Semantic coherence facilitates distributional learning

Long Ouyang, Lera Boroditsky, Michael C. Frank

Stanford University

Author Note

Corresponding author: Long Ouyang (longouyang@post.harvard.edu)

Abstract

Abstract here.

Semantic coherence facilitates distributional learning

Learning the meanings of words is a hard problem. Fortunately, for learners, this hard problem is softened by a multiplicity of sources of information about word meaning. The external world, other people, and language itself provide useful information to the learner. Specific examples of informative sources include vision (Kay & McDaniel, 1978; Rescorla, 1980; Booth & Waxman, 2002; Regier & Zheng, 2007), social experience (Baldwin, 1991; Bloom, 2002; Tomasello, 2003), phonology (Bloomfield, 1895; Kohler, 1929; Ramachandran and Hubbard, 2001; Maurer, Pathman, & Mondloch, 2006; Graf-Estes et al., 2007; Parault & Parkinson, 2008; Shukla, White, & Aslin, 2011), syntax (Gleitman, 1990; Fisher et al., 1994), morphology (Carr & Johnston, 2001), and linguistic context. Our chief interest in this paper is this last source – linguistic context.

Researchers believe that people may use knowledge of linguistic contexts as a clue toward meaning. For example, one might infer that "postman" and "mailman" have similar meanings based solely on the fact that they occur in similar linguistic contexts (e.g., they both occur in sentences about delivering packages and driving mail trucks). Learning from linguistic context in this fashion is called *distributional learning*. While researchers believe that distributional learning is a powerful learning mechanism, human experiments using artificial language learning suggest that people may have limitations on their ability to do distributional learning.

In this paper, we help characterize some of the conditions under which people may successfully perform distributional learning. Human experiments typically expose learners to artificial languages where all words are *novel*. Our experiments suggest that distributional learning of meaning 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 distributional learning. We then introduce the specific language structure we explore in this paper, the MNPQ language, and discuss some of successes and the failures in finding evidence of distributional learning in experiments that use this language. In Experiment 1, we present

evidence that semantic coherence can facilitate MNPQ learning. 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 semantic coherence effect.

Evidence for distributional learning

The intellectual roots of distributional learning do not, in fact, directly deal with *learning*. The scientific investigation of learning drew inspiration from distributional theories of *meaning* from philosophy and linguistics. 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" (Firth, 1957). 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[1]. 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) elicited synonym judgments for pairs of concrete nouns and compared these judgments against co-occurrence statistics in a corpus generated by an independent group of subjects. They found a positive relationship between synonymy and degree of linguistic context match. Clark (1968) found a correlation between the degree of substibutility and the degree of context overlap for prepositions. Szalay & Bryson (1974) found correlations between word substitution ratings and judgments of similarity and grouping for both concrete and abstract nouns.

TODO: Geffroy et al.; Berry-Rogghe; Jones & Sinclair

(cf. syntagmatic relations) have respectable correlations (0.4-0.6) with judgments of similarity and grouping and some other measures. studied concrete and abstract nouns (food, hunger, rice; money, poverty, beggar; school, knowledge, education; greeting, politeness, manners)

M. (1973). Lexicometric Analysis of Cooccurrences. In A. J. Aitken, R. W. Bailey, & N.

---

[1]"... it frequently happens that when we do not rest with the explanation that something is due to meaning, we discover that it has a formal regularity or 'explanation.' It may still be 'due to meaning' in one sense, but it accords with a distributional regularity." (1970, p.785)

Hamilton-Smith (Eds.), The Computer and Literary Studies. Edinburgh: Edinburgh University Press.al., 1973 V., & Reich, P. (1971). Some eliciting and for descriptive semantics. In P Kay (Ed.), mathematical anthropology. Cambridge, Mass. MIT Press. Berry-Rogghe, G. (1973). The computation of collocations and relevance in lexical studies. In A. J. Aitken, R. W. Bailey, & N. Hamilton-Smith (Eds.), The Computer and Literary Studies (pp. 103-112). Edinburgh: Edinburgh University Press. & Sinclair, J. (1974). English lexical collocations: computational linguistics. Cahiers de Lexicologie, 24(1),

classification: RCF98, Cartwright + brent, 1996, mintz et al., 2002, topic models. (bottom line: computational proof of concept)

This early empirical work was at a limited scale – human subjects gave judgments for a small set of words and computational analyses considered small corpora. In the 1980's, computer scientists devised techniques that paved the way for larger 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 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 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 –

	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

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. An interesting feature of the learned clusters is that they often had semantic organization – in the cluster of adjectives, the color words and number words formed separate clusters (p448). We bring up this syntactic example to demonstrate that distributional learning has been posited as a learning mechanism in a variety of domains, from word segmentation (e.g., Meylan et al., 2012) to visual feature learning (e.g., Austerweil & Griffiths, 2011), to visuomotor category formation (e.g., Hunt & Aslin, 2010).

	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

The success of early computational models led to a proliferation of models that learn from distributional 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 distributional 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. Similarly, Bedny et al. (2011) presented sighted and congenitally blind subjects with pairs of visual verbs and asked them to make similarity judgments. Judgments between the two groups were quite similar, raising the possibility that blind subjects learned verb meanings from linguistic information (however, Bedny et al. argue that judgments are not based on word co-occurrence, as blind and sighted ratings were more

correlated with each other than with ratings derived from LSA).

These demonstrations in the blind, while suggestive, are only indirect evidence of distributional learning. Researchers studying distributional 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 distributional 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 distributional structure known as the MNPQ language.

A puzzle: the MNPQ language

The MNPQ language contains four categories of words, M, N, P, and Q, and sentences that subjects hear take one of two forms, MN and PQ. Early investigations (Braine, 1966; Smith, 1966) found that subjects tend to endorse novel but grammatical MN and PQ sentences as having come from the language they heard. However, subjects 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 (either syntactic or semantic) 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 strongly indicate that information sufficient for categorization exists in the linguistic input.

However, many of same these studies have also demonstrated that MNPQ learning is possible when distributional information is partially or completely correlated with another cue. For example, Braine (1987) found that successful MNPQ learning results when distributional information is partially correlated with natural gender. In this experiment, subjects 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 distributional 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*.

Recently, this suggestion received attention in the work of Lany and Saffran (2010), who found that 22-month old infants successfully learned MNPQ when distributional regularities were 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. These demonstrations that distributional information may be useful to the learner when combined with other information sources has motivated research on how disparate information sources might be integrated to facilitate learning (e.g., Monaghan, Chater, & Christiansen, 2005; Johns & Jones, 2011).

In this paper, we add to this literature by exploring a new information source: semantic coherence. All of the studies referenced above were conducted using 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 distributional learning?

To explore this possibility, we presented subjects with an MNPQ language where sentences took the form "M and N" or "P and Q". We hypothesized that distributional 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 distributional learning. In Experiment 2, we compared semantic coherence to phonological coherence. In Experiment 3, we compared semantic coherence to semantic incoherence, the presence of known words that do not adhere to some semantic organization.

Experiment 1: Semantic coherence

In all the experiments reported in this paper, we presented subjects 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.

After subjects were exposed to the language, we tested them on three measures of MNPQ learning – sentence memory, similarity rating, and a referent assignment task.

Method

Subjects.   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 subjects. 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 subject. 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[2]. 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 subject 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 – subjects heard either 56, 126, 196, or 392 sentences. Before starting the experiment, we asked subjects to turn on their speakers and click a button, which played a spoken English word ("airplane"). Subjects were required to correctly type the word to continue.

---

[2]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 4.

$$
\boxed{
\begin{array}{ccc|ccc}
\underline{m_1\ n_1} & m_1\ n_2 & m_1\ n_3 & \underline{p_1\ q_1} & p_1\ q_2 & p_1\ q_3 \\[6pt]
m_2\ n_1 & \underline{m_2\ n_2} & m_2\ n_3 & p_2\ q_1 & \underline{p_2\ q_2} & p_2\ q_3 \\[6pt]
m_3\ n_1 & m_3\ n_2 & \underline{m_3\ n_3} & p_3\ q_1 & p_3\ q_2 & \underline{p_3\ q_3}
\end{array}
}
$$

Figure 3. The MNPQ language. Underlined sentences were withheld from exposure.

The experiment had four phases – exposure, similarity, memory, and referent assignment. Below, we detail these phases (for exposition, we have switched the order of memory and similarity).

Exposure. Subjects 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, subjects had to click a button to hear each sentence.

Memory. Subjects 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[3], instead of a sentence from the MNPQ language, we played

---

[3]In early runs of the experiment, 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 subjects did not receive catch trials. In [3/3;56], 30 out of 43 subjects did not receive catch trials. In [3/3;126], 30 out of 40 subjects did not receive catch trials. In [3/3;196], 30 out of 40 subjects did not receive catch
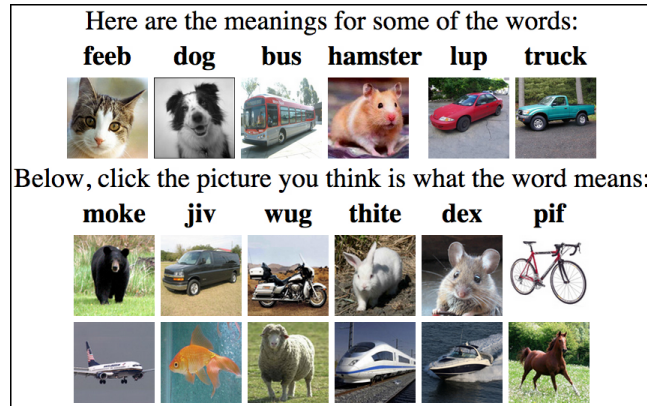
Figure 4. The referent assignment task.

a non-repeatable audio instruction to press a specific response button. If subjects 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 subjects 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** ◀ᵈ and **thite** ◀ᵈ ?"); 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, subjects were asked to press the response button corresponding to the solution of a simple arithmetic problem. If subjects 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. Subjects 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

trials.

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); subjects made a choice by clicking on one of the two pictures. If subjects 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 subjects 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-subject 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.   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 × coherence interactions. A significant interaction (i.e., the E × 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 an subject 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 ~ exposures × 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 learning (i.e., sensitivity to the amount of exposure) in the 3/3 condition.
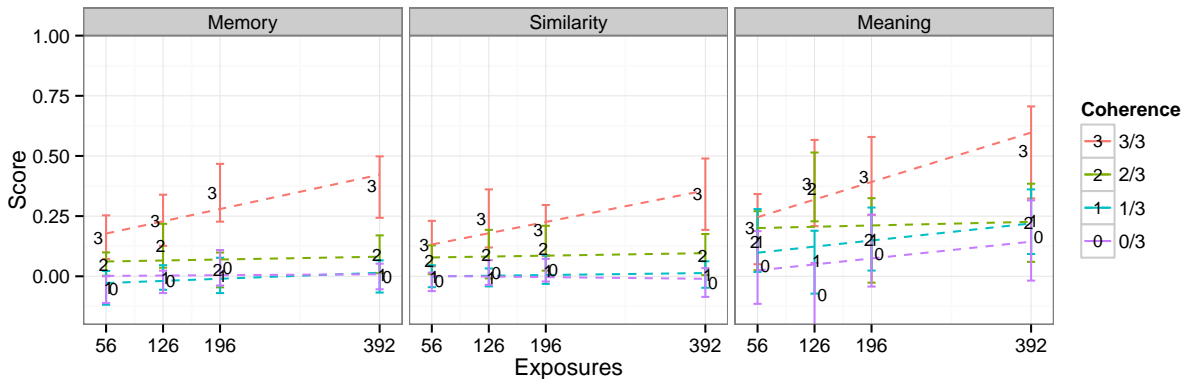

Figure 5. Experiment 1 results. Each plot shows data for one measure (memory, similarity, meaning) in Experiment 1. Points show condition means, error bars show 95% CIs, and dashed lines show the best-fitting linear regression.

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.

0.00422 **

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 do successful MNPQ learning 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

Subjects.   530 MTurk workers participated in the study.

Materials.   The three types of phonological coherence[4] 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.

---

[4] 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 2) 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.
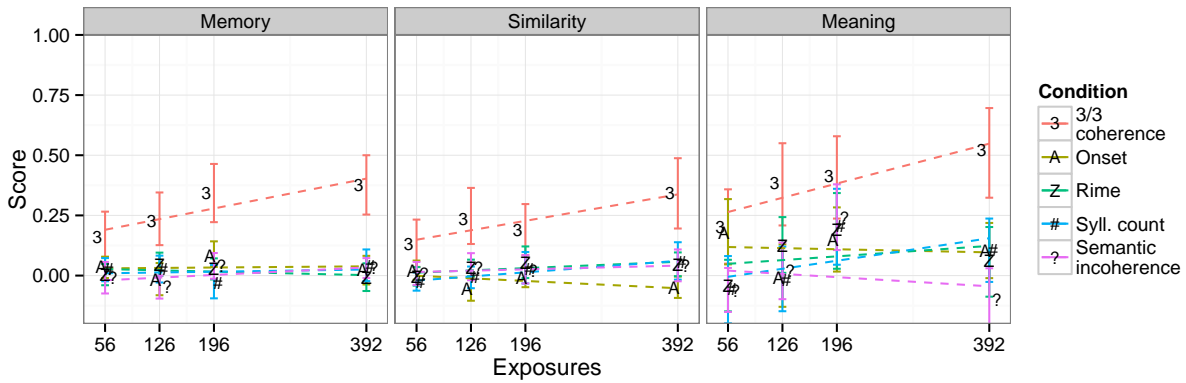
Figure 6. Experiments 2 and 3 results. Each plot shows data for one measure (memory, similarity, meaning). Points show condition means, error bars show 95% CIs, and dashed lines show the best-fitting linear trend. For comparison, we also included the 3/3 coherence condition from Experiment 1.

Results and Discussion

We discarded the 42 subjects who did not pass all the catch trials. Results are graphed in Figure 6. Using an interactive regression model (score $\sim$ exposures $\times$ condition), 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.

Similarity. 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. 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. 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.

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 incoherence) aids MNPQ learning.

Experiment 3: Semantic incoherence

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 distributional learning, or whether the mere presence of known words is sufficient – that is, whether a semantically *incoherent* language facilitates distributional learning.

Methods

Subjects.   162 MTurk workers.

Materials.   In the semantically incoherent language, the specific M and P words were drawn randomly for each subject 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 subjects who did not pass all the catch trials. Results are graphed in Figure 6. See Table 2 for regression results.

Memory.   Incoherent scores were not significantly different from 0/3 scores and incoherent efficiency was not significantly different from 0/3 efficiency.

Similarity.   Incoherent scores were not significantly different from 0/3 scores and incoherent efficiency was not significantly different from 0/3 efficiency.

Referent assignment.   Incoherent scores were not significantly different from 0/3 scores and incoherent efficiency was not significantly different from 0/3 efficiency.

Apparently, the familiar but semantically incoherent linguistic 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

In this paper, we contributed to the literature on human experiments that seeks to understand the conditions under which people can successfully perform distributional learning. For the MNPQ language, previous work has found successful learning in cases where distributional cues are correlated with natural gender (Braine, 1987) or phonological cues (e.g., Lany & Saffran, 2010). Learning experiments in this literature have typically presented subjects 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. In our experiments, we have found 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, sharper inductive bias in similarity and meaning judgments. Additionally, for the memory and similarity measures, we found evidence of greater statistical efficiency – adding sentences for 3/3 subjects resulted in a larger performance gain. 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 incoherent 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 incoherence condition), learners failed to learn distributional regularities in the language.

Implications for empirical and computational research

Our experiments highlight a limitation of artificial language learning. Researchers using entirely artificial languages may be severely limiting the power of distributional learning mechanisms, 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*").

Furthermore, our results add to a body of experimental work that can serve as a useful testbed for computational models. Surprisingly, work on distributional learning of semantics (and syntax) is generally either empirical or computational – the two are rarely combined (but see Tian et al., 2012 for a recent and welcome exception). 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 other domains (e.g., the word segmentation work of Meylan et al., 2012), analyzing model behavior on the experimental stimuli that people receive has proven useful, raising constraints on the kinds of representations and processes that could account for behavioral patterns; we believe that the domain of semantics could benefit from this approach as well.

Mechanism: memory and what else?

Using mediation analysis, we found evidence that effect of semantic coherence is partially one of reducing memory demands. However, reduced memory load doesn't appear to drive the entire effect – what else is happening? One possibility is that learners use semantic coherence to infer the topic of discourse (cf. the topic learning models of Griffiths & Steyvers, 2007). Then, learners attach meaning to novel words on the basis of co-occurrences with these topics. It may be through such a process that we come to acquire inchoate meanings for 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).

Table 1

Regression model for Experiment 1

Regressor	β	Std. Error	t	p
Memory				
Intercept	0.03593	0.01876	1.91	0.056
Condition: 1/3 – (0/3)	-0.00053	0.02742	-0.01	0.984
Condition: 2/3 – (0/3,1/3)	0.03094	0.01478	2.09	<0.05*
Condition: 3/3 – (0/3,1/3,2/3)	0.03963	0.01084	3.65	<0.001*
Exposures	0.00023	0.00008	2.81	<0.005*
E × C: 1/3 – (0/3)	0.00000	0.00012	0.02	0.977
E × C: 2/3 – (0/3,1/3)	-0.00002	0.00006	-0.32	0.743
E × C: 3/3 – (0/3,1/3,2/3)	0.00013	0.00004	2.73	<0.01*
Similarity				
Intercept	0.05018	0.01895	2.64	<0.01*
Condition: 1/3 – (0/3)	-0.00707	0.02768	-0.25	0.798
Condition: 2/3 – (0/3,1/3)	0.02374	0.01493	1.59	0.112
Condition: 3/3 – (0/3,1/3,2/3)	0.02244	0.01095	2.04	<0.05*
Exposures	0.00014	0.00008	1.77	0.077
E × C: 1/3 – (0/3)	0.00005	0.00012	0.42	0.671
E × C: 2/3 – (0/3,1/3)	0.00002	0.00006	0.31	0.754
E × C: 3/3 – (0/3,1/3,2/3)	0.00013	0.00004	2.80	<0.01*
Referent assignment				
Intercept	0.10876	0.03606	3.01	<0.005*
Condition: 1/3 – (0/3)	0.06647	0.05275	1.26	0.208
Condition: 2/3 – (0/3,1/3)	0.06163	0.02838	2.17	<0.05*
Condition: 3/3 – (0/3,1/3,2/3)	0.03598	0.02084	1.72	0.084
Exposures	0.00045	0.00015	2.86	<0.005*
E × C: 1/3 – (0/3)	-0.00009	0.00023	-0.39	0.690
E × C: 2/3 – (0/3,1/3)	-0.00012	0.00012	-0.99	0.319
E × C: 3/3 – (0/3,1/3,2/3)	0.00012	0.00009	1.38	0.167

Table 2

Regression model for Experiments 2 and 3

Regressor	β	Std. Error	t	p
Memory				
Intercept	-0.03411	0.02749	-1.24	0.215
Condition: Onset – 0/3	0.06266	0.03806	1.64	0.100
Condition: Rime – 0/3	0.06528	0.03574	1.82	0.068
Condition: Syllable count – 0/3	0.04102	0.03856	1.06	0.287
Condition: Semantic incoherent – 0/3	0.00559	0.03817	0.14	0.883
Exposures	0.00012	0.00011	1.00	0.316
E × C: Onset – 0/3	-0.00009	0.00017	-0.55	0.577
E × C: Rime – 0/3	-0.00019	0.00016	-1.20	0.228
E × C: Syllable count – 0/3	-0.00007	0.00016	-0.43	0.660
E × C: Semantic incoherent – 0/3	0.00003	0.00016	0.23	0.811
Similarity				
Intercept	0.01107	0.02539	0.43	0.662
Condition: Onset – 0/3	-0.00316	0.03515	-0.08	0.928
Condition: Rime – 0/3	-0.00640	0.03301	-0.19	0.846
Condition: Syllable count – 0/3	-0.04553	0.03561	-1.27	0.201
Condition: Semantic incoherent – 0/3	-0.00078	0.03525	-0.02	0.982
Exposures	-0.00006	0.00011	-0.54	0.582
E × C: Onset – 0/3	-0.00009	0.00015	-0.59	0.550
E × C: Rime – 0/3	0.00019	0.00014	1.30	0.191
E × C: Syllable count – 0/3	0.00030	0.00015	1.97	<0.05*
E × C: Semantic incoherent – 0/3	0.00014	0.00015	0.93	0.350
Referent assignment				
Intercept	-0.05533	0.06831	-0.81	0.418
Condition: Onset – 0/3	0.17766	0.09442	1.88	0.060
Condition: Rime – 0/3	0.09165	0.08868	1.03	0.301
Condition: Syllable count – 0/3	0.02348	0.09581	0.24	0.806
Condition: Semantic incoherent – 0/3	0.08683	0.09478	0.91	0.359
Exposures	0.00054	0.00029	1.84	0.066
E × C: Onset – 0/3	-0.00061	0.00042	-1.45	0.146
E × C: Rime – 0/3	-0.00032	0.00040	-0.80	0.418
E × C: Syllable count – 0/3	-0.00006	0.00041	-0.16	0.867
E × C: Semantic incoherent – 0/3	-0.00074	0.00040	-1.83	0.066

References

Foss, D. J., & Jenkins, J. J. (1966). Mediated stimulus equivalence as a function of the number of converging stimulus items. *Journal of Experimental Psychology*.

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*.

Imai, M., Kita, S., Nagumo, M., & Okada, H. (2008). Sound symbolism facilitates early verb learning. *Cognition*.