest that co-occurrence statistics may be particularly powerful, offering leverage on many kinds of linguistic information. Although devised to learn grammatical categories, the model

often learned clusters that had semantic organization. For example, in the cluster of adjectives, color words and number words formed separate clusters (p448).

The success of this early computational work led to a proliferation of models that learn from co-occurrence statistics (see Riordan & Jones, 2010 for an overview and comparison of state-of-the-art models). The computational evidence is quite suggestive: in principle, statistical patterns of co-occurrence can drive learning lexical category learning. However, while suggestive, this work is only indirect evidence of distributional learning. Researchers studying acquisition of grammatical categories have employed methods that can in principle provide stronger evidence. These studies expose learners to artificial languages with certain co-occurrence regularities and measure whether learners form categories on the basis of these regularities. In the next section, we discuss results of studies that have examined one co-occurrence structure called the MNPQ language.

A puzzle: the MNPQ language

The MNPQ language contains four categories of words, M (which includes m_1, m_2, and m_3), N (n_1, n_2, n_3), P (p_1, p_2, p_3), and Q (q_1, q_2, q_3). Participants hear two types of training sentences[2]: MN and PQ. Thus, sentences like $m_1 n_3$ are grammatical while sentences like $m_1 q_3$ are illegal. Early investigations (Braine, 1966; K. H. Smith, 1966) found that participants tend to endorse novel grammatical sentences as familiar. However, participants also endorse ungrammatical MQ and PN sentences, suggesting that they learn position regularities (that M/P come first and N/Q come second) but not co-occurrence regularities (that M co-occurs with N but not Q and that P occurs with Q but not N). This failure to learn categories on the basis of pure co-occurrence has been reliably observed in a number of studies (Braine, 1987; Brooks et al., 1993; Frank & Gibson, 2011; Frigo & McDonald, 1998; Gerken et al., 2005; Kempe & Brooks, 2001; Lany & Saffran, 2010). However, Reeder, Newport, and Aslin (2009, Experiment

---

[2] We use the term "sentence" following its meaning in formal language theory, where it refers to a sequence of symbols derived from a formal gramar. The sentences we examine in our work are clearly impoverished relative to real sentences found in natural languages.

5) report successful learning of a related language, (Q)AXB(R). In this language, there are optional Q and R categories that serve to deconfound co-occurrence regularities from positional regularities (i.e., in MNPQ, Ms and Ps are always sentence-initial and Ns and Qs are always sentence-final). Thus, this work by Reeder et al. suggests that positional regularities may override co-occurrence regularities in the MNPQ language. Nevertheless, the MNPQ failures are puzzling, given that computational models suggest that distributional learning is a powerful mechanism.

MNPQ learning is possible, however, when co-occurrence information is partially or completely correlated with another cue. For example, Braine (1987) found successful MNPQ learning when co-occurrence information is partially correlated with natural gender. In this experiment, participants acquired an artificial language by learning to name pictures of referents. In the experimental condition, all pictures of men were labeled by Ms and all pictures of women were labeled by Ps (though not all Ms referred to men and not all Ps referred to women). Learning of the co-occurrence regularities was significantly higher in the experimental condition than in a control condition where natural gender was not correlated with M/P membership. Though Braine's experiment combined co-occurrence cues with natural gender, he suggested that phonological cues might better serve real-world language learners. For instance, Spanish and Italian speakers might learn grammatical gender categories by taking advantage of the fact that feminine nouns often end with -a, while masculine nouns often end with -o. More generally, demonstrations that co-occurrence information can be useful in concert with other information sources have encouraged researchers to study how disparate sources might be integrated to facilitate learning (e.g., Johns & Jones, 2012; Monaghan, Chater, & Christiansen, 2005).

Nearly all of this empirical work has interpreted the results of human experiments with reference to learning grammatical category. To our knowledge, only one study has examined word meaning. Recently, Lany and Saffran (2010) investigated Braine's proposal of correlating co-occurrence and phonological cues in a study of meaning acquisition. They found that 22-month old infants successfully learned an MNPQ language when co-occurrence was aligned with the number of syllables in a word (in particular, when N words were disyllabic and Q words

were monosyllabic) but *not* when the number of syllables was not predictive of N/Q membership. This suggests that distributional information may play a common role in acquisition of grammatical category and word meaning, at least for the MNPQ language.

In our current work, we add to the literature on distributional learning of word meaning by exploring a new information source: semantic coherence. To date, all studies have used the artificial language learning paradigm, with a vocabulary consisting of all novel words. Thus, at the beginning of the experiments, learners did not know the meanings of any of the words. Real learners, by contrast, typically know the meanings of some (if not most) words they hear and such words tend to relate to a single topic of discourse. Put another way, the language that real learners encounter tends to have semantic coherence; some words are known and adhere to some semantic organization. We ask: does semantic coherence facilitate distributional learning of lexical categories?

To explore this possibility, we presented participants with an MNPQ language where sentences took the form "M and N" or "P and Q". Note that we used the explicit English conjunction "and" between the two words, which imbued sentences with some semantic content (i.e., our stimuli were not merely a syntactic ordering). We hypothesized that a sufficient level of semantic coherence (specifically, a taxonomic coherence where Ms are animals and Ps were vehicles) would yield successful distributional learning for N and Q words. For instance, hearing the four sentences:

1. dog and dax

2. dog and ziv

3. car and wug

4. car and pif

might allow learners to infer that daxes and zivs belong to the same category, as both words co-occur with "dog", and that wugs and pifs belong to the same category, as both words co-occur with "car."

In Experiment 1, we tested whether semantic coherence facilitated distributional learning of lexical categories and whether learners might use these categories to make inferences about word meaning. In Experiment 2, we compared semantic coherence to phonological coherence. In Experiment 3, we compared semantic coherence to a semantic baseline that used known words that did not adhere to any obvious semantic organization.

Experiment 1: Semantic Coherence

In all the experiments reported in this paper, we exposed participants to auditory sentences from an MNPQ language and then assessed learning using three different measures. We refer to the Ms and Ps as *context words* and we refer to Ns and Qs as *target words*. Context words are found at the beginnings of sentences in our language, while target words are found at the ends of the sentences in our language (Figure 2). We systematically varied properties of the context words and then measured learning of the target words.

In Experiment 1, we parametrically varied two independent properties of the context words. First, we varied semantic coherence—the fraction of M/P words obeying a taxonomic organization (M = animal words, P = vehicle words). Second, as one hallmark of statistical learning is sensitivity to the amount of evidence observed, we varied the amount of exposure to the language to measure the efficiency of learning. After exposure to the language, we tested participants on three measures of MNPQ learning—similarity rating, sentence memory, and a referent assignment task. The similarity rating and sentence memory tasks measure whether participants acquire lexical categories, whereas the referent assignment measures whether participants can use acquired lexical categories to inform online decisions about referents for novel words.

In the similarity rating task, participants rated the similarity of pairs of target words. In the sentence memory task, participants rated the familiarity of four kinds of sentences: familiar sentences, novel grammatical sentences, and two kinds of ungrammatical sentences. . In the referent assignment task, we provided visual referents for the context words (e.g., "feeb" refers to

a cat) and asked participants to assign the target words to referents (e.g., choose whether "chuv" refers to a horse or a bicycle).

Method

Participants.   678 Amazon Mechanical Turk (MTurk) workers. Using MTurk's worker qualifications, we limited participation to workers located in the United States and with a previous HIT approval rate greater than or equal to 90%. We chose MTurk workers as our participants because the number of experimental conditions required a large number of participants. Work by Buhrmester, Kwang, and Gosling (2010) and Crump, McDonnell, and Gureckis (2013) suggests that MTurk is a valid platform for web-based learning experiments.

Materials.   Sentences took the form "M and N" or "P and Q" (Figure 2). Note that sentences literally included the word "and" in the middle. We generated the actual lexical items randomly for each participant. Ns and Qs were always novel nonsense words and were drawn without replacement from the set {moke, thite, jiv, pif, dex, wug}. Ms and Ps could be either novel or familiar. Novel Ms were drawn from {feeb, bim, lup} and novel Ps were drawn from {zabe, vap, chuv}. Familiar Ms and Ps obeyed a taxonomic organization—familiar Ms were drawn from {hamster, cat, dog} and familiar Ps were drawn from {car, bus, truck}.

To create the audio files, we input the sentences as "X. and. Y." (e.g., "car. and. chuv.", including periods) into an American English text-to-speech engine using a female voice.[3] The periods between words introduced substantial pauses ranging in length from 150 to 300 ms; piloting revealed that without pauses, it was difficult for participants to distinguish the words. Sentences using only monosyllabic words were around 2 seconds long. Sentences using the sole disyllabic word, hamster, were around 3 seconds long. The referent assignment task involved visual referents. For the context words, we used 128x128 pixel images of a cat, dog, hamster, car, bus, and truck. For the target words, we used 100x100 pixel images of a horse, rabbit, sheep, bear,

---

[3] We programmatically submitted all of our sentences to the text-to-speech web service that powers Google Translate. During our investigation, the set of voices on the web service changed, which required us to synthesize our stimuli using different software. See also Footnote 7.

Exposure sentences			Memory items		Similarity items	
$m_1\ n_1$	$m_1\ n_2$	$m_1\ n_3$	Sentence type	Example	Pair type	Example
$m_2\ n_1$	$m_2\ n_2$	$m_2\ n_3$	Familiar	$m_1\ n_2$	Within-category	n_1, n_2
$m_3\ n_1$	$m_3\ n_2$	$m_3\ n_3$	Withheld	$m_1\ n_1$	Cross-category	n_1, q_1
$p_1\ q_1$	$p_1\ q_2$	$p_1\ q_3$	Co-occurrence violation	$m_1\ q_2$		
$p_2\ q_1$	$p_2\ q_2$	$p_2\ q_3$	Position violation	$m_1\ m_2$		
$p_3\ q_1$	$p_3\ q_2$	$p_3\ q_3$				

Figure 2. The MNPQ language and test items for memory and similarity. Underlined sentences were withheld from exposure.

goldfish, mouse, boat, van, train, motorcycle, plane, and bicycle. Images are shown in Figure 3.

Design and Procedure.   We parametrically varied coherence. The language for a participant contained either 0/3, 1/3, 2/3, or 3/3 familiar M and P words each. We also varied the amount of exposure to the language—participants heard either 56, 126, 196, or 392 sentences. Before starting the experiment, we asked participants to turn on their speakers and click a button, which played a spoken English word ("airplane"). We required participants to type the word correctly to continue. The experiment had four tasks: exposure, similarity, memory, and referent assignment. Below, we detail each of these tasks.

Exposure.   Participants listened to sentences from the language. We withheld six sentences from exposure (Figure 2), yielding 14 unique sentences in the exposure set. Each sentence was heard either 4, 9, 14, or 28 times, giving 56, 126, 196, or 392 total trials. We presented the sentences in random order subject to the constraint that there were no repeated words between consecutive trials (pilot testing suggested that repeated words between trials substantially afforded learning). To encourage compliance, participants had to click a button to hear each sentence.
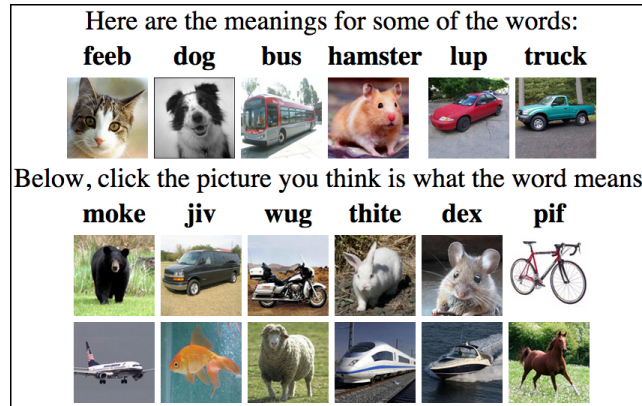
Figure 3. A screenshot of the referent assignment task.

Similarity.   For each pair of words in the union of N and Q, we asked participants to rate on a 5 point scale how similar they believed the two words to be in meaning. In particular, we instructed participants to "tell us how similar you think they are in terms of meaning". This resulted in within-category ratings (e.g., n_1 vs. n_2) and cross-category ratings (e.g., n_1 vs. q_1). We presented the pairs in a fixed pseudorandom order containing no repeated words between consecutive trials. Though exposure was entirely auditory, for convenience, we presented these similarity questions as text (e.g., "How similar are **pif** ◀◉ and **thite** ◀◉ ?"); to facilitate mapping between visual and spoken word forms, the speaker button next to each word played the spoken word when clicked. In two catch trials, we asked participants to press the response button corresponding to the solution of a simple arithmetic problem. If participants learned the MN and PQ co-occurrence relationships *and* used these relationships as a basis for lexical categorization, then we expected that participants would rate within-category pairs of words as more similar than cross-category pairs.

Memory.   Participants listened to sentences and rated on a 5 point scale how confident they were that they had heard the sentence during exposure. We tested four types of sentences:

- *Familiar* sentences heard during exposure.

- *Withheld* sentences not heard during exposure but conforming to the MNPQ structure.

- *Co-occurrence violation* sentences of the form MQ and PN.

- *Position violation* sentences of the form MM, NN, PP, and QQ.[4]

We presented sentences in random order such that there were no repeated words between consecutive trials. In two catch trials,[5] instead of a sentence from the MNPQ language, we played an audio instruction (which could not be repeated by participants) to press a specific response button. If participants learned the MN and PQ co-occurrence relationships, then we expected that they would rate novel grammatical sentences as more familiar than the co-occurrence sentences.

Referent assignment. We provided participants with referents for all of the Ms and Ps and asked them to choose referents for the Ns and Qs (Figure 3). At the top of the screen, we displayed the Ms and Ps in random order and we provided referents by displaying a single image underneath each word. The M and P referents were either animals (cat, dog, hamster) or vehicles (car, bus, truck); either Ms were animals and Ps were vehicles, or vice versa. Recall that some conditions contained *familiar* M and P words; in these cases, we paired the known words with the obvious referents (e.g., "dog" was always paired with an image of a dog). Below the M and P words and their meanings, we displayed a row containing the N and Q words. Under each word, we displayed a two-alternative referent choice between an animal (the "correct" choice for N words) and vehicle words (the "correct" choice for Q words); participants made a choice by clicking on one of the two pictures. If participants learned the MN and PQ co-occurrence relationships *and* used them to form nascent lexical categories *and* used these lexical categories as a basis for inferences about word meaning, then we expected that referent assignment scores would reflect a tendency to choose similar referents for words in the same category. In addition,

---

[4] We only used position violation sentences that did not repeat the same word twice, e.g., we never presented $m_1 - m_1$.

[5] In initial data collection, we did not include catch trials. Let [A/3;B] denote the experimental condition with A/3 coherence and B exposures (e.g., [2/3;196] refers to the 2/3 coherence level with 196 exposures). In [0/3;196], 18 out of 40 participants did not receive catch trials. In [3/3;56], 30 out of 43 participants did not receive catch trials. In [3/3;126], 30 out of 40 participants did not receive catch trials. In [3/3;196], 30 out of 40 participants did not receive catch trials.
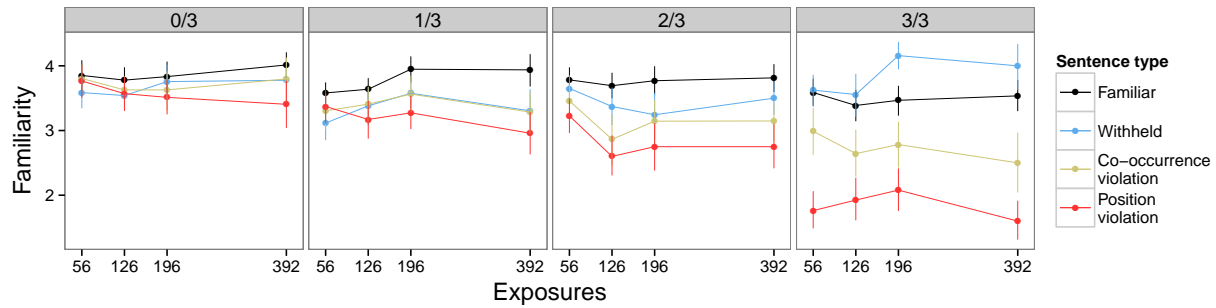
Figure 4. Mean familiarity ratings for differente sentence types in Experiment 1.

we hypothesized that, under these circumstances, participants would pick animals for Ns and vehicles for Ps. This is because Ns consistently co-occurred with Ms, which we indicated referred to animals, and Qs consistently co-occurred with Ps, which we indicated referred to vehicles.

Results and Discussion

We excluded the 55 participants who did not correctly answer all of the catch trials.[6] We first present descriptive statistics of the memory task before analyzing the influence of coherence and exposure on distributional learning as measured by all three tasks.

Descriptive statistics for the memory task. Figure 4 shows mean familiarity ratings broken down by sentence type, coherence, and exposures. Within each coherence level, we looked for evidence of a familiarity ordering on sentence types. For each participant, we computed their mean familiarity rating for each sentence type. Then, at each coherence level, we ran 6 paired-sample t-tests comparing mean familiarity ratings across sentence types and corrected for multiple comparisons using the Holm procedure (Figure B1). This yielded the following familiarity orderings:

0/3: F > W, P

---

[6] Recall that some participants – mostly in the 3/3 conditions – did not receive catch trials (see Footnote 5). This does not bias the results in favor of our hypothesis; if anything, this weakens any effect of semantic coherence, as it introduces noise at the highest level of semantic coherence but not at any other level. Indeed, our results are not affected by including subjects who had catch trials and failed them.

1/3: F > W, C, P   **&**   C > P

2/3: F > W > C > P

3/3: W > F > C > P

These familiarity orderings indicate that some learning occurred at every level of coherence. For example, participants at all levels of coherence rated familiar sentences as more familiar than position violation sentences. However, only participants at the 2/3 and 3/3 levels distinguished between withheld and co-occurrence violation sentences. This lends some support to the idea that positional regularities in MNPQ may be more influential than co-occurrence regularities.

Effect of coherence and exposure.   For each dependent measure—memory, similarity, and referent assignment—we defined a within-participant score representing the sensitivity to the co-occurrence regularities in the language. We scaled all scores to lie in the interval [-1, 1], which allows for the following interpretation: scores below 0 indicate incorrect learning (e.g., preferring co-occurrence violation sentences over withheld sentences), while scores above 0 indicate correct behavior.

The memory score measured the difference in familiarity ratings for withheld versus co-occurrence violation sentences. For each participant, we computed their memory score as the mean rating for withheld sentences (e.g., $m_1 - n_1$) minus their mean rating for novel co-occurrence violation sentences (e.g., $m_1 - q_1$). This difference had a minimum value of -4 and a maximum value of 4; we divided this difference by 4 to rescale it to lie in the interval [-1, 1]. Note that memory score did not include ratings for familiar or position-violation sentences.

The similarity score measured the difference between similarity ratings for within-category and cross-category pairs of words on the similarity task. For each participant, we computed their similarity score as their mean rating for within category pairs (e.g., $m_1 - n_1$) minus their mean rating for cross category pairs (e.g., $m_1 - q_1$). As with the memory scores, this yields a measure that lies in the interval [-4, 4], which we scaled to [-1, 1].

The referent assignment score measured the total number of correct choices in the referent assignment task. The lowest score was 0 and the highest possible score was 6. We rescaled this to

the interval [-1, 1] by taking the raw number correct, subtracting 3, and then dividing by 3.

We analyzed the effects of coherence and exposure using linear models. First, we looked for main effects of coherence on score (i.e., the Condition coefficients in Table 1). Second, as we were interested in the relationship between amount of exposure and score, we looked for exposure $\times$ coherence interactions. A significant interaction (e.g., the E $\times$ C coefficients in Table 1) would indicate a difference in how *efficiently* the statistical learning process makes use of evidence at different coherence levels. For all scores, we coded coherence as a categorical variable and analyzed the data using a regression which modeled the mean score in a participant group (e.g., 3/3-56) as an interactive function of the number of exposures (e.g., 56) times the condition (e.g., full coherence). In other words, our regression equation was score $\sim$ exposures $\times$ condition. Results are shown in Figure 5.

To examine the differences between the different coherence levels, we used Helmert contrasts analyzing (i) the difference between the 1/3 and 0/3 conditions, (ii) the difference between the 2/3 condition and the 0/3 and 1/3 conditions combined, and (iii) the difference between the 3/3 condition and the 0/3, 1/3, and 2/3 conditions combined. Results of these analyses are shown in Table 1. Before detailing the results for each measure, we will first state the two broad patterns of results. First, learning was highest in 3/3 condition. Second, we found the strongest evidence of statistical efficiency (i.e., sensitivity to the amount of exposure) in the 3/3 condition.

Task results.   We report performance on each of the three tasks separately.

Memory.   More exposure led to higher memory scores, as there was a main effect of exposure. Higher coherence also led to higher memory scores; 2/3 scores were significantly higher than scores from the 0/3 and 1/3 conditions combined and 3/3 scores were significantly higher than scores from the rest of the conditions combined. Because 2/3 and 3/3 scores both outperformed all the respective lower levels of coherence, we also computed this model using coherence as a continuous variable; the continuous coherence regressor significantly predicted increases in score, $\beta = 0.06$, $t(619) = 3.86$, $p < 0.0005$, suggesting that parametrically
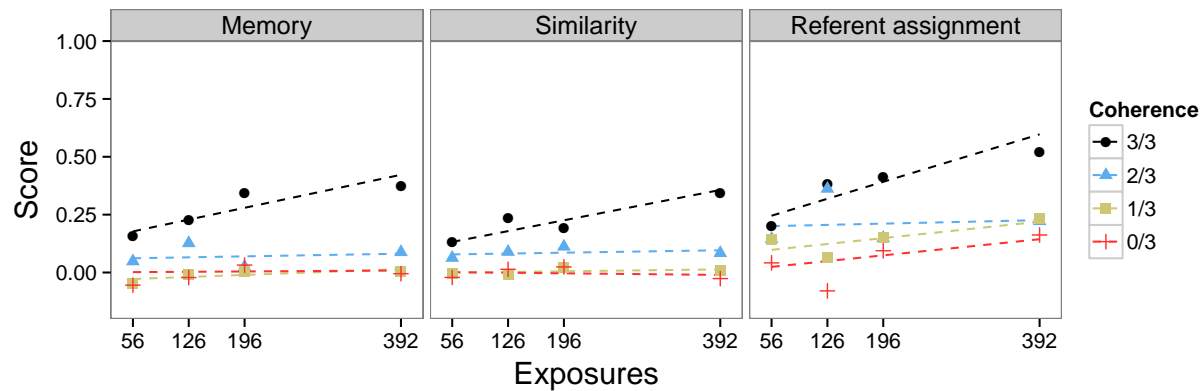
Figure 5. Experiment 1 results. Each plot shows data for one measure (memory, similarity, referent assignment) in Experiment 1. Scores ranged from -1 to 1. Data points show condition means, error bars show 95% CIs, and dashed lines show the best-fitting linear regression.

increasing coherence results in parametric increases in memory score.

Additionally, the effect of exposures on memory score was significantly higher in 3/3 than in the other conditions combined (i.e., there was a significant exposure $\times$ coherence interaction), suggesting greater efficiency of statistical learning in 3/3. Thus, more semantically coherent linguistic input (1) bolstered memory for the MN and PQ co-occurrence regularities and (2) increased the efficiency of the statistical learning process responsible for learning those regularities, at least in the 3/3 condition.

Similarity.   Similarity scores were higher in the 3/3 condition than the other conditions combined. In addition, the effect of exposures on score was significantly higher in the 3/3 condition than in the other conditions combined, as there was a significant exposure $\times$ coherence interaction. Thus, more coherent linguistic input (1) increased the distinction between within-category and cross-category pairs of words and (2) increased the efficiency of the statistical learning of such distinctions, at least in the 3/3 condition.

Referent assignment.   2/3 scores were significantly higher than 0/3 and 1/3 scores combined. 3/3 scores were marginally higher than the rest of the scores combined, possibly because 3/3 coherence and 2/3 coherence may confer comparable advantages on this task. We

also computed this model using coherence as a continuous variable; the continuous coherence regressor significantly predicted increases in score, $\beta = 0.09$, $t(619) = 2.87$, $p < 0.005$, suggesting that *parametrically* increasing coherence results in *parametric* increases in referent assignment score. None of the interaction terms reached significance, indicating that the amount of exposure to the language and greater coherence independently increased the ability to assign N and Q words to the correct referents.

Summary.   In Experiment 1, we found that semantic coherence (1) increased ability to distinguish novel grammatical sentences from sentences violating co-occurrence regularities in the memory task, (2) sharpened sensitivity to lexical category boundaries based on the co-occurrence regularities in the similarity task, and (3) increased inductive bias in associating target words with objects in the referent assignment task.

The context words that participants heard in Experiment 1 included words with familiar meanings that were organized into coherent categories. Would it be enough to merely have familiar words without coherence or the reverse? How does the effect of semantic coherence depend on each? In Experiment 2 we test for the effect of meaning and in Experiment 3, we test for the effect of coherence. In Experiment 2, we remove familiar meaning by exposing learners to languages with phonological, as opposed to semantic, coherence. In Experiment 3, we remove coherence by exposing learners to languages with context words that are familiar but do not adhere to any obvious semantic organization.

Experiment 2: Phonological coherence

In Experiment 2, we investigated whether learners could learn the language used in Experiment 1 when the context words (Ms and Ps) exhibited phonological, rather than semantic, coherence. It is worth emphasizing that we manipulated phonological coherence in the *context* words, whereas previous work has manipulated phonological coherence of the *target* words. Previous work has shown that learning is facilitated when target words have both distributional and phonological regularities. By contrast, we are interested in how distributional learning for one

category of words varies as a function of *other* words in the linguistic environment.

We tested three types of coherence: onset, rime, and syllable count. We selected these kinds of coherence because work in other areas suggests that these kinds of coherence may reliably occur in natural language or facilitate distributional learning when applied to *target* words. Onset coherence might facilitate learning due to sound symbolism (e.g., words that start with *gl* have similar meanings: glimmer, glow, gleam). Rime coherence occurs in languages like Russian, which mark the gender of a noun by modifying the rime of the corresponding verb. Finally, syllable count coherence has been shown to facilitate learning when applied to target words. Thus, we investigated it to determine whether it also facilitates learning when applied to context words.

Method

Participants. 530 MTurk workers, recruited as in Experiment 1, participated in the study.

Materials. The three types of phonological coherence[7] were:

- *Onset*. Ms all started with one consonant cluster (pladge, plaaf, plab) and Ps all started with another (zof, zawd, zawsh).

- *Rime*. Ms all ended with one vowel (calo, pawmo, marfo) and Ps all ended with another (zaygee, kaisee, tetchee).

- *Syllable count*. Ms were disyllabic (coomo, fengle, kaisee) and Ps were monosyllabic (gope, jic, skeege).

Design and Procedure. The method was identical to that of Experiment 1.

---

[7] The stimuli for the rime and syllable count conditions differ from those in the rest of our conditions. For the rest of the conditions, we used a text-to-speech web service provided by Google to generate the audio stimuli (see Footnote 3) for the bulk of the conditions. However, the available voices on this service changed during our experiment. Thus, we generated new stimuli for the rime and syllable count conditions using commercially available software, NaturalReader 10. To ensure that the old and new stimuli were comparable, we performed a partial replication of Experiment 1 using the new synthesis engine; the difference old and new stimuli did not appear to make a substantial difference. See Appendix A for comparisons.
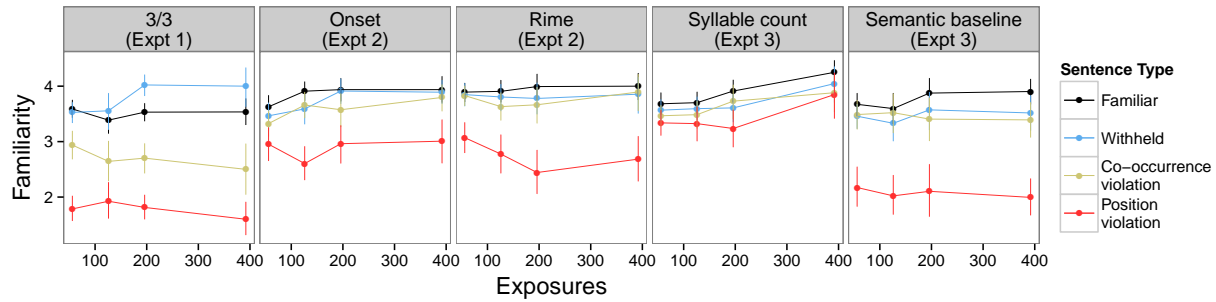
Figure 6. Mean familiarity ratings for different sentence types in Experiments 2 and 3. For comparison, we also include the 3/3 coherence condition from Experiment 1.

Results and Discussion

We discarded the 42 participants who did not pass all the catch trials. Results are graphed in Figure 7.

Descriptive statistics for the memory task.   Figure 6 shows mean familiarity ratings broken down by sentence type, coherence, and exposures. As in the analyses for Experiment 1, we computed a familiarity ordering on sentence types within each phonological coherence condition using paired sample t-tests (Figure B1):

Onset: $F > W > C > P$

Rime: $F > W \approx C > P$

Syllable count: $F > W \approx C > P$

Thus, participants in the rime and syllable count conditions distinguished between all types of sentences except for withheld and co-occurrence violation sentences, indicating some learning of the language but not of the co-occurrence regularities. By contrast, participants in the onset condition did appear to learn the co-occurrence regularities, rating withheld sentences as significantly more familiar than co-occurrence violation sentences. The ordering for onset phonological coherence is the same as the ordering for 2/3 semantic coherence. To understand the degree of co-occurrence structure learning for the onset condition, we compared onset memory scores with 2/3 and 3/3 memory scores. Onset scores (M = 0.03, SD = 0.17) were not
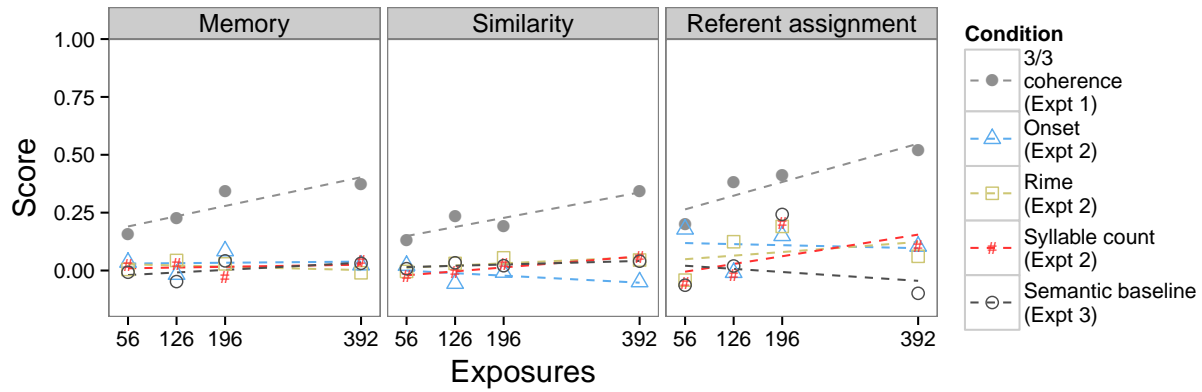
Figure 7. Experiments 2 and 3 results. Each plot shows data for one measure (memory, similarity, referent assignment). Scores ranged from -1 to 1. Points show condition means, error bars show 95% CIs, and dashed lines show the best-fitting linear trend. For comparison, we also include the 3/3 coherence condition from Experiment 1.

significantly different from 2/3 scores (M = 0.06, SD = 0.23), $t(314)$ = -1.5, p = 0.13, but they were significantly lower than 3/3 scores (M = 0.27, SD = 0.36), $t(302)$ = -7.31, $p < 0.001$.

Effects of condition and exposure.   Using a regression model with main effects of exposure and condition and an exposure × condition interaction, we compared each phonological condition with the 0/3 condition of Experiment 1 using a regression model (Table 2).

Memory.   There was no effect of exposure or condition—none of the phonological condition scores were significantly different from 0/3 scores. None of the exposure by condition interaction terms were significant.

Similarity.   Again, we found no main effect of exposure or condition. One interaction term was significant: there appeared to be greater efficiency of statistical learning in the syllable count condition than in the 0/3 condition.

Referent assignment.   Again, we found no main effect of exposure or condition. None of the exposure by condition interaction terms were significant.

In sum, with one exception, participants in all three phonological conditions appeared not to have acquired separate N and Q categories; phonological coherence was virtually

indistinguishable from 0/3 coherence in terms of facilitating MNPQ learning. The one exception is that participants in the onset coherence condition were able to distinguish withheld from co-occurrence violation sentences, matching 2/3 participants' performance in this regard. However, unlike 2/3 coherence, onset coherence did not appear to confer benefits in the similarity or referent assignment tasks. Thus, this broad pattern of results suggests that the mere presence of *some* kind of coherence in the context words is not what drives the effects of semantic coherence. In Experiment 3, we consider whether the mere presence of familiar words (semantic baseline) aids MNPQ learning.

Experiment 3: Semantic baseline

In Experiment 1, Ms and Ps were all familiar words obeying a taxonomic organization. In Experiment 3, we explored a language with familiar words but no semantic organization—that is, whether a semantic baseline language facilitates distributional learning. We might expect this baseline condition to facilitate learning due to lower memory demands—known words tax the memory system less, which might free learners to identify co-occurrence regularities.

Methods

Participants. 162 MTurk workers, recruited as in Experiment 1. Note that there were many fewer participants in Experiment 3 because there was only one condition, compared to four conditions in Experiment 1 and three conditions in Experiment 2.

Materials. In the semantic baseline language, the specific M and P words were drawn randomly for each participant from the pool {*shelf*, *glove*, *rain*, *leash*, *card*, *ball*}. In the referent assignment task, these known words were paired with images of the obvious referents (e.g., *card* with a picture of a card). Compared to the context words in Experiment 1, these words are fairly well matched in frequency (see 3).

Design and procedure. The method was identical to that of Experiment 1.

Results and Discussion

We discarded the 18 participants who did not pass all the catch trials.

Descriptive statistics of the memory task.   As in Experiment 1, we computed an ordering on the sentence types in the memory task: $F > W \approx C > P$. This indicates that participants were able to learn some regularities in the language, but not the co-occurrence regularities.

Effect of exposure.   Results are graphed in Figure 7. See Table 2 for regression results.

Memory.   Baseline scores were not significantly different from 0/3 scores and baseline efficiency was not significantly different from 0/3 efficiency.

Similarity.   Baseline scores were not significantly different from 0/3 scores and baseline efficiency was not significantly different from 0/3 efficiency.

Referent assignment.   Baseline scores were not significantly different from 0/3 scores and baseline efficiency was not significantly different from 0/3 efficiency.

Apparently, the baseline input appeared to have provided no benefit compared to the novel words of the 0/3 condition, suggesting that the presence of known words by itself does not aid MNPQ learning.

General Discussion

Can you learn lexical categories from co-occurrence statistics? Previous work on distributional learning based on co-occurrence presented a paradox: While computational models suggest that distributional statistics are a powerful source of information, experiments with humans show consistent failures. Our experiments suggest a partial resolution. For people, distributional learning is facilitated by semantic coherence: Knowing that Alice associated with X and Y tells us something about Alice only if X and Y are already meaningful and meaningfully related to one another. Removing either the meanings of the individuals or the coherence of the relationship between them (cf. "word salad") removes the facilitation.

What is the mechanism by which semantic coherence facilitates learning? Perhaps learners use semantic coherence to infer the topic of discourse and then tag novel words with these topics.

For example, in the highest semantic coherence condition of our experiments, participants may have learned that the topic of discourse is either animal-related or vehicle-related and then tracked co-occurrence between these topics and novel words (Frank, Tenenbaum, & Fernald, 2013). Such a proposal is congruent with the "preferential acquisition" idea of Hills and colleagues (2010; 2009), in which the easiest words to learn are those that connect well with others in the learning environment.

Our finding that phonological coherence does not facilitate learning may at first appear at odds with experiments that have reached the opposite conclusion (Frigo & McDonald, 1998; Lany & Saffran, 2010; Monaghan et al., 2005). But previous work has applied phonological regularities to the target words (the words that experimenters measure learning for). In contrast, in order to permit comparison with the semantic coherence conditions, we applied the regularities to the context words (the words that co-occurred with the target words). In other words, previous experiments used target coherence, while we used context coherence. In our experiment, the target categories themselves did not have phonological regularity but they reliably co-occurred with context categories that did. We speculate that this subtle but substantial difference explains the different patterns we observed.

Limitations and Future Directions

The limitations of our approach may serve as inspiration for future work. First, we explored a single artificial language, MNPQ, because past research has shown that it is largely resistant to distributional learning. While less is known about other languages, Reeder et al. (2009) found successful distributional learning for the (Q)AXB(R) language, which fixes a positional confound in MNPQ (Ms and Ps are always sentence-initial and Ns and Qs are always sentence-final), suggesting that learning is sensitive to the language structure. How do variations across language structure interact with factors like semantic and phonological coherence?

Second, we investigated a single type of semantic coherence, where context words fell into two different taxonomic categories, and a single syntactico-semantic construction that biases

interpretation towards concrete nouns (e.g., "cat and wug"). These choices likely affected the magnitude of the effects we observed. While the simplicity and transparency of our semantic categories likely facilitated learning, our use of concrete noun targets might even have understated the effects of coherence. Distributional learning through semantic coherence might be especially useful for words like verbs and adjectives (Redington et al., 1998), where referential grounding provides less information about meaning (Gleitman, 1990).

Finally, while our study addressed one shortcoming of much artificial language research—the lack of semantic grounding—it is still congruent with other artificial language research in scale (6 words in a single exposure session). Small-scale experiments do not allow strong inferences about the importance of different factors in lexical learning (Frank et al., 2013; Romberg & Saffran, 2010). A mixture of computational analyses and larger-scale experiments are necessary to understand the role semantic coherence—and distributional learning more generally—plays in language learning.

Conclusion

Since Firth's edict that we shall know a word "by the company it keeps," two lines of research have yielded apparently conflicting results. Although computational models suggest that distributional information is a powerful cue towards categorization, artificial language learning experiments have suggested a more dubious outlook. Our results here may help to close this gap between the experimental and the computational. On the basis of our findings, we suggest a small modification to Firth's maxim: You shall know a word *if you know* the company it keeps.

Table 1

Regression model for Experiment 1. For readability, exposure values were divided by 1000.

Regressor	β	Std Error	t	p
Memory: $R^2 = 0.18$, $F(7, 615) = 20.01$				
Intercept	0.0359	0.018	1.91	0.056
Condition: 1/3 – (0/3)	-0.0005	0.027	-0.01	0.984
Condition: 2/3 – (0/3,1/3)	0.0309	0.014	2.09	<0.05*
Condition: 3/3 – (0/3,1/3,2/3)	0.0396	0.010	3.65	<0.001*
Exposures	0.2346	0.083	2.81	<0.005*
E × C: 1/3 – (0/3)	0.0033	0.120	0.02	0.977
E × C: 2/3 – (0/3,1/3)	-0.0215	0.065	-0.32	0.743
E × C: 3/3 – (0/3,1/3,2/3)	0.1326	0.048	2.73	<0.01*
Similarity: $R^2 = 0.12$, $F(7, 615) = 12.09$				
Intercept	0.0501	0.018	2.64	<0.01*
Condition: 1/3 – (0/3)	-0.0070	0.027	-0.25	0.798
Condition: 2/3 – (0/3,1/3)	0.0237	0.014	1.59	0.112
Condition: 3/3 – (0/3,1/3,2/3)	0.0224	0.010	2.04	<0.05*
Exposures	0.1487	0.084	1.77	0.077
E × C: 1/3 – (0/3)	0.0514	0.121	0.42	0.671
E × C: 2/3 – (0/3,1/3)	0.0208	0.066	0.31	0.754
E × C: 3/3 – (0/3,1/3,2/3)	0.1373	0.048	2.80	<0.01*
Referent assignment: $R^2 = 0.06$, $F(7, 611) = 6.45$				
Intercept	0.1087	0.036	3.01	<0.005*
Condition: 1/3 – (0/3)	0.0664	0.052	1.26	0.208
Condition: 2/3 – (0/3,1/3)	0.0616	0.028	2.17	<0.05*
Condition: 3/3 – (0/3,1/3,2/3)	0.0359	0.020	1.72	0.084
Exposures	0.4577	0.159	2.86	<0.005*
E × C: 1/3 – (0/3)	-0.0919	0.230	-0.39	0.690
E × C: 2/3 – (0/3,1/3)	-0.1257	0.126	-0.99	0.319
E × C: 3/3 – (0/3,1/3,2/3)	0.1292	0.093	1.38	0.167

Table 2

Regression model for Experiments 2 and 3. For readability, exposure values were divided by 1000.

Predictor	β	Std Error	t	p
Memory: $R^2 = 0.009$, $F(7, 631) = 0.88$				
Intercept	-0.0341	0.027	-1.24	0.215
Condition: Onset – 0/3	0.0626	0.038	1.64	0.100
Condition: Rime – 0/3	0.0652	0.035	1.82	0.068
Condition: Syllable count – 0/3	0.0410	0.038	1.06	0.287
Condition: Semantic incoherent – 0/3	0.0055	0.038	0.14	0.883
Exposures	0.1201	0.119	1.00	0.316
E × C: Onset – 0/3	-0.0948	0.170	-0.55	0.577
E × C: Rime – 0/3	-0.1950	0.161	-1.20	0.228
E × C: Syllable count – 0/3	-0.0735	0.167	-0.43	0.660
E × C: Semantic incoherent – 0/3	0.0388	0.162	0.23	0.811
Similarity: $R^2 = 0.02$, $F(7, 631) = 2.51$				
Intercept	0.0110	0.025	0.43	0.662
Condition: Onset – 0/3	-0.0031	0.035	-0.08	0.928
Condition: Rime – 0/3	-0.0064	0.033	-0.19	0.846
Condition: Syllable count – 0/3	-0.0455	0.035	-1.27	0.201
Condition: Semantic incoherent – 0/3	-0.0007	0.035	-0.02	0.982
Exposures	-0.0607	0.110	-0.54	0.582
E × C: Onset – 0/3	-0.0937	0.157	-0.59	0.550
E × C: Rime – 0/3	0.1951	0.149	1.30	0.191
E × C: Syllable count – 0/3	0.3049	0.154	1.97	<0.05*
E × C: Semantic incoherent – 0/3	0.1401	0.150	0.93	0.350
Referent assignment: $R^2 = 0.01$, $F(7, 629) = 1.12$				
Intercept	-0.0553	0.068	-0.81	0.418
Condition: Onset – 0/3	0.1776	0.094	1.88	0.060
Condition: Rime – 0/3	0.0916	0.088	1.03	0.301
Condition: Syllable count – 0/3	0.0234	0.095	0.24	0.806
Condition: Semantic incoherent – 0/3	0.0868	0.094	0.91	0.359
Exposures	0.5462	0.296	1.84	0.066
E × C: Onset – 0/3	-0.6122	0.421	-1.45	0.146
E × C: Rime – 0/3	-0.3243	0.400	-0.80	0.418
E × C: Syllable count – 0/3	-0.0692	0.414	-0.16	0.867
E × C: Semantic incoherent – 0/3	-0.7403	0.402	-1.83	0.066

Table 3

Context word frequencies from Experiments 1 and 3 sorted by decreasing frequency. Frequencies were retrieved from Google N-grams data for the year 1999.

Word	Frequency (%)	Experiment
Car	0.0091	1
Dog	0.0042	1
Card	0.0035	3
Ball	0.0035	3
Rain	0.0030	3
Bus	0.0026	1
Cat	0.0021	1
Truck	0.0017	1
Shelf	0.00094	3
Glove	0.00029	3
Leash	0.00015	3
Hamster	0.000099	1

References

Baldwin, D. A. (1993). Early referential understanding: Infants' ability to recognize referential acts for what they are. *Developmental Psychology*, *29*(5), 832.

Berry-Rogghe, G. L. M. (1973). The Computation of Collocations and their Relevance in Lexical Studies. In A. Aitken, R. Bailey, & N. Hamilton-Smith (Eds.), *The computer and literary studies* (pp. 103–112). Edinburgh: Edinburgh University Press.

Braine, M. D. S. (1966). Learning the positions of words relative to a marker element. *Journal of Experimental Psychology*, *72*(4), 532–540.

Braine, M. D. S. (1987). What is learned in acquiring word classes: A step toward an acquisition theory. In *Mechanisms of language acquisition* (pp. 65–87). Hillsdale, NJ: Lawrence Erlbaum.

Brooks, P. J., Braine, M. D. S., Catalano, L., Brody, R. E., & Sudhalter, V. (1993). Acquisition of Gender-like Noun Subclasses in an Artificial Language: The Contribution of Phonological Markers to Learning. *Journal of Memory and Language*, *32*(1), 76–95.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2010). Amazon's Mechanical Turk. *Perspectives on Psychological Science*.

Church, K., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*(1), 22–29.

Clark, E. V. (1988). On the logic of contrast. *Journal of Child Language*, *15*(02), 317.

Clark, H. (1968). On the use and meaning of prepositions. *Journal of Verbal Learning and Verbal Behavior*, *7*(2), 421–431.

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PloS One*, *8*(3), e57410.

Firth, J. R. (1957). A Synopsis of Linguistic Theory 1930-1955. In *Studies in linguistic analysis* (pp. 1–32). Oxford: Blackwell.

Frank, M. C., & Gibson, E. (2011). Overcoming memory limitations in rule learning. *Language Learning and Development*, *7*(2), 130–148.

Frank, M. C., Tenenbaum, J. B., & Fernald, A. (2013). Social and Discourse Contributions to the Determination of Reference in Cross-Situational Word Learning. *Language Learning and Development*, *9*(1), 1–24.

Frigo, L., & McDonald, J. L. (1998). Properties of Phonological Markers That Affect the Acquisition of Gender-Like Subclasses. *Journal of Memory and Language*, *39*(2), 218–245.

Geffroy, A., Lafon, P., Seidel, G., & Tournier, M. (1973). Lexicometric analysis of cooccurrences. In A. Aitken, R. Bailey, & N. Hamilton-Smith (Eds.), *The computer and literary studies*. Edinburgh: Edinburgh University Press.

Gerken, L., Wilson, R., & Lewis, W. (2005). Infants can use distributional cues to form syntactic categories. *Journal of Child Language*.

Gleitman, L. (1990). The Structural Sources of Verb Meanings. *Language Acquisition*, *1*(1), 3–55.

Harris, Z. S. (1951). *Methods in structural linguistics*. Chicago: University of Chicago Press.

Hills, T. T., Maouene, J., Riordan, B., & Smith, L. B. (2010). The associative structure of language: Contextual diversity in early word learning. *Journal of Memory and Language*, *63*(3), 259–273.

Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. (2009). Longitudinal Analysis of Early Semantic Networks: Preferential Attachment or Preferential Acquisition? *Psychological Science*, *20*(6), 729–739.

Hollich, G. J., Hirsh-Pasek, K., & Golinkoff, R. M. (2000). Breaking the Language Barrier: An Emergentist Coalition Model for the Origins of Word Learning. *Monographs of the Society for Research in Child Development*, *65*(3), i–135.

Johns, B. T., & Jones, M. (2012). Perceptual Inference Through Global Lexical Similarity. *Topics in Cognitive Science*, *4*(1), 103–120.

Kempe, V., & Brooks, P. J. (2001). The Role of Diminutives in the Acquisition of Russian Gender: Can Elements of Child-Directed Speech Aid in Learning Morphology? *Language*

Learning, *51*(2), 221–256.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211.

Lany, J., & Saffran, J. (2010). From Statistics to Meaning: Infants' Acquisition of Lexical Categories. *Psychological Science*, *21*(2), 284.

Maratsos, M. P., & Chalkley, M. A. (1980). The internal language of children's syntax: The ontogenesis and representation of syntactic categories. In K. E. Nelson (Ed.), *Children's language* (pp. 127–214). Gardner.

Markman, E. M. (1991). *Categorization and naming in children: Problems of induction*. Cambridge, MA: MIT Press.

Monaghan, P., Chater, N., & Christiansen, M. H. (2005). The differential role of phonological and distributional cues in grammatical categorisation. *Cognition*.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, *22*(4), 425–469.

Reeder, P., Newport, E., & Aslin, R. (2009). The role of distributional information in linguistic category formation. *Proceedings of the 31st Annual Meeting of the Cognitive Science Society*.

Riordan, B., & Jones, M. N. (2010). Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*, *3*(2), 303–345.

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(6), 906–914.

Rubenstein, H., & Goodenough, J. (1965). Contextual correlates of synonymy. *Communications of the ACM*, *8*(10), 627–633.

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

Saffran, J., Newport, E., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*(4), 606–621.

Salton, G., & McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc.

Schütze, H. (1992). Dimensions of meaning. *Proceedings of the 1992 ACM/IEEE conference on Supercomputing*, 787–796.

Smith, K. H. (1966). Grammatical intrusions in the recall of structured letter pairs: Mediated transfer or position learning? *Journal of Experimental Psychology*, *72*(4), 580–588.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558–1568.

Stefflre, V., & Reich, P. (1971). Some eliciting and computational procedures for descriptive semantics. In P. Kay (Ed.), *Explorations in mathematical anthropology.* Cambridge, Mass.: MIT Press.

Szalay, L., & Bryson, J. (1974). Psychological meaning: Comparative analyses and theoretical implications. *Journal of Personality and Social Psychology*.

Wittgenstein, L. (1953/1997). *Philosophical Investigations*. Oxford, UK; Malden, Mass.: Blackwell.

Yu, C., & Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, *70*(13-15), 2149–2165.

Appendix A

Old versus new stimuli comparison

We performed a partial replication of Experiment 1 with the new stimuli. We collected data in four out of the sixteen experimental conditions and the results mirror those found with the old stimuli:
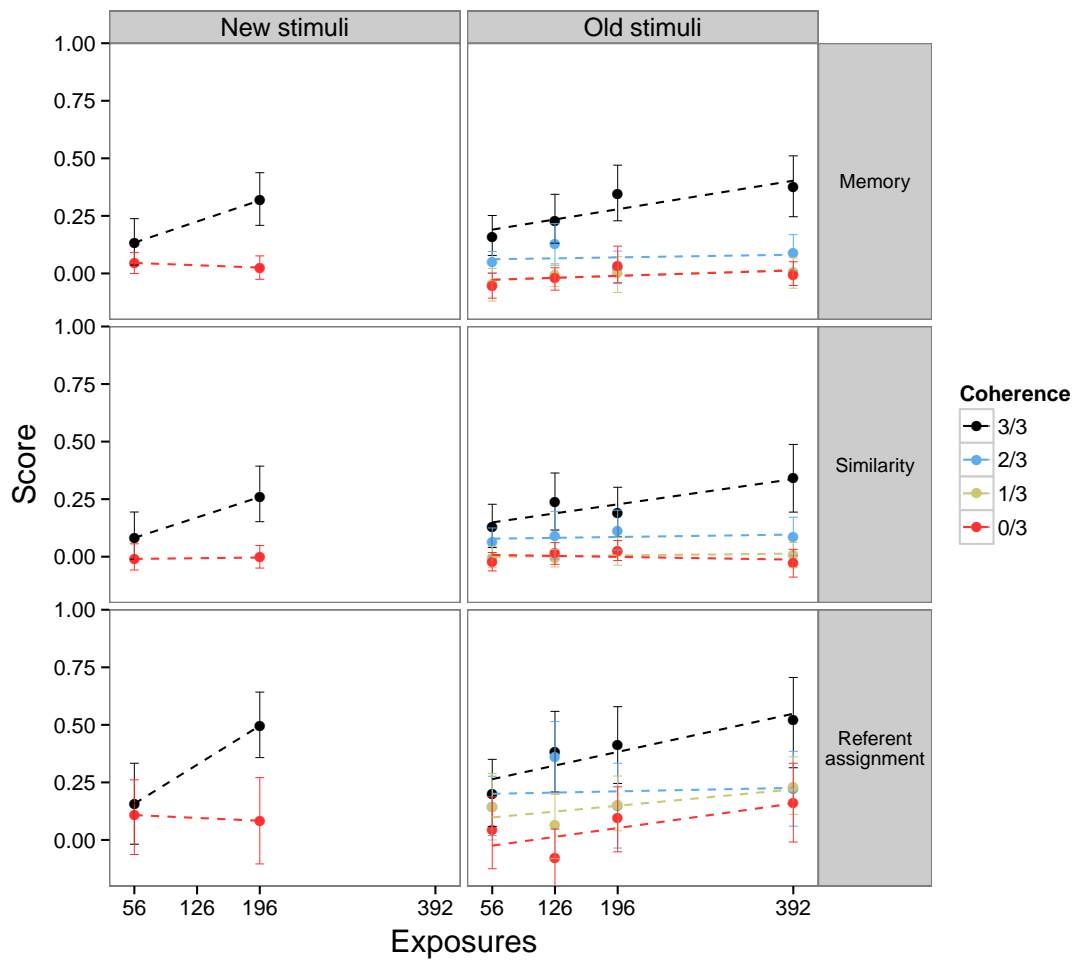
Appendix B

t-tests for familiarity orderings in the memory task

0/3 coherence: df = 150

	F	W	C
W	$p = 0.01*$		
C	$p = 0.06$	$p = 0.48$	
P	$p < 0.001*$	$p = 0.21$	$p = 0.08$

1/3 coherence: df = 153

	F	W	C
W	$p < 0.001*$		
C	$p < 0.001*$	$p = 0.52$	
P	$p < 0.001*$	$p = 0.08$	$p = 0.01*$

2/3 coherence: df = 164

	F	W	C
W	$p < 0.001*$		
C	$p < 0.001*$	$p < 0.001*$	
P	$p < 0.001*$	$p < 0.001*$	$p < 0.001*$

3/3 coherence: df = 152

	F	W	C
W	$p < 0.001*$		
C	$p < 0.001*$	$p < 0.001*$	
P	$p < 0.001*$	$p < 0.001*$	$p < 0.001*$

Onset: df = 150

	F	W	C
W	$p = 0.04*$		
C	$p < 0.001*$	$p = 0.04*$	
P	$p < 0.001*$	$p < 0.001*$	$p < 0.001*$

Rime: df = 187

	F	W	C
W	$p = 0.09$		
C	$p = 0.002*$	$p = 0.15$	
P	$p < 0.001*$	$p < 0.001*$	$p < 0.001*$

Syllable count: df = 148

	F	W	C
W	$p < 0.006*$		
C	$p < 0.001*$	$p = 0.30$	
P	$p < 0.001*$	$p = 0.001*$	$p = 0.01*$

Semantic baseline: df = 143

	F	W	C
W	$p < 0.001*$		
C	$p < 0.001*$	$p = 0.83$	
P	$p < 0.001*$	$p < 0.001*$	$p < 0.001*$

Figure B1. *p* values for paired-sample t-tests of familiarity ratings for (F)amiliar, (W)ithheld, (C)o-occurrence violation, and (P)osition violation sentences in Experiments 1, 2, and 3. *p* values have been adjusted for multiple comparisons using the Holm procedure.