Semantic coherence facilitates learning word meanings through contextual co-occurrence

Long Ouyang, Lera Boroditsky, Michael C. Frank

Department of Psychology, Stanford University

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

Please address correspondence to Long Ouyang, Department of Psychology, Stanford University, 450 Serra Mall (Jordan Hall), Stanford, CA, 94305, tel: (650) XYZ-XYZZ, email: louyang@stanford.edu

Abstract

Words' contextual co-occurrence with each other may be an important source of information about their meanings. For example, a learner might notice that *postman* and *mailman* occur in contexts with very similar sets of other words and infer that they have similar meanings. Computational models demonstrate that contextual co-occurrence learning is a possible information source for language learning, but artificial language learning investigations have suggested that human capacity for this sort of learning is quite limited (unless co-occurrence regularities are signaled by another redundant cue). Experiments in this paradigm typically present learners with words that are all unknown to start, however. By contrast, real language learners typically encounter input that contains some known words that adhere to a semantic organization. In three experiments, we show that the use of known semantic reference points facilitates contextual co-occurrence learning and that this effect crucially depends both on the presence of known words and the adherence of these known words to some semantic organization.

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

Learning the meanings of words is, in principle, an arbitrarily hard problem (Quine, 1960; Carey, 1978). Fortunately for learners, this hard problem is eased by a multiplicity of sources of information about word meaning: physical, social, conceptual, and linguistic cues all can provide useful information for learning different kinds of words (Clark, 1988; Markman, 1991; Gleitman, 1990; Baldwin, 1993; Hollich, Hirsh-Pasek, & Golinkoff, 2000). Much recent research has focused on information that learners can glean from distributional properties of language input. 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 (Yu & Ballard, 2007; L. B. Smith & Yu, 2008). In the current paper, we explore a more sophisticated type of distributional learning: learning from co-occurrences between words and their contexts (K. Smith, 1966; Maratsos & Chalkley, 1980; Braine, 1987; Redington, Crater, & Finch, 1998).

We explore the notion that learners use patterns of co-occurrence between words as a clue towards word meaning. In particular, learners might exploit the fact that words that occur in similar linguistic contexts tend to have similar meanings (as suggested by Firth, 1957, above). For example, one might infer that "postman" and "mailman" have similar meanings based solely on the fact that they tend to occur with words like "deliver", "package", and "truck". Note that this inference does not derive from *direct* co-occurrence between "postman" and "mailman" (indeed, we might expect these terms to only rarely co-occur), but rather from the similar *patterns* of co-occurrence with *other* words. We refer to such learning as contextual co-occurrence learning. This mechanism is not limited to learning word meanings – indeed, most research has focused on contextual co-occurrence learning of syntactic properties like grammatical category. Throughout this work, however, we take contextual co-occurrence learning to be concerned with meaningunless otherwise specified.

~~While~~ contextual co-occurrence learning is often thought to be important for language learning (TODO: cites), a number of experiments, conducted mainly on learning *syntactic category* from contextual co-occurrence, suggest that human learners' capacities are limited (TODO: cites). Our current study addresses this mismatch.

~~In this paper,~~ we here characterize some of the conditions under which people may successfully perform contextual co-occurrence learning of meaning. Human experiments typically expose learners to artificial languages where all words are *novel*. Our experiments suggest that contextual co-occurrence learning is facilitated by semantic coherence, the presence of *known* words adhering to some semantic organization. To preview our paper: first, we briefly review some of the general evidence for contextual co-occurrence learning. We then introduce the specific language structure we explore in this paper and discuss some of successes and the failures in finding evidence of contextual co-occurrence learning in experiments that use this language. In Experiment 1, we present evidence that semantic coherence can facilitate contextual co-occurrence. In Experiments 2 and 3, we further explore this effect by isolating each component of semantic coherence – meaning and coherence – and find that neither meaning alone nor coherence alone is sufficient to facilitate MNPQ 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.

Evidence for contextual co-occurrence learning

Initial proposal about learnign from contextual co-occurrence came from philosophical and linguistic research on the nature of meaning. In philosophy, Wittgenstein (1953/1997) objected to the idea that words have precise, formal definitions and argued that meaning derives from usage, i.e., patterns of usage in language. To illustrate, he considered the word “game”, a term used to describe activies as varied as board games, card games, ball games, Olympic games, 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 tidy circle around which we could circumscribe a boundary, but rather a “complicated network of similarities overlapping and crisscrossing” (§66). Wittgenstein argued that mapping this network by describing usage is critical to understanding meaning.

In linguistics, Firth (1957) similarly argued for a theory of meaning meanings based on patterns of “habitual collocation”. For example (p12), he asserts that part of the meaning of the “cow” is its co-occurrence with ~~(e.g.,)~~ “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” (p13)

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 (1951). Harris argued that, even in cases where word meaning was determined by extralinguistic influences, such influences would have distributional correlates.

	Document X	Document Y	...
Word A	Frequency of word A in document X	Frequency of word A in document Y	...
Word B	Frequency of word B in document X	Frequency of word B in document Y	...
...	...	...	...

Figure 1. Word-document matrix

Thus, meaning could be divined by quantitatively analyzing purely linguistic information.

These proposals about the distributional theory of meaning stimulated early empirical work ~~with humans~~. This ~~early~~ work typically demonstrated the validity of the distributional hypothesis using small samples of human judgments and corpora (Rubenstein & Goodenough, 1965; Clark, 1968; Stefflre & Reich, 1971; Geffroy et al., 1973; Berry-Rogghe, 1973; Szalay & Bryson, 1974).

In the 1980's computer scientists devised techniques that paved the way for larger scale investigations of contextual co-occurrence learning. Motivated by practical issues in the field of information retrieval, they considered the relationships between words and documents. A typical problem was that of retrieving relevant documents from a database in response to a query with certain search terms. 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 as rows, columns as documents, and particular cells encoding the frequency with which a particular word occurs in a particular document (~~see~~ Figure 1).

While such matrices can be interpreted as representing documents in terms of their

	Word X (-1)	Word X (+1)	Word X (+2)	...
Word A	Frequency of X 1 word before A	Frequency of X 1 word after A	Frequency of X 2 words after A	...
Word B	Frequency of X 1 word before B	Frequency of X 1 word after B	Frequency of X 2 words after B	...
...	...	...	...	...

Figure 2. Word-word matrix

constituent words, they can also be interpreted as representing words in terms of their patterns of use across documents, or, a proxy for the linguistic contexts that a word participates in. Put another way, such a matrix can be thought of either as a model of document meaning (i.e., a document's meaning is its column in this matrix) or, more importantly, for our purposes, a model of word meaning (i.e., a word's meaning is its row in this matrix).

Note that this word-document approach represents documents as "bags" of words – information about the relative position of words is discarded. The word-word approach retains this information (Church & Hanks, 1990; Schutze, 1992). The word-word matrix has words as both rows and columns; 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 (~~see~~ Figure 2).

~~In what proved to be a widely influential research program,~~ researchers found that both word-document and word-word models were capable of ~~robustly~~ learning semantic properties of words. As one example, Landauer and Dumais (1997) developed a model called Latent Semantic Analysis (LSA), which builds a word-document matrix from a corpus, applies a dimensionality reduction technique (singular value decomposition), and computes the similarity between words using a cosine measure. After training an a corpus of encyclopedia articles, LSA closely matched the performance of non-native English speakers on a synonym test.

As another example, Redington, Chater, and Finch (1998) developed a model that performs hierarchical clustering on a word-word matrix. This model was able to learn syntactic (as opposed to semantic) categories like noun, verb, and adjective. An interesting feature of the learned clusters is that they often had semantic organization. For example, in the cluster of adjectives, the color words and number words formed separate clusters (p448).

The success of early computational models led to a proliferation of models that learn from co-occurrence information (see Riordan & Jones, 2010 for an overview and comparison of state-of-the-art models). The computational evidence is quite strong: statistical patterns of co-occurrence can, in principle, be used to learn *inter alia* some aspects of word meaning.

We have computational proofs of concept, but do humans actually use co-occurrence information to learn meaning? Empirical research on knowledge of visual and color terms in the congenitally blind offers some indirect evidence. Shepard and Cooper (1992) presented congenitally blind, color-blind, and normally sighted participants with pairs of color words and asked them to make similarity ratings. As might be predicted, the ratings of blind participants did not correlate highly with normally sighted individuals. However, their ratings did appear to preserve the local similarity relationships – violet was rated most similar to purple, teal to green, and so forth. How might congenitally blind individuals have learned anything about color relationships? One possibility, raised by Shepard and Cooper, is that the blind may have extracted these relationships from their linguistic input. More recent work by Bedny et al. (2011) has shown similar results with knowledge of visual verbs.

These demonstrations in the blind, while suggestive, are only indirect evidence of contextual co-occurrence learning. Researchers studying learning of syntactic categories (as opposed to word meanings) have employed methods that in principle can 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 kind of co-occurrence structure, known as 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). Sentences that participants hear take one of two forms, MN and PQ. Thus, sentences like m_1n_3 are grammatical while sentences like m_1q_3 are illegal. Early investigations (Braine, 1966; Smith, 1966) found that participants tend to endorse novel grammatical MN and PQ sentences as having come from the language they heard. 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; Frigo & McDonald, 1998; Kempe & Brooks, 2001; Gerken, Gomez & Wilson, 2005; Lany & Saffran, 2010; Frank & Gibson, 2011). These results are puzzling, given that computational models suggest that contextual co-occurrence learning is a powerful mechanism.

However, many of same these studies have also demonstrated that MNPQ learning is possible when co-occurrence information is partially or completely correlated with another cue.

For example, Braine (1987) found that successful MNPQ learning results 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 (though not all Ms referred to men) and all pictures of women were labeled by P words (though 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 may be useful when combined with other information sources have motivated research on how disparate information sources might be integrated to facilitate learning (e.g., Monaghan, Chater, & Christiansen, 2005; Johns & Jones, 2011).

~~It is worth reiterating that~~ nearly all empirical work has interpreted the results of human experiments with reference to learning grammatical category, rather than word meaning. 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 and found that 22-month old infants successfully learned MNPQ 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 result suggests that contextual co-occurrence learning of grammatical category and word meaning may be quite similar, at least for the MNPQ language.

In this paper, we add to the literature on learning word meaning by exploring a new information source: semantic coherence. To date, all studies have used the artificial language learning paradigm. 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 contextual co-occurrence learning?

To explore this possibility, we presented participants with an MNPQ language where sentences took the form “M and N” or “P and Q”. We hypothesized that contextual co-occurrence learning for N and Q words would be afforded, given a certain level of semantic coherence (specifically, a taxonomic coherence where M’s are animals and P’s were vehicles). 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 contextual co-occurrence learning. In Experiment 2, we compared semantic coherence to phonological coherence. In Experiment 3, we compared semantic coherence to a semantic baseline, which had known words that did not adhere to any obvious semantic organization.

Experiment 1: Semantic coherence

In all the experiments reported in this paper, we presented participants with auditory sentences from an MNPQ language. In different conditions, we varied properties of the M's (which co-occur with N's) and P's (which co-occur with Q's). We will call the M's and P's *context words*. We measured learning for the N's and Q's, which we will call the *target words*.

In Experiment 1, we parametrically varied two independent properties of the context words. First, we manipulated 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 manipulated the amount of exposure to the language, in order to measure the efficiency of learning with respect to exposure amount. After participants were exposed to the language, we tested them on three measures of MNPQ learning – sentence memory, similarity rating, and a referent assignment task.

Method

Participants. 678 Amazon Mechanical Turk (MTurk) workers. Using MTurk's worker qualifications, we limited participaton to workers located in the United States and with a previous HIT approval rate greater than or equal to 90%. We chose MTurk workers because the number of experimental conditions required a large number of participants. Work by Burhmester, Kwang, and Gosling (2011) and Crump, McDonald, 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" (~~see~~ Figure 3). Note that sentences literally included the word "and" in the middle. We generated the actual lexical items randomly for each participant. N’s and Q’s were always novel nonsense words and were drawn without replacement from the set {moke, thite, jiv, pif, dex, wug}. M’s and P’s could be either novel or familiar. Novel M’s were drawn from {feeb, bim, lup} and novel P’s were drawn from {zabe, vap, chuv}. Familiar M’s and P’s obeyed a taxonomic organization – familiar M’s were drawn from {hamster, cat, dog} and familiar P’s 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[1]. 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, goldfish, mouse, boat, van, train, motorcycle, plane, and bicycle.

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”). Participants were required to correctly type the word to continue. The experiment had four phases: exposure, similarity, memory, and referent

[1] ~~In particular~~, we programatically submitted all of the text sentences to the text-to-speech web service that powers Google Translate; when we initially performed the synthesis, this web service used one set of voices. The available voices have changed since we created our initial stimuli. See also Footnote 3.

Exposure sentences

$\underline{m_1\ n_1}$	$m_1\ n_2$	$m_1\ n_3$
$m_2\ n_1$	$\underline{m_2\ n_2}$	$m_2\ n_3$
$m_3\ n_1$	$m_3\ n_2$	$\underline{m_3\ n_3}$
$\underline{p_1\ q_1}$	$p_1\ q_2$	$p_1\ q_3$
$p_2\ q_1$	$\underline{p_2\ q_2}$	$p_2\ q_3$
$p_3\ q_1$	$p_3\ q_2$	$\underline{p_3\ q_3}$

Memory items

Sentence type	Example
Familiar	$m_1\ n_2$
Withheld	$m_1\ n_1$
Category violation	$m_1\ q_2$
Position violation	$m_1\ m_1$

Similarity items

Pair type	Example
Within-category	n_1, n_2
Cross-category	n_1, q_1

Figure 3. The MNPQ language and test items for memory and similarity. Underlined sentences were withheld from exposure.

assignment. Below, we detail these phases (for exposition, we have switched the order of memory and similarity).

Exposure Participants listened to sentences from the language. We withheld six sentences from exposure (see Figure 3), 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.

Memory Participants listened to sentences and judged on a 5 point scale how confident they were that they had previously 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.
- *Cross-category* sentences of the form MQ and PN.
- *Position-violation* sentences of the form MM, NN, PP, and QQ.

Sentences were presented in random order such that there were no repeated words between consecutive trials. In two catch trials[2], instead of a sentence from the MNPQ language, we played a non-repeatable audio instruction 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 cross-category sentences.

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. This resulted in within-category judgments (e.g., n_1 vs. n_2) and cross-category judgments (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

[2] 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.

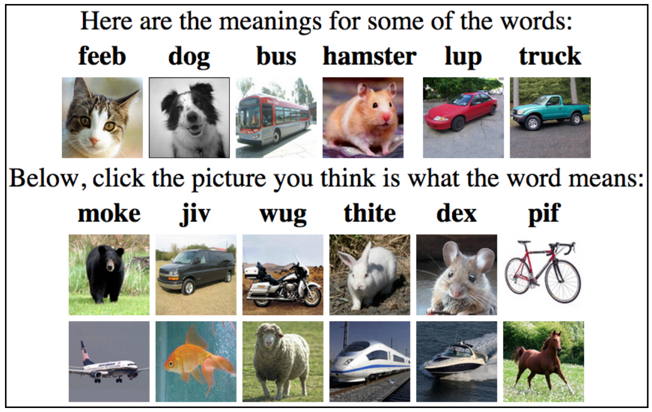

Figure 4. The referent assignment task.

word when clicked. In two catch trials, participants were asked 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 within-category pairs of words would be judged to be more similar than cross-category pairs.

Referent assignment Participants made 2AFC referent assignments for the N and Q words (~~see~~ Figure 4). At the top of the screen, we displayed the M and P words in random order. Under each word, we showed an image of an associated referent. The referents corresponded to the familiar pools for M and P words: CAT, DOG, HAMSTER, CAR, BUS, and TRUCK. Familiar words were always associated with the obvious referents (e.g., "dog" was always paired with an image of a dog). Below the "seeded" word meanings, we displayed a row containing the N and Q words. Under each word, we displayed a 2AFC 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 on the basis of the taxonomic categories of the co-occurring words (e.g., N's should be animals because they co-occur with M's, which are known to be animals).

To summarize, we devised three measures of learning: (1) memory for sentences, (2) similarity between target words, and (3) inductive bias in referent assignment.

Results and Discussion

We excluded the 55 participants who did not correctly answer all of the catch trials. Results are shown in Figure 5. For each dependent measure – memory, similarity, and meaning – we defined a within-participant score representing the sensitivity to the co-occurrence regularities in the language. Memory score was the difference in mean ratings between novel withheld sentences (e.g., m_1–n_1) and novel category violation sentences (e.g., m_1–q_1). Similarity score was the difference between mean ratings of within-category (e.g., N–N) and cross-category (e.g N–Q) ratings. Referent assignment score was the total number of correct choices in the referent assignment task. All scores were normalized to the interval [-1, 1].

Analysis approach. Using linear models, we analyzed two aspects of the data. First, we were interested in main effects of coherence on score (i.e., the Condition coefficients in Table 1). Second, we were interested in the relationship between amount of exposure and score. Accordingly, we looked for exposure $\times$ coherence interactions. A significant interaction (i.e., 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

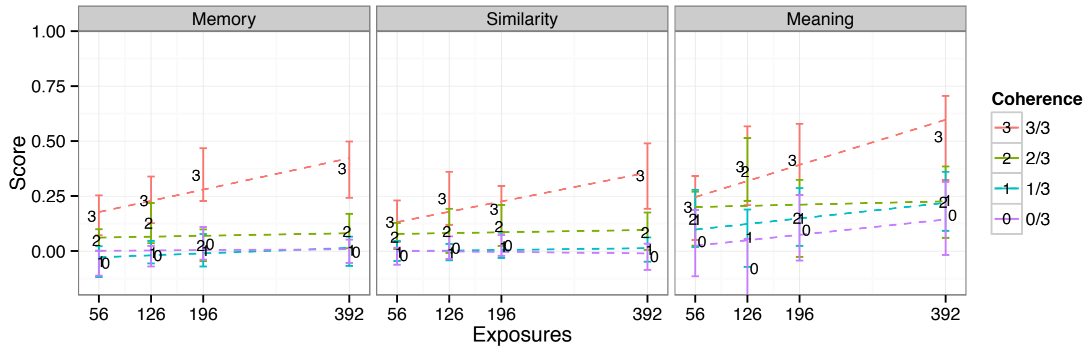

Figure 5. Experiment 1 results. Each plot shows data for one measure (memory, similarity, meaning) in Experiment 1. Points show condition means, error bars show 95% CIs, and dashed lines show the best-fitting linear regression.

a categorical variable and analyzed the data using a regression which modeled the mean score in an 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.

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. ~~Here,~~ we separately report performance on each of the three tasks.

Memory There was a significant main effect of exposure, with greater exposure resulted in better memory scores. There was also a significant main effect of condition; 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* increasing coherence results in *parametric* increases in memory score.

Additionally, there was a significant exposure $\times$ condition interaction; the effect of exposures on score was significantly higher in 3/3 than in the other conditions combined, 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 There was a significant main effect of condition: 3/3 scores were significantly higher than in the other conditions combined. Additionally, there was a significant exposure $\times$ condition interaction; the effect of exposures on score was significantly higher in 3/3 than in the other conditions combined. 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 process involved in making such distinctions, at least in the 3/3 condition.

Referent assignment There were significant main effects of exposure and condition. 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, $\beta = 0.03, t(615) = 1.72, p = 0.08$, 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.

Integrative results. Why does semantic coherence facilitate MNPQ learning? Frank & Gibson (2011) have shown that MNPQ learning can be bolstered by easing working memory demands. Additionally, there is evidence that novel words tax the memory system more, as they are encoded in terms of smaller phonological units (Treiman & Danis, 1988). So it is possible that semantic coherence improved MNPQ learning by reducing memory demands.

We tested for this possibility in our data using mediation analyses. In particular, we tested whether memory scores mediated the effect of coherence on either (1) similarity scores or (2) referent assignment scores. In both cases, we found partial mediation. After controlling for memory, the regression coefficient relating coherence and similarity decreased significantly from 0.07 to 0.03, Sobel z = 7.80, $p < 0.005$; this reduced value was significantly greater than zero, $t(620) = 3.50, p < 0.001$, indicating partial mediation. After controlling for memory, the regression coefficient relating coherence and referent assignment score decreased significantly from 0.10 to 0.05, Sobel z = 5.33, $p < 0.005$; this reduced value was significantly greater than zero, $t(616) = 3.15, p < 0.005$, again indicating partial mediation. Thus, improved memory can explain some, but not all, of the increase in similarity and referent assignment scores due to semantic coherence.

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 words with objects in the referent assignment task. Using mediation analysis, we found that evidence that semantic coherence boosts learning in part because it eases memory demands.

Semantic coherence has two components – meaning and coherence. How does the effect of semantic coherence depend on each? In Experiments 2 and 3, we test for the effect of meaning and coherence respectively. In Experiment 2, we remove 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 (M's and P's) exhibited phonological, rather than semantic, coherence. We tested three types of coherence: onset, rime, and syllable count.

Method

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

Materials. The three types of phonological coherence[3] were:

- *Onset*. M's all started with one consonant cluster (pladge, plaaf, plab) and P's all started with another (zof, zawd, zawsh).
- *Rime*. M's all ended with one vowel (calo, pawmo, marfo) and P's all ended with another (zaygee, kaisee, tetchee).
- *Syllable count*. M's were disyllabic (coomo, fengle, kaisee) and P's were monosyllabic (gope, jic, skeege).

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

Results and Discussion

We discarded the 42 participants who did not pass all the catch trials. Results are graphed in Figure 6. Using an 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 (~~see~~ Table 2).

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

[3] 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 1) 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.

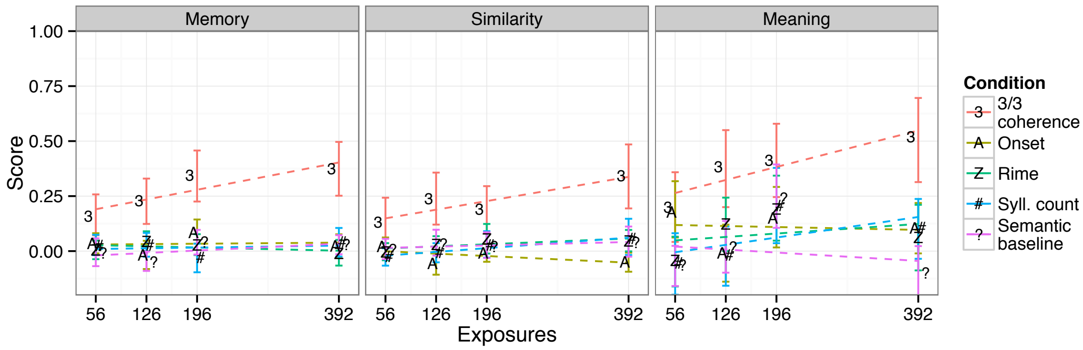

Figure 6. Experiments 2 and 3 results. Each plot shows data for one measure (memory, similarity, meaning). Points show condition means, error bars show 95% CIs, and dashed lines show the best-fitting linear trend. For comparison, we also included the 3/3 coherence condition from Experiment 1.

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.

Across the three models, there were no significant predictors, save the one interaction term for syllable count versus 0/3 on the similarity measure, which can be plausibly attributed to chance. This suggests that phonological coherence was virtually indistinguishable from the 0/3 condition in terms of facilitating MNPQ learning. This indicates that mere coherence is not what drives the effects of semantic coherence. In Experiment 3, we consider whether the mere presence of known words (semantic baseline) aids MNPQ learning.

Experiment 3: Semantic baseline

In Experiment 1, M's and P's were all familiar words obeying a taxonomic organization. In Experiment 3, we explored whether mere coherence is sufficient for facilitation of contextual co-occurrence learning, or whether the mere presence of known words is sufficient – that is, whether a semantically baseline language facilitates contextual co-occurrence 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.

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).

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. Results are graphed in Figure 6. 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

How people learn word meanings remains a central problem in cognitive science. ~~It is clear, however, that people draw on various information sources during learning.~~ In this paper, we have explored whether people might learn meanings from language itself, using patterns of co-occurrence as a clue towards word meaning. Research using computational models suggests that contextual co-occurrence is a powerful source of information, yet human experiments suggest that contextual co-occurrence learning does not succeed without correlated information sources.

We contributed to the literature on human experiments that seeks to understand the conditions under which people can successfully perform contextual co-occurrence learning. For the MNPQ language, previous work has found successful learning in cases where co-occurrence cues are correlated with natural gender (Braine, 1987) or phonological cues (e.g., Lany & Saffran, 2010). Learning experiments in this literature have typically presented participants with linguistic input that is entirely novel. This is not representative of the conditions that most real language learners confront. Instead, real learners typically know the meanings of some of the words, which may have some amount of semantic organization, a factor we call semantic coherence. Our experiments provide evidence that semantic coherence facilitates MNPQ learning.

In Experiment 1, we showed that semantic coherence facilitates MNPQ learning; 3/3 semantic coherence resulted in better memory for the co-occurrence structure of the language and sharper inductive bias in similarity and meaning judgments. Additionally, for the memory and similarity measures, we found evidence of greater statistical efficiency – using regression, we found that the 3/3 rate of learning was higher than the other conditions. In Experiments 2 and 3, we investigated whether the effect of semantic coherence was driven by either meaning or coherence alone. In Experiment 2, we kept coherence but removed meaning by testing languages with phonological coherence. These languages did not confer any learning benefits, indicating that coherence alone does not drive the effect of semantic coherence. In Experiment 3, we kept meaning but removed coherence by testing a baseline language with known words but no obvious semantic organization. This language did not confer any learning benefits, indicating that meaning alone does not drive the effect of semantic coherence. It appears that the interaction of both meaning and coherence drives the effect; when we separately removed meaning (as in the phonological conditions) and coherence (as in the semantic baseline condition), learners failed to learn co-occurrence regularities in the language.

Our experiments highlight a limitation of artificial language learning. Researchers using entirely artificial languages may be severely limiting the power of contextual co-occurrence learning, which our experiments show to be greatly enhanced by the presence of known words that adhere to some semantic organization. In our analyses, memory was a significant (albeit partial) mediator of the effect of semantic coherence on similarity and referent assignment scores. Thus, artificial languages may place too high a memory burden on learners. Also, to the extent that semantic coherence works through factors other than memory, entirely artificial languages may deprive these pathways as well. To better match experimental settings with conditions that real language learners find themselves in, we argue for more research using “semi-artificial” languages. Alternately, it may be possible to mimic semantic organization in entirely artificial languages by seeding sentences with nonce topics (e.g., "the sentence you are about to hear is about *chylu*").

Our empirical results can serve as a useful testbed for computational models. Surprisingly, work on contextual co-occurrence learning of semantics (and syntax) is generally either empirical or computational – the two are rarely combined (cf. Qian et al., 2012). In computational work, researchers typically train models on large corpora, rather than the artificial languages for which there is more detailed human learning data. In the future, researchers should investigate model behavior for these smaller "corpora" derived from artificial languages. These are simpler to collect and can serve as stronger constraints on theorizing, because (e.g.,) detailed human judgments about sentence familiarity are available.

Using mediation analysis, we found evidence that reduced memory load appears to drive some, but not all, of the effect of semantic coherence. What other factors are in play? One possibility is that learners use semantic coherence to infer the topic of discourse. Then, learners attach meaning to novel words on the basis of co-occurrences with these topics. For example, in the 3/3 condition of our experiments, participants may have learned that the topic of discourse is either animals or vehicles and then tracked co-occurrence between these taxonomic topics and novel words.

It may be through such a process that we come to acquire meanings for abstract words (e.g., *system*) or inchoate meanings for concrete words (e.g., people may know that *brigade* is a military term but they tend not to know its precise meaning). If such a mechanism were at work, it might also predict that word learning would have a "contiguous" character, with faster learning for words that occur in more coherent contexts. Indeed, the original LSA work by Landauer & Dumais (1997) anticipated this possibility – learning was faster when the model already "knew" many words compared to when it knew very few (p229).

So what role does contextual co-occurrence learning play in word learning? The earliest empirical work on the MNPQ language suggested a quite limited role. Later research on correlated cues suggested that learning might be more effective when combined with other cues. Our research highlights the effectiveness of a new cue: semantic coherence. ~~We conclude by posing some questions and potential approaches for future research:~~

- In our experiments, we manipulated semantic coherence by manipulating the fraction of sentences in which target words occurred with known coherent words. However, our sentences were simple and short ("[context] and [target]"). How can we formalize semantic coherence for more realistic sentences, which are longer and may contain mixtures of known and unknown words? We suggest that semantic coherence might be formalized within a probabilistic inference framework as the certainty about the topic of discourse (or perhaps the speaker's meaning).
- It seems likely that the earliest words acquired by learners, concrete nouns, are acquired through mechanisms that exploit visual access to referents and social cues. Thus, contextual co-occurrence learning may be better suited to learning abstract words or concrete nouns for which referents are not available (e.g., *poi* and *breadfruit*). Do adult speakers' representations of such words reflect co-occurrence statistics? Can we manipulate such representations experimentally?
- To what extent is real linguistic input semantically coherent? Does this change through development? Crucially, is this level of coherence sufficient to support acquisition solely by contextual co-occurrence learning?

References

Baldwin, D. (1993). Early referential understanding: Infants' ability to recognize referential acts for what they are. *Developmental Psychology*, *29*, 832–843.

Braine, M. (1987). What is learned in acquiring word classes: A step toward an acquisition theory. *Mechanisms of language acquisition*, 65–87.

Carey, S. (1978). The child as word learner. In *Linguistic theory and psychological reality* (pp. 264–293). Cambridge, MA: MIT Press.

Clark, E. (1988). On the logic of contrast. *Journal of Child Language*, *15*, 317–335.

Firth, J. R. (1957). *A synopsis of linguistic theory, 1930-1955*.

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

Hollich, G., Hirsh-Pasek, K., & Golinkoff, R. (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*.

Maratsos, M., & Chalkley, M. (1980). The internal language of children's syntax: The ontogenesis and representation of syntactic categories. *Children's language*, *2*, 127–214.

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

Quine, W. V. O. (1960). *Word and object*. Cambridge, MA: MIT press.

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

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

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

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

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

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

Table 1

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

Regressor	β	Std Error	t	p
		Memory		
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		
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		
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				
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 $\times$ C: Onset – 0/3	-0.0948	0.170	-0.55	0.577
E $\times$ C: Rime – 0/3	-0.1950	0.161	-1.20	0.228
E $\times$ C: Syllable count – 0/3	-0.0735	0.167	-0.43	0.660
E $\times$ C: Semantic incoherent – 0/3	0.0388	0.162	0.23	0.811
Similarity				
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 $\times$ C: Onset – 0/3	-0.0937	0.157	-0.59	0.550
E $\times$ C: Rime – 0/3	0.1951	0.149	1.30	0.191
E $\times$ C: Syllable count – 0/3	0.3049	0.154	1.97	¡0.05*
E $\times$ C: Semantic incoherent – 0/3	0.1401	0.150	0.93	0.350
Referent assignment				
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 $\times$ C: Onset – 0/3	-0.6122	0.421	-1.45	0.146
E $\times$ C: Rime – 0/3	-0.3243	0.400	-0.80	0.418
E $\times$ C: Syllable count – 0/3	-0.0692	0.414	-0.16	0.867
E $\times$ C: Semantic incoherent – 0/3	-0.7403	0.402	-1.83	0.066