Semantic coherence facilitates distributional learning

Long Ouyang

Department of Psychology, Stanford University

Lera Boroditsky

Department of Cognitive Science, University of California San Diego

Michael C. Frank

Department of Psychology, Stanford University

Word count: 8,365

Author Note

Abstract

Computational models have shown that purely statistical knowledge about words' linguistic contexts is sufficient to learn many properties of words, including syntactic and semantic category. For example, models can infer that "postman" and "mailman" are semantically similar because they have quantitatively similar patterns of association with *other* words (e.g., they both tend to occur with words like "deliver", "truck", "package"). Contra these computational results, artificial language learning experiments suggest that distributional statistics *alone* do not facilitate learning of linguistic categories. However, experiments in this paradigm expose participants to entirely novel words, whereas real language learners encounter input that contains some known words that are semantically organized. In three experiments, we show that (1) the presence of familiar semantic reference points facilitates distributional learning and (2) this effect crucially depends both on the presence of known words and the adherence of these known words to some semantic organization.

Keywords: distributional learning; word learning; semantic coherence

Semantic coherence facilitates distributional learning

1 Introduction

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

How do people learn language? Learners use many information sources, including physical, social, conceptual, and linguistic cues (Baldwin, 1993; E. V. Clark, 1988; Gleitman, 1990; Hollich, Hirsh-Pasek, & Golinkoff, 2000; Markman, 1991). But in addition to external sources of information, the distributional properties of language itself can be informative about both its structure and meaning. A wide variety of experiments show that learners are sensitive to these distributional properties. For example, learners can group sounds together into word forms based on their statistical co-occurrence (Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996) and pair word forms with their referents based on consistent associations (L. Smith & Yu, 2008; Yu & Ballard, 2007). In the current paper, we explore a more sophisticated type of distributional learning: learning about a word using evidence about the linguistic contexts that it occurs (Braine, 1987; Maratsos & Chalkley, 1980; Redington, Chater, & Finch, 1998; K. H. Smith, 1966).

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

Distributional learning is a general mechanism thought to be important throughout language learning, including acquisition of word meaning (Landauer & Dumais, 1997) and grammatical category (Redington et al., 1998). Nevertheless, a number of experiments, conducted mainly on acquisition of grammatical category, suggest that human learners' capacities are limited in this regard (Braine, 1987; Brooks, Braine, Catalano, Brody, & Sudhalter, 1993; Frank & Gibson, 2011; Frigo & McDonald, 1998; Gerken, Wilson, & Lewis, 2005; Kempe & Brooks, 2001). Our study addresses this mismatch by attempting to isolate conditions under which people succeed in distributional learning.

We focus here on the effects of existing linguistic knowledge on distributional learning. Prior experiments typically exposed learners to artificial languages composed of entirely novel words. Our experiments, which we view as part of of a recent trend to address mismatches between artificial language learning studies and natural language, suggest that *semantic coherence*, the presence of known words adhering to some semantic organization, facilitates distributional learning. Put another way, a small semantic hook allows people to better leverage distributional information for learning. Phrased in terms of the Alice–Bob example, we stand a better chance of learning about Alice from her associates if those associates play roles that are already meaningful to us. Knowing that Alice associates with professors and college sophomores gives us a clearer picture of her than, say, knowing that she associates with taphonomists and bryologists.[1] Learning a word from its company is easier if this company is already familiar.

We begin by briefly reviewing some of the computational evidence for distributional learning. We then introduce the specific language structure, MNPQ, that we explore in this paper. Next, we discuss research that uses the MNPQ structure to study distributional learning of grammatical categories. In Experiments 1a and 1b, we present evidence that semantic coherence can facilitate distributional learning and that learners make use of acquired distributional information when performing an online referent assignment task. In Experiments 2 and 3, we

---

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

further explore this effect by isolating two components of semantic coherence—familiarity of words and presence of an overarching semantic organization—and find that neither alone facilitates learning. We conclude by discussing the limitations of purely artificial language learning, possibilities for future empirical and computational work, and potential mechanisms for the effects we observe.

1.1 Computational evidence for distributional learning

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

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

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

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

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

In the 1980s, computer scientists devised techniques that paved the way for large scale investigations of distributional learning. Motivated by practical issues in the field of information retrieval, they considered the relationships between words and documents. A typical problem was to retrieve documents relevant to a user query with certain search terms (i.e., a search engine). One solution to this problem is to represent documents as points in a high dimensional space whose dimensions are frequencies for different words (Salton & McGill, 1983). This approach lends itself quite naturally to a matrix representation with words labeling rows, documents labeling columns, and cells encoding how often a particular word occurs in a particular document (Fig. 1). While we can interpret such matrices as representing documents in terms of their component words, we can also interpret them representing words in terms of their frequency of use across different documents (a useful form for distributional learning). To make an analogy with our Alice–Bob example, such *word-document* matrices license inferences about Alice and Bob based on *where* they appear (rather than *who* they appear with).
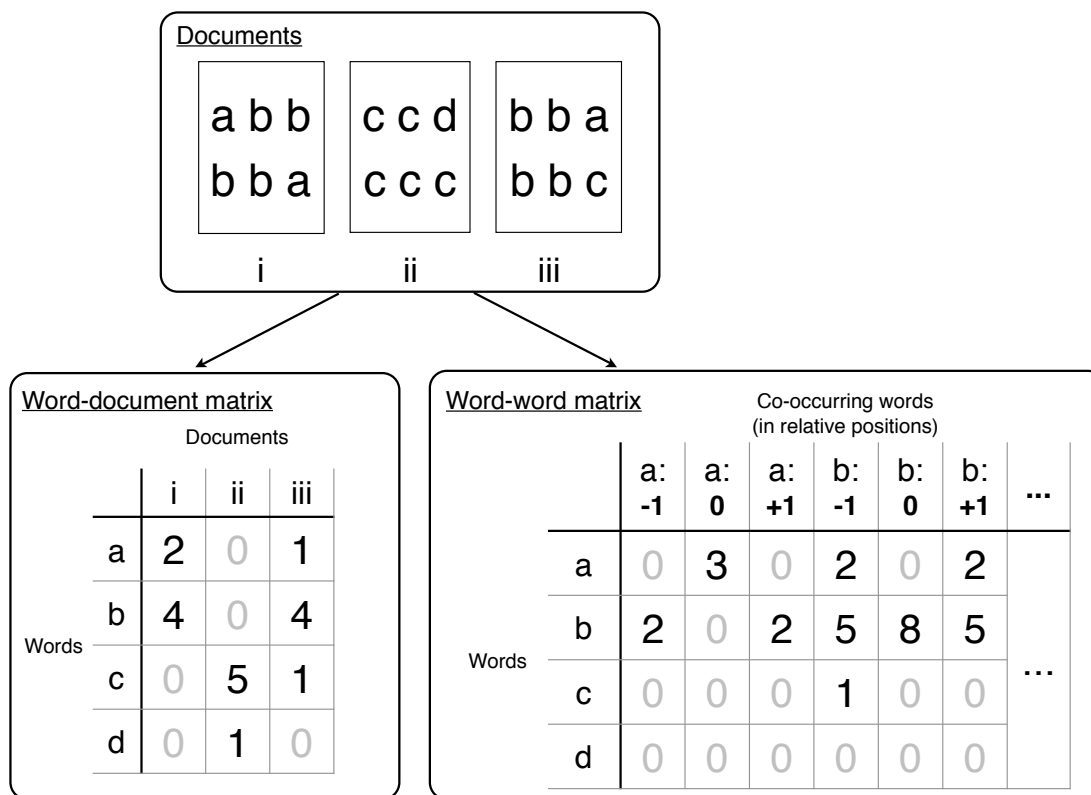
Documents

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

Word-document matrix
Documents

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

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

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

Fig. 1. Word-document and word-word matrices. Word-document matrices measure how often words occur in particular documents. Word-word matrices measure how often pairs of words occur in certain relative positions (e.g., the cell in the third row and fourth column of the word-word matrix indicates that "b" occurred before "c" once in the corpus).

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

Computational models operating on such co-occurrence representations can acquire both semantic and syntactic properties of words. In the domain of semantics, Landauer and Dumais

(1997) built a word-document matrix from a corpus of encyclopedia articles, applied a dimensionality reduction technique (singular value decomposition), and computed the similarity between words using a cosine measure. This Latent Semantic Analysis model (LSA) was able to pass a TOEFL synonym test successfully. In the domain of syntax, Redington et al. (1998) developed a model that performs hierarchical clustering on a word-word matrix to that successfully acquires grammatical categories like noun, verb, and adjective. In fact, these results suggest that co-occurrence statistics may be particularly powerful, offering leverage on many kinds of linguistic information. Although devised to learn grammatical categories, the model often learned clusters that had semantic organization. For example, in the cluster of adjectives, color words and number words formed separate clusters (p448).

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

1.2 A puzzle: the MNPQ language

The MNPQ language contains four categories of words, M (which includes m_1, m_2, and m_3), N (n_1, n_2, n_3), P (p_1, p_2, p_3), and Q (q_1, q_2, q_3). Participants hear two types of training sentences[2]: MN and PQ. Thus, sentences like $m_1 n_3$ are grammatical while sentences like $m_1 q_3$

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

are illegal. Early investigations (Braine, 1966; K. H. Smith, 1966) found that participants tend to endorse novel grammatical sentences as familiar. However, participants also endorse ungrammatical MQ and PN sentences, suggesting that they learn position regularities (that M/P come first and N/Q come second) but not co-occurrence regularities (that M co-occurs with N but not Q and that P occurs with Q but not N). This failure to learn categories on the basis of pure co-occurrence has been reliably observed in a number of studies (Braine, 1987; Brooks et al., 1993; Frank & Gibson, 2011; Frigo & McDonald, 1998; Gerken et al., 2005; Kempe & Brooks, 2001; Lany & Saffran, 2010). However, Reeder, Newport, and Aslin (2009, Experiment 5) report successful learning of a related language, (Q)AXB(R). In this language, there are optional Q and R categories that serve to deconfound co-occurrence regularities from positional regularities (i.e., in MNPQ, Ms and Ps are always sentence-initial and Ns and Qs are always sentence-final). Thus, this work by Reeder et al. suggests that positional regularities may override co-occurrence regularities in the MNPQ language. Nevertheless, the MNPQ failures are puzzling, given that computational models suggest that distributional learning is a powerful mechanism.

Of course, studies using artificial languages create learning conditions that differ in many ways from natural language learning. In the studies described above, people are presented with impoverished languages in which only co-occurrence information is available, and the question is whether people can learn the embedded structure under these conditions. The artificial input contrasts starkly with real input. To name just a few differences, real linguistic input (1) is absorbed over many years, (2) comes, at least during development, from knowledgeable and helpful human beings who are physically copresent, (3) exhibits non-uniform distributions, (4) has complex syntactic structure, (5) contains useful morphological structure, (6) refers to actual referents, and (7) contains already-known words. The failure to find learning from artificial input lacking these (and other) features is therefore to be taken with a grain of salt. Modulo issues of timescale and social interaction, these studies provide evidence against the quite strong hypothesis that people can learn from *purely* distributional information. But they do not adjudicate against weaker hypothesis classes. For example, it is possible that other information

sources must be present that synergize with distributional information.

Indeed, including multiple features of natural language in artificial language can lead to synergistic effects on learning. For instance, in work on segmentation, Kurumada, Meylan, and Frank (2013) found that presenting participants with Zipf-distributed input facilitated segmentation and Hay, Pelucchi, Estes, and Saffran (2011) found that presenting participants with natural language (Italian) containing peaked transitional probabilities facilitated subsequent word learning, bolstering previous findings that had used artificial languages. In what follows, we outline some research on MNPQ learning that has adopted the strategy of "naturalizing" artificial language by introducing correlated information sources. Of course, introducing some naturalistic features into artificial stimuli does not magically transform them into natural languages. These "semi-artificial" languages are therefore subject to many of the same concerns as fully artificial languages. Still, parametrically adding naturalistic features to fully artificial languages can serve as a useful empirical tool for measuring the impact of different information sources.

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

facilitate learning (e.g., Johns & Jones, 2012; Monaghan, Chater, & Christiansen, 2005).

Nearly all of this empirical work has interpreted the results of human experiments with reference to learning grammatical category. To our knowledge, only one study has examined word meaning. Recently, Lany and Saffran (2010) investigated Braine's proposal of correlating co-occurrence and phonological cues in a study of meaning acquisition. They found that 22-month old infants successfully learned an MNPQ language when co-occurrence was aligned with the number of syllables in a word (in particular, when N words were disyllabic and Q words were monosyllabic) but *not* when the number of syllables was not predictive of N/Q membership. This suggests that distributional information may play a common role in acquisition of grammatical category and word meaning, at least for the MNPQ language.

1.3   The current study

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

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

1. dog and dax

2. dog and ziv

3. car and wug

4. car and pif

might allow learners to infer that daxes and zivs are similar, as both words co-occur with "dog," and that wugs and pifs are similar, as both words co-occur with "car."

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

Our design takes our experiments out of the realm of purely artificial language learning, mixing familiar words with novel ones. At the same time, the impoverished learning materials we provide our participants cannot be taken as a parallel to the richness of natural language. It seems likely that participants reason about our materials using the same processes as in complex, artificial language tasks that include explicit semantic content (Braine, 1987). While artificial (or partially artificial) languages suitable for use in experiments cannot fully mimic natural language, they are useful as tools for investigating the relative ease or difficulty of learning from different kinds of information sources in model systems of this type.

2   Experiment 1a: Semantic Coherence

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

the sentences in our language (Table 1). We systematically varied properties of the context words and then measured learning of the target words.

In Experiment 1a, we parametrically varied two independent properties of the context words. First, we varied semantic coherence—the fraction of M/P words obeying a taxonomic organization (M = animal words, P = vehicle words). Second, as one hallmark of statistical learning is sensitivity to the amount of evidence observed, we varied the amount of exposure to the language to measure the efficiency of learning. After exposure to the language, we tested participants on three measures of MNPQ learning—similarity rating, sentence memory, and a referent assignment task. The sentence memory task measures whether participants learned the MNPQ co-occurrence structure. The similarity rating and referent assignment tasks measure whether participants use this information in judging similarity and assigning referents, respectively.

In the sentence memory task, participants rated the familiarity of four kinds of sentences: familiar sentences, novel grammatical sentences, and two kinds of ungrammatical sentences. In the similarity rating task, participants rated the similarity of pairs of target words. In the referent assignment task, we provided visual referents for the context words (e.g., "feeb" refers to a cat) and asked participants to assign the target words to referents (e.g., choose whether "chuv" refers to a horse or a bicycle).

2.1   Method

2.1.1   Participants

678 Amazon Mechanical Turk (MTurk) workers. Using MTurk's worker qualifications, we limited participation to workers located in the United States and with a previous task approval rate greater than or equal to 90%. We chose MTurk workers as our participants because the number of experimental conditions required a large number of participants. Work by Buhrmester, Kwang, and Gosling (2011) and Crump, McDonnell, and Gureckis (2013) suggests that MTurk is a valid platform for web-based learning experiments.

2.1.2 Materials

Sentences took the form "M and N" or "P and Q" (Table 1). Note that sentences literally included the word "and" in the middle. We generated the actual lexical items randomly for each participant. Ns and Qs were always novel nonsense words and were drawn without replacement from the set {moke, thite, jiv, pif, dex, wug}. Ms and Ps could be either novel or familiar. Novel Ms were drawn from {feeb, bim, lup} and novel Ps were drawn from {zabe, vap, chuv}. Familiar Ms and Ps obeyed a taxonomic organization—familiar Ms were drawn from {hamster, cat, dog} and familiar Ps were drawn from {car, bus, truck}.

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

2.1.3 Design and Procedure

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

---

[3] We programmatically submitted all of our sentences to the text-to-speech web service that powers Google Translate. During our investigation, the set of voices on the web service changed, which required us to synthesize our stimuli using different software. See also Footnote 9.

Table 1

The MNPQ language and test items for memory and similarity.

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

[a] Underlined sentences were withheld from exposure.

referent assignment (we presented memory and referent assignment later because we were concerned that earlier testing might give opportunities for additional learning). Below, we detail each of these tasks.

Exposure. Participants listened to sentences from the language. We withheld six sentences from exposure (Table 1), 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.

Similarity. For each pair of words in the union of N and Q, we asked participants to rate on a 5 point scale how similar they believed the two words to be in meaning. In particular, we instructed participants to "tell us how similar you think they are in terms of meaning". This resulted in within-category ratings (e.g., n_1 vs. n_2) and cross-category ratings (e.g., n_1 vs. q_1). We
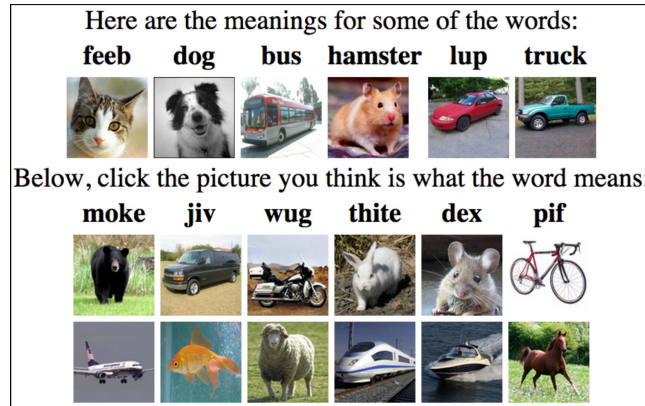
Fig. 2. A screenshot of the referent assignment task.

presented the pairs in a fixed pseudorandom order containing no repeated words between consecutive trials. Though exposure was entirely auditory, we presented these similarity questions as text (e.g., "How similar are **pif** ◀ŵ and **thite** ◀ŵ ?"); to facilitate mapping between visual and spoken word forms, the speaker button next to each word played the spoken word when clicked.[4] In two catch trials, we asked participants to press the response button corresponding to the solution of a simple arithmetic problem. If participants learned the MN and PQ co-occurrence relationships *and* used these relationships as a basis for similarity judgments, then we expected that participants would rate within-category pairs of words as more similar than cross-category pairs.

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

- *Familiar* sentences heard during exposure.

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

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

---

[4] We presented these questions as text because we were concerned that playing auditory questions would further train participants (furthermore, some of this unintended training would be on non-grammatical sequences).

- *Position violation* sentences of the form MM, NN, PP, and QQ.[5]

We presented sentences in random order such that there were no repeated words between consecutive trials. In two catch trials,[6] instead of a sentence from the MNPQ language, we played an audio instruction (which could not be repeated by participants) to press a specific response button. If participants learned the MN and PQ co-occurrence relationships, then we expected that they would rate novel grammatical sentences as more familiar than the co-occurrence violation sentences.

Referent assignment. We provided participants with referents for all of the Ms and Ps and asked them to choose referents for the Ns and Qs (Fig. 2). At the top of the screen, we displayed the Ms and Ps in random order and we provided referents by displaying a single image underneath each word. The M and P referents were either animals (cat, dog, hamster) or vehicles (car, bus, truck); either Ms were animals and Ps were vehicles, or vice versa. Recall that some conditions contained *familiar* M and P words; in these cases, we paired the known words with the obvious referents (e.g., "dog" was always paired with an image of a dog). Below the M and P words and their meanings, we displayed a row containing the N and Q words. Under each word, we displayed a two-alternative referent choice between an animal (the "correct" choice for N words) and vehicle words (the "correct" choice for Q words); participants made a choice by clicking on one of the two pictures. If participants learned the MN and PQ co-occurrence relationships *and* and used them as a basis for inferences about word meaning, then we expected that referent assignment scores would reflect a tendency to choose similar referents for words in the same category. In addition, we hypothesized that, under these circumstances, participants

---

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

[6] 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.
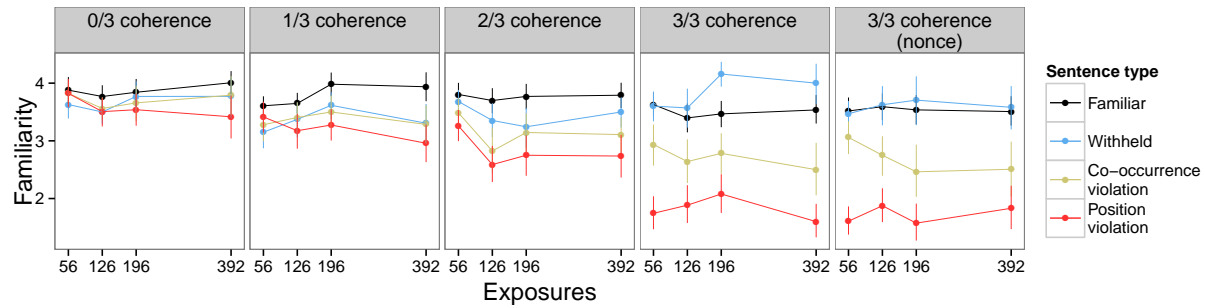
Fig. 3. Mean familiarity ratings for different sentence types in Experiment 1a and 1b. The far right panel shows results from Experiment 1b.

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

2.2   Results and Discussion

We excluded the 72 participants who did not correctly answer all of the catch trials.[7] We first present descriptive statistics of the memory task before analyzing the influence of coherence and exposure on distributional learning as measured by all three tasks. Anonymized data and analysis scripts are available at https://github.com/longouyang/semantic-coherence.

2.2.1   Descriptive statistics for the memory task

Fig. 3 shows mean familiarity ratings broken down by sentence type, coherence, and exposures. Within each coherence level, we looked for evidence of a familiarity ordering on sentence types. For each participant, we computed their mean familiarity rating for each sentence type. Then, at each coherence level (collapsing across exposure levels), we ran 6 paired-sample t-tests

---

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

comparing mean familiarity ratings across sentence types and corrected for multiple comparisons using the Holm procedure (Table B1). This yielded the following familiarity orderings:

$$0/3: F > W, C, P$$
$$1/3: F > W, C, P$$
$$2/3: F > W > C > P$$
$$3/3: W > F > C > P$$

These familiarity orderings indicate that some learning occurred at every level of coherence. For example, participants at all levels of coherence rated familiar sentences as more familiar than position violation sentences. However, only participants at the 2/3 and 3/3 levels distinguished between withheld and co-occurrence violation sentences. This lends some support to the idea that positional regularities in MNPQ may be more influential than co-occurrence regularities. It is interesting that, in the 3/3 condition, withheld sentences were rated significantly higher than *actually* familiar sentences. This may be related to prototype-enhancement effects found in category learning (e.g., J. D. Smith & Minda, 2002).

2.2.2 Effect of coherence and exposure

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

The memory score measured the difference in familiarity ratings for withheld versus co-occurrence violation sentences. For each participant, we computed their memory score as the mean rating for withheld sentences (e.g., m_1–n_1) minus their mean rating for novel co-occurrence violation sentences (e.g., m_1–q_1). Because ratings range from 1 to 5, this difference had a minimum value of -4 and a maximum value of 4; we divided this difference by 4 to rescale it to

lie in the interval [-1, 1]. Note that memory score did not include ratings for familiar or position-violation sentences.

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

The referent assignment score measured the total number of correct choices in the referent assignment task. The lowest score was 0 and the highest possible score was 6. We rescaled this to the interval [-1, 1] by taking the raw number correct, subtracting 3, and then dividing by 3.

We analyzed the effects of coherence and exposure using linear models. First, we looked for main effects of coherence on score (i.e., the Condition coefficients in Table 2). Second, as we were interested in the relationship between amount of exposure and score, we looked for exposure × coherence interactions. A significant interaction (e.g., the E × C coefficients in Table 2) would indicate a difference in how *efficiently* the statistical learning process makes use of evidence at different coherence levels. For all scores, we coded coherence as a categorical variable and analyzed the data using a regression which modeled the mean score in a participant group (e.g., 3/3-56) as an interactive function of the number of exposures (e.g., 56) times the condition (e.g., full coherence). In other words, our regression equation was score $\sim$ exposures × condition. Results are shown in Fig. 4.

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 2. 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
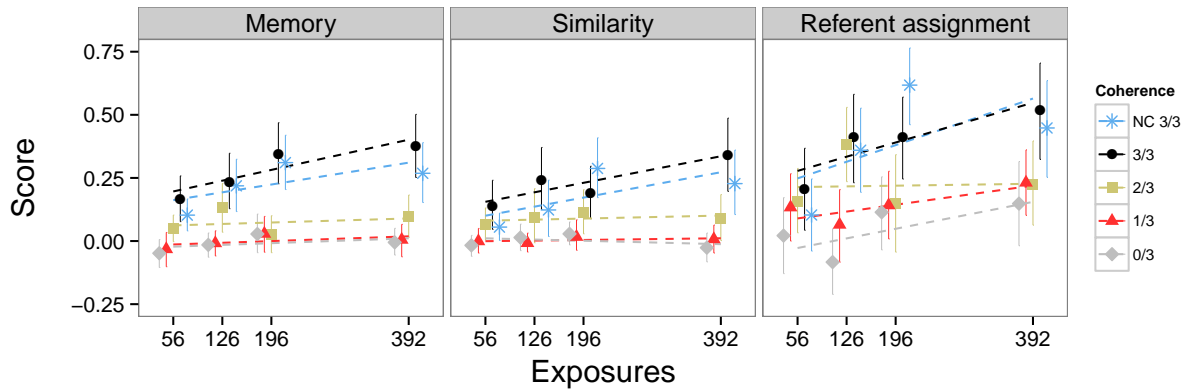
Fig. 4. Experiment 1a and 1b results. Each plot shows data for one measure (memory, similarity, referent assignment). Scores ranged from -1 to 1. Data points show condition means, error bars show bootstrapped 95% CIs, and dashed lines show the best-fitting linear regression.

condition.

Table 2

Regression model for Experiment 1a

Predictor	β	Std Error	t	p
Memory: $R^2 = 0.18$, $F(7, 598) = 19.26$				
Intercept	0.0432	0.019	2.26	<0.05*
Condition: 1/3 – (0/3)	0.0043	0.027	0.15	0.874
Condition: 2/3 – (0/3,1/3)	0.0269	0.015	1.78	0.074
Condition: 3/3 – (0/3,1/3,2/3)	0.0396	0.010	3.62	<0.001*
Exposures	0.2215	0.084	2.63	<0.01*
E × C: 1/3 – (0/3)	-0.0012	0.121	-0.00	0.992
E × C: 2/3 – (0/3,1/3)	-0.0050	0.067	-0.07	0.939
E × C: 3/3 – (0/3,1/3,2/3)	0.1294	0.048	2.65	<0.01*
Similarity: $R^2 = 0.12$, $F(7, 598) = 11.85$				
Intercept	0.0536	0.019	2.76	<0.01*
Condition: 1/3 – (0/3)	-0.0085	0.028	-0.30	0.763

Condition: 2/3 – (0/3,1/3)	0.0237	0.015	1.54	0.123
Condition: 3/3 – (0/3,1/3,2/3)	0.0239	0.011	2.13	<0.05*
Exposures	0.1407	0.085	1.63	0.101
E × C: 1/3 – (0/3)	0.0517	0.124	0.41	0.676
E × C: 2/3 – (0/3,1/3)	0.0258	0.068	0.37	0.705
E × C: 3/3 – (0/3,1/3,2/3)	0.1320	0.049	2.65	<0.01*

Referent assignment: $R^2 = 0.07$, $F(7, 594) = 6.70$

Intercept	0.1136	0.036	3.10	<0.005*
Condition: 1/3 – (0/3)	0.0629	0.053	1.17	0.240
Condition: 2/3 – (0/3,1/3)	0.0689	0.028	2.38	<0.05*
Condition: 3/3 – (0/3,1/3,2/3)	0.0397	0.021	1.88	0.059
Exposures	0.4434	0.161	2.74	<0.01*
E × C: 1/3 – (0/3)	-0.0784	0.233	-0.33	0.736
E × C: 2/3 – (0/3,1/3)	-0.1421	0.128	-1.10	0.268
E × C: 3/3 – (0/3,1/3,2/3)	0.1201	0.094	1.27	0.202

Note. For readability, we divided the exposure values by 1,000.

2.2.3 Task results

We report performance on each of the three tasks separately.

Memory. More exposure led to higher memory scores, as there was a main effect of exposure. Higher coherence also led to higher memory scores; 3/3 scores were significantly higher than scores from the rest of the conditions combined.

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

regularities, at least in the 3/3 condition.

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

Referent assignment. 2/3 scores were significantly higher than 0/3 and 1/3 scores combined. 3/3 scores were marginally higher than the rest of the scores combined, possibly because 3/3 coherence and 2/3 coherence may confer comparable advantages on this task. We also computed this model using coherence as a continuous variable; the continuous coherence regressor significantly predicted increases in score, $\beta = 0.09$, $t(619) = 2.87$, $p < 0.005$, suggesting that *parametrically* increasing coherence results in *parametric* increases in referent assignment score. None of the interaction terms reached significance, indicating that the amount of exposure to the language and greater coherence independently increased the ability to assign N and Q words to the correct referents.

2.2.4 Summary

In Experiment 1a, 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 category boundaries based on the co-occurrence regularities in the similarity task, and (3) increased inductive bias in associating target words with referents.

However, one possible issue with our experiments is that the context words were not the only known words. In particular, we linked these words using the conjuctive predicate "and." Arguments to "and" typically belong to the same syntactic category and may often share semantic category membership as well Benor and Levy (2006). An English speaker could use knowledge of this regularity to infer probable semantic categories for target words. For instance, in the phrase "dog and dax," she could infer that daxes belong to a semantic category that is similar to

dogs. Indeed, previous studies have shown that people use selectional restrictions of predicates to constrain inferences about semantic properties of novel arguments (Altmann & Kamide, 1999; Köhne & Crocker, 2010).[8] This raises the question: do our results hinge on the use of the familiar connective "and"? Or would we also observe these results using a nonce connective? In Experiment 1b, we investigate this question.

3   Experiment 1b: Semantic coherence with a nonce connective

In Experiment 1b, we investigated whether the familiar "and" connective drove the effects of coherence that we found in Experiment 1a. We modified the materials from the 3/3 coherence condition, replacing "and" with a nonce connective, "tezz."

3.1   Method

3.1.1   Participants

163 MTurk workers.

3.2   Materials

The materials were identical to the 3/3 condition of Experiment 1a, but with "and" replaced with "tezz."

3.3   Design and Procedure

The design and procedure were identical to that of Experiment 1a.

3.4   Results and Discussion

We excluded 15 subjects who did not pass all trials.

---

[8]We thank an anonymous reviewer for this observation.

Table 3

Regression model from Experiment 1b

Predictor	β	Std Error	t	p
Memory $R^2 = 0.03$, $F(1, 146) = 4.60$				
Intercept	0.16	0.03	4.42	<0.001*
Exposures	0.44	0.20	2.14	<0.05*
Similarity $R^2 = 0.03$, $F(1, 146) = 5.61$				
Intercept	0.10	0.03	2.60	<0.05*
Exposures	0.51	0.21	2.37	<0.05*
Referent assignment $R^2 = 0.04$, $F(1, 144) = 7.04$				
Intercept	0.24	0.06	4.00	<0.001*
Exposures	0.94	0.35	2.65	<0.01*

Note. For interpretability of the intercept, we subtracted the lowest level of exposure from all exposure values. For readability, we divided exposure levels by 1,000.

3.4.1 Descriptive statistics for the memory task

Fig. 3 shows mean familiarity ratings broken down by sentence type, coherence, and exposures. As in Experiment 1a, we used Holm-corrected pairwise t-tests to compute a familiarity ordering on sentence types (Table B1). The ordering for the nonce connective 3/3 condition was:

$$(F \approx W) > C > P$$

Thus, participants who saw a version of the 3/3 language with "and" replaced by "tezz" successfully distinguished between novel grammatical (i.e., withheld) sentences and ungrammatical (i.e., co-occurrence violation) sentences. If results of Experiment 1a hinged on the use of the "and" connective, participants should not have been able to distinguish between withheld and co-occurrence violation sentences.

It is informative to compare this nonce connective ordering with the orderings from the 2/3 and 3/3 conditions of Experiment 1a:

2/3:   F > W > C > P

Nonce: (F ≈ W) > C > P

3/3:   W > F > C > P

In 2/3, familiar sentences were rated as more familiar than withheld grammatical sentences, whereas the opposite was true in 3/3. Our nonce connective 3/3 condition appears to be in between these two cases; the difference in ratings between familiar and withheld sentences was not statistically significant.

3.4.2   Task results

Similar to Experiment 1a, we performed a regression analyzing score on each task as a function of exposures. Note that in Experiment 1a, we also included coherence as a predictor but we did not include it here as there was only a single coherence level – 3/3.

Regression results are shown in Table 3. Note that in our model, we subtracted 56 from all exposure levels so that the intercept could be interpreted as performance at the lowest exposure level. For all three tasks, the intercept was significant, indicating that participants performed above chance at the lowest exposure level. Additionally, for all three tasks, there was evidence of statistical learning, as the effect of exposure level was significant—more exposures were associated with higher task scores.

3.5   Conclusion

Our analyses indicate that the results of Experiment 1a were not entirely driven by the presence of the "and" connective. Even with the nonce connective "tezz," we observed the same hallmarks of learning: distinguishing between withheld grammatical sentences and co-occurrence violation sentences and sensitivity to the amount of exposure. It is possible that "and" *further* facilitates learning beyond the effects of semantic coherence. However, our results here demonstrate that the presence of a known connective is not *necessary* for distributional learning. Thus, taken together, Experiments 1a and 1b suggest that semantic coherence facilitates

distributional learning.

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

4   Experiment 2: Phonological coherence

In Experiment 2, we investigated whether learners could learn the language used in Experiment 1a when the context words (Ms and Ps) exhibited phonological, rather than semantic, coherence. It is worth emphasizing that we manipulated phonological coherence in the *context* words, whereas previous work (e.g., Lany & Saffran, 2010) has manipulated phonological coherence of the *target* words. This previous work has shown that learning is facilitated when target words have both distributional and phonological regularities. By contrast, we are interested in how distributional learning for one category of words varies as a function of *other* words in the linguistic environment.

We tested three types of coherence: onset, rime, and syllable count. We selected these kinds of coherence because work in other areas suggests that these kinds of coherence may reliably occur in natural language or facilitate distributional learning when applied to *target* words. Onset coherence might facilitate learning due to sound symbolism (e.g., words that start with *gl* have similar meanings: glimmer, glow, gleam). Rime coherence can occur in languages that require gender-marking agreement between nouns and verbs. Finally, syllable count coherence has been shown to facilitate learning when applied to target words (Lany & Saffran, 2010). Thus, we investigated it to determine whether it also facilitates learning when applied to context words.

4.1 Method

4.1.1 Participants

530 MTurk workers.

4.1.2 Materials

The three types of phonological coherence[9] were:

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

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

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

4.1.3 Design and Procedure

The design and procedure were identical to that of Experiment 1a.

4.2 Results and Discussion

We discarded the 51 participants who did not pass all the catch trials. Results are graphed in Fig. 6.
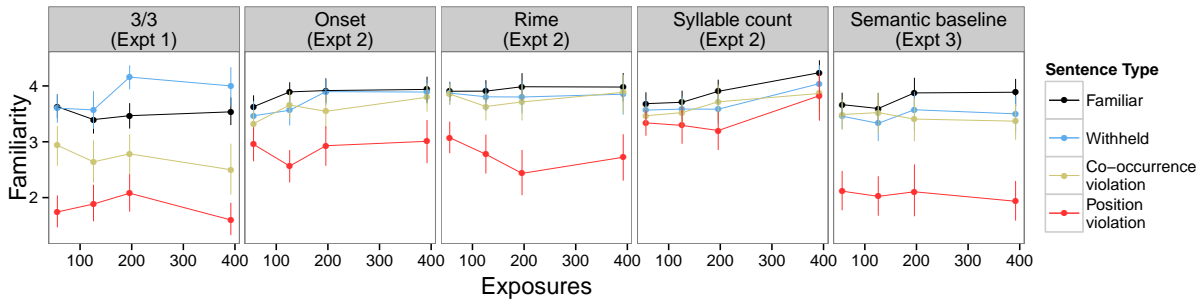
Fig. 5. Mean familiarity ratings for different sentence types in Experiments 2 and 3. For comparison, we also include the 3/3 coherence condition from Experiment 1a.

4.2.1 Descriptive statistics for the memory task

Fig. 5 shows mean familiarity ratings broken down by sentence type, coherence, and exposures. As in previous analyses, we computed a familiarity ordering on sentence types within each phonological coherence condition using paired sample t-tests (Table B1):

Onset: $F \approx W$ $F > C, P$ $(W \approx C) > P$

Rime: $F \approx W$ $F > C, P$ $(W \approx C) > P$

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

Participants in these conditions all failed to distinguish between withheld and co-occurrence violation sentences; they exhibited some learning of the language but not of the co-occurrence regularities.

4.2.2 Effects of condition and exposure

---

[9] The stimuli for the rime and syllable count conditions differ from those in the rest of our conditions. For the rest of the conditions, we used a text-to-speech web service provided by Google to generate the audio stimuli (see Footnote 3) for the bulk of the conditions. However, the available voices on this service changed during our experiment. Thus, we generated new stimuli for the rime and syllable count conditions using commercially available software, NaturalReader 10. To ensure that the old and new stimuli were comparable, we performed a partial replication of Experiment 1a using the new synthesis engine; the difference old and new stimuli did not appear to make a substantial difference. See Appendix A for comparisons.
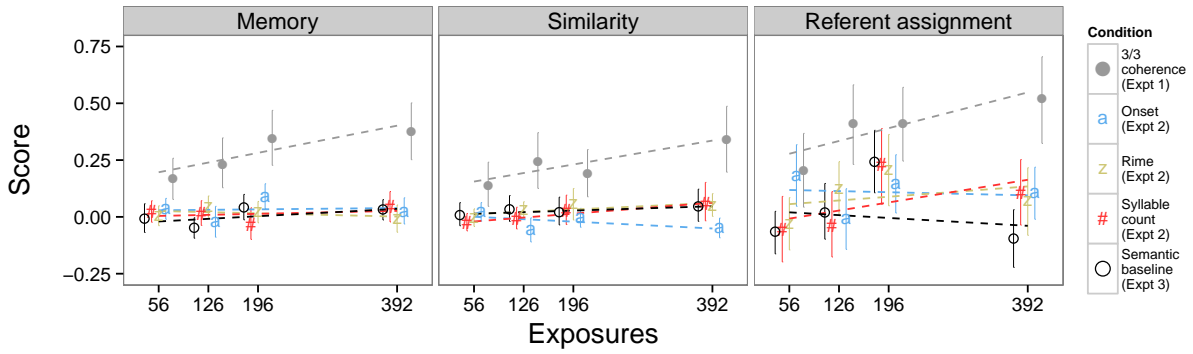
Fig. 6. Experiments 2 and 3 results. Each plot shows data for one measure (memory, similarity, referent assignment). Scores ranged from -1 to 1. Points show condition means, error bars show 95% CIs, and dashed lines show the best-fitting linear trend. For comparison, we also include the 3/3 coherence condition from Experiment 1a.

Table 4

Regression model for Experiments 2 and 3

Predictor	β	Std Error	t	p
Memory: $R^2 = 0.01$, $F(9, 756) = 0.89$				
Intercept	-0.0277	0.028	-0.98	0.322
Condition: Onset – 0/3	0.0550	0.038	1.42	0.153
Condition: Rime – 0/3	0.0578	0.036	1.59	0.111
Condition: Syllable count – 0/3	0.0283	0.039	0.72	0.469
Condition: Semantic baseline – 0/3	-0.0016	0.038	-0.04	0.966
Exposures	0.0984	0.121	0.80	0.419
E × C: Onset – 0/3	-0.0686	0.171	-0.39	0.689
E × C: Rime – 0/3	-0.1713	0.164	-1.04	0.297
E × C: Syllable count – 0/3	-0.0324	0.169	-0.19	0.848
E × C: Semantic baseline – 0/3	0.0667	0.165	0.40	0.687

Similarity: $R^2 = 0.02$, $F(9, 756) = 2.10$

Intercept	0.0146	0.026	0.56	0.573
Condition: Onset – 0/3	-0.0054	0.035	-0.15	0.878
Condition: Rime – 0/3	-0.0111	0.033	-0.33	0.740
Condition: Syllable count – 0/3	-0.0487	0.036	-1.33	0.180
Condition: Semantic baseline – 0/3	-0.0060	0.035	-0.16	0.866
Exposures	-0.0688	0.113	-0.60	0.542
E × C: Onset – 0/3	-0.0847	0.159	-0.53	0.595
E × C: Rime – 0/3	0.2168	0.152	1.42	0.155
E × C: Syllable count – 0/3	0.3152	0.157	1.99	<0.05*
E × C: Semantic baseline – 0/3	0.1662	0.153	1.08	0.279

Referent assignment: $R^2 = 0.01$, $F(9, 752) = 1.42$

Intercept	-0.0580	0.069	-0.83	0.404
Condition: Onset – 0/3	0.1804	0.095	1.88	0.059
Condition: Rime – 0/3	0.1010	0.089	1.12	0.261
Condition: Syllable count – 0/3	0.0218	0.097	0.22	0.821
Condition: Semantic baseline – 0/3	0.0880	0.096	0.91	0.360
Exposures	0.5439	0.301	1.80	0.071
E × C: Onset – 0/3	-0.6143	0.425	-1.44	0.148
E × C: Rime – 0/3	-0.3114	0.406	-0.76	0.443
E × C: Syllable count – 0/3	-0.0351	0.420	-0.08	0.933
E × C: Semantic baseline – 0/3	-0.7207	0.409	-1.76	0.078

Note. For readability, we divided exposure values by 1000.

Using a regression model with main effects of exposure and condition and an exposure ×
condition interaction, we compared each phonological condition with the 0/3 condition of
Experiment 1a using a regression model (Table 4).

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

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

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

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

5   Experiment 3: Semantic baseline

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

5.1 Methods

5.1.1 Participants

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

5.1.2 Materials

In the semantic baseline language, the specific M and P words were drawn randomly for each participant from the pool {*shelf*, *glove*, *rain*, *leash*, *card*, *ball*}. In the referent assignment task, these known words were paired with images of the obvious referents (e.g., *card* with a picture of a card). Compared to the context words in Experiment 1a, these words are fairly well matched in frequency (see Table 5).

5.1.3 Design and procedure

The design and procedure were identical to that of Experiment 1a.

5.2 Results and Discussion

We discarded the 21 participants who did not pass all the catch trials.

5.2.1 Descriptive statistics of the memory task

As in Experiment 1a, we computed an ordering on the sentence types in the memory task:

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

(see Table B1). This indicates that participants were able to learn some regularities in the language, but not the co-occurrence regularities.

5.2.2 Effect of exposure

Results are graphed in Fig. 6. See Table 4 for regression results.

Table 5

Context word frequencies from Experiments 1 and 3

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

Note. Frequencies were retrieved from Google

N-grams data for the year 1999.

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.

6   General Discussion

Can people learn about words from co-occurrence statistics? Previous work on distributional learning based on co-occurrence presented a paradox: While computational models suggest that distributional statistics *per se* are a powerful source of information, experiments with humans show consistent failures without additional correlated features. Our experiments suggest a partial resolution. For people, distributional learning is facilitated by semantic coherence: Knowing that Alice associated with X and Y tells us something about Alice only if X and Y are already meaningful and meaningfully related to one another. Removing either the meanings of the individuals or the coherence of the relationship between them removes the facilitation.

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

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

An additional way that our work contrasts with prior artificial language learning work is that in our tasks participants may have conceptualized their task as acquiring new words in a language they already know (English), whereas prior studies with purely artificial stimuli may have encouraged participants to conceptualize their task as acquiring an entirely new language. Thus, comparing our results with purely artificial experiments is only possible to the extent that the learning mechanisms in these two cases (extending a familiar language versus acquiring a new language) are similar.

6.1   Limitations and Future Directions

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

Second, we investigated a single type of semantic coherence, where context words fell into two different taxonomic categories, and a single syntactico-semantic construction that biases interpretation towards concrete nouns (e.g., "cat and wug"). Although Experiment 1b speaks against the interpretation that the predicate "and" was responsible for our results, nevertheless these choices likely affected the magnitude of the effects we observed. While the simplicity and transparency of our semantic categories likely facilitated learning, our use of concrete noun targets might even have understated the effects of coherence. Distributional learning through semantic coherence might be especially useful for words like verbs and adjectives (Redington et al., 1998), where referential grounding provides less information about meaning (Gleitman, 1990).

Finally, while our study addressed one shortcoming of much artificial language

research—the lack of semantic grounding—it is still congruent with other artificial language research in scale (6 words in a single exposure session). Small-scale experiments do not allow strong inferences about the importance of different factors in lexical learning (Frank et al., 2013; Romberg & Saffran, 2010). Moving forward, we believe that a mixture of computational analyses and larger-scale experiments are necessary to understand the roles that semantic coherence and distributional learning play in language learning.

6.2   Conclusions

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

References

Altmann, G. T. M., & Kamide, Y. (1999, dec). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264. `doi: 10.1016/s0010-0277(99)00059-1`

Baldwin, D. A. (1993). Early referential understanding: Infants' ability to recognize referential acts for what they are. *Developmental Psychology*, *29*(5), 832. `doi: 10.1037/0012-1649.29.5.832`

Benor, S., & Levy, R. (2006). The chicken or the egg? a probabilistic analysis of english binomials. *Language*, *82*(2), 233–278. `doi: 10.1353/lan.2006.0077`

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

Braine, M. D. S. (1966). Learning the positions of words relative to a marker element. *Journal of Experimental Psychology*, *72*(4), 532–540. `doi: 10.1037/h0023763`

Braine, M. D. S. (1987). What is learned in acquiring word classes: A step toward an acquisition theory. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 65–87). Hillsdale, NJ: Lawrence Erlbaum. `doi: 10.4324/9781315798721`

Brooks, P. J., Braine, M. D. S., Catalano, L., Brody, R. E., & Sudhalter, V. (1993). Acquisition of gender-like noun subclasses in an artificial language: The contribution of phonological markers to learning. *Journal of Memory and Language*, *32*(1), 76–95. `doi: 10.1006/jmla.1993.1005`

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5. `doi: 10.1037/e527772014-223`

Church, K., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*(1), 22–29. `doi: 10.3115/981623.981633`

Clark, E. V. (1988). On the logic of contrast. *Journal of Child Language*, *15*(02), 317. `doi: 10.1017/s0305000900012393`

Clark, H. (1968). On the use and meaning of prepositions. *Journal of Verbal Learning and Verbal Behavior*, *7*(2), 421–431. `doi: 10.1016/s0022-5371(68)80027-1`

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating amazon's mechanical turk as a tool for experimental behavioral research. *PloS One*, *8*(3), e57410. `doi: 10.1371/journal.pone.0057410`

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

Frank, M. C., & Gibson, E. (2011). Overcoming memory limitations in rule learning. *Language Learning and Development*, *7*(2), 130–148. `doi: 10.1080/15475441.2010.512522`

Frank, M. C., Tenenbaum, J. B., & Fernald, A. (2013). Social and discourse contributions to the determination of reference in cross-situational word learning. *Language Learning and Development*, *9*(1), 1–24. `doi: 10.1080/15475441.2012.707101`

Frigo, L., & McDonald, J. L. (1998). Properties of phonological markers that affect the acquisition of gender-like subclasses. *Journal of Memory and Language*, *39*(2), 218–245. `doi: 10.1006/jmla.1998.2569`

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

Gerken, L., Wilson, R., & Lewis, W. (2005, may). Infants can use distributional cues to form syntactic categories. *Journal of Child Language*, *32*, 249–268. `doi: 10.1017/s0305000904006786`

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, *1*(1), 3–55. `doi: 10.1207/s15327817la0101_2`

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

Hay, J. F., Pelucchi, B., Estes, K. G., & Saffran, J. (2011). Linking sounds to meanings: Infant statistical learning in a natural language. *Cognitive Psychology*, *63*(2), 93–106. `doi: 10.1016/j.cogpsych.2011.06.002`

Hills, T. T., Maouene, J., Riordan, B., & Smith, L. B. (2010). The associative structure of

language: Contextual diversity in early word learning. *Journal of Memory and Language*, *63*(3), 259–273. doi: 10.1016/j.jml.2010.06.002

Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. (2009). Longitudinal analysis of early semantic networks: Preferential attachment or preferential acquisition? *Psychological Science*, *20*(6), 729–739. doi: 10.1111/j.1467-9280.2009.02365.x

Hollich, G. J., Hirsh-Pasek, K., & Golinkoff, R. M. (2000). Breaking the language barrier: An emergentist coalition model for the origins of word learning. *Monographs of the Society for Research in Child Development*, *65*(3), i–135.

Johns, B. T., & Jones, M. (2012). Perceptual inference through global lexical similarity. *Topics in Cognitive Science*, *4*(1), 103–120. doi: 10.1111/j.1756-8765.2011.01176.x

Kempe, V., & Brooks, P. J. (2001). The role of diminutives in the acquisition of russian gender: Can elements of child-directed speech aid in learning morphology? *Language Learning*, *51*(2), 221–256. doi: 10.1111/j.0023-8333.2005.00297.x

Köhne, J., & Crocker, M. W. (2010). Sentence processing mechanisms influence cross-situational word learning. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd annual meeting of the cognitive science society* (pp. 2548–2463). Austin.

Kurumada, C., Meylan, S. C., & Frank, M. C. (2013). Zipfian frequency distributions facilitate word segmentation in context. *Cognition*, *127*(3), 439–453. doi: 10.1016/j.cognition.2013.02.002

Landauer, T. K., & Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211. doi: 10.1037/0033-295x.104.2.211

Lany, J., & Saffran, J. (2010). From statistics to meaning: Infants' acquisition of lexical categories. *Psychological Science*, *21*(2), 284. doi: 10.1177/0956797609358570

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

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

Cambridge, MA: MIT Press.

Monaghan, P., Chater, N., & Christiansen, M. H. (2005). The differential role of phonological and distributional cues in grammatical categorisation. *Cognition*, *96*(2), 143–182. `doi: 10.1016/j.cognition.2004.09.001`

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, *22*(4), 425–469. `doi: 10.1016/s0364-0213(99)80046-9`

Reeder, P., Newport, E., & Aslin, R. (2009). The role of distributional information in linguistic category formation. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st annual meeting of the cognitive science society* (pp. 2564–2569). Amsterdam.

Riordan, B., & Jones, M. N. (2010). Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*, *3*(2), 303–345. `doi: 10.1111/j.1756-8765.2010.01111.x`

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(6), 906–914. `doi: 10.1002/wcs.78`

Rubenstein, H., & Goodenough, J. (1965). Contextual correlates of synonymy. *Communications of the ACM*, *8*(10), 627–633. `doi: 10.1145/365628.365657`

Saffran, J., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928. `doi: 10.1126/science.274.5294.1926`

Saffran, J., Newport, E., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*(4), 606–621. `doi: 10.1006/jmla.1996.0032`

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill, Inc.

Schütze, H. (1992). Dimensions of meaning. In R. Werner (Ed.), *Proceedings of the 1992 acm/ieee conference on supercomputing* (pp. 787–796). Minneapolis, MN. `doi: 10.1109/superc.1992.236684`

Smith, J. D., & Minda, J. P. (2002). Distinguishing prototype-based and exemplar-based processes in dot-pattern category learning. *Journal of Experimental Psychology: Learning,*

Memory, and Cognition, *28*(4), 800–811. doi: `10.1037/0278-7393.28.4.800`

Smith, K. H. (1966). Grammatical intrusions in the recall of structured letter pairs: Mediated
     transfer or position learning? *Journal of Experimental Psychology*, *72*(4), 580–588. `doi:`
     `10.1037/h0023768`

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational
     statistics. *Cognition*, *106*(3), 1558–1568. `doi:`
     `10.1016/j.cognition.2007.06.010`

Stefflre, V., & Reich, P. (1971). Some eliciting and computational procedures for descriptive
     semantics. In P. Kay (Ed.), *Explorations in mathematical anthropology* (pp. 79–116).
     Cambridge, Mass.: MIT Press.

Szalay, L. B., & Bryson, J. A. (1974). Psychological meaning: Comparative analyses and
     theoretical implications. *Journal of Personality and Social Psychology*, *30*(6), 860–870.
     `doi:  10.1037/h0037607`

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

Yu, C., & Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical
     and social cues. *Neurocomputing*, *70*(13-15), 2149–2165. `doi:`
     `10.1016/j.neucom.2006.01.034`

Appendix A

Old versus new stimuli comparison

We performed a partial replication of Experiment 1a with the new stimuli. We collected data in

four out of the sixteen experimental conditions and the results mirror those found with the old
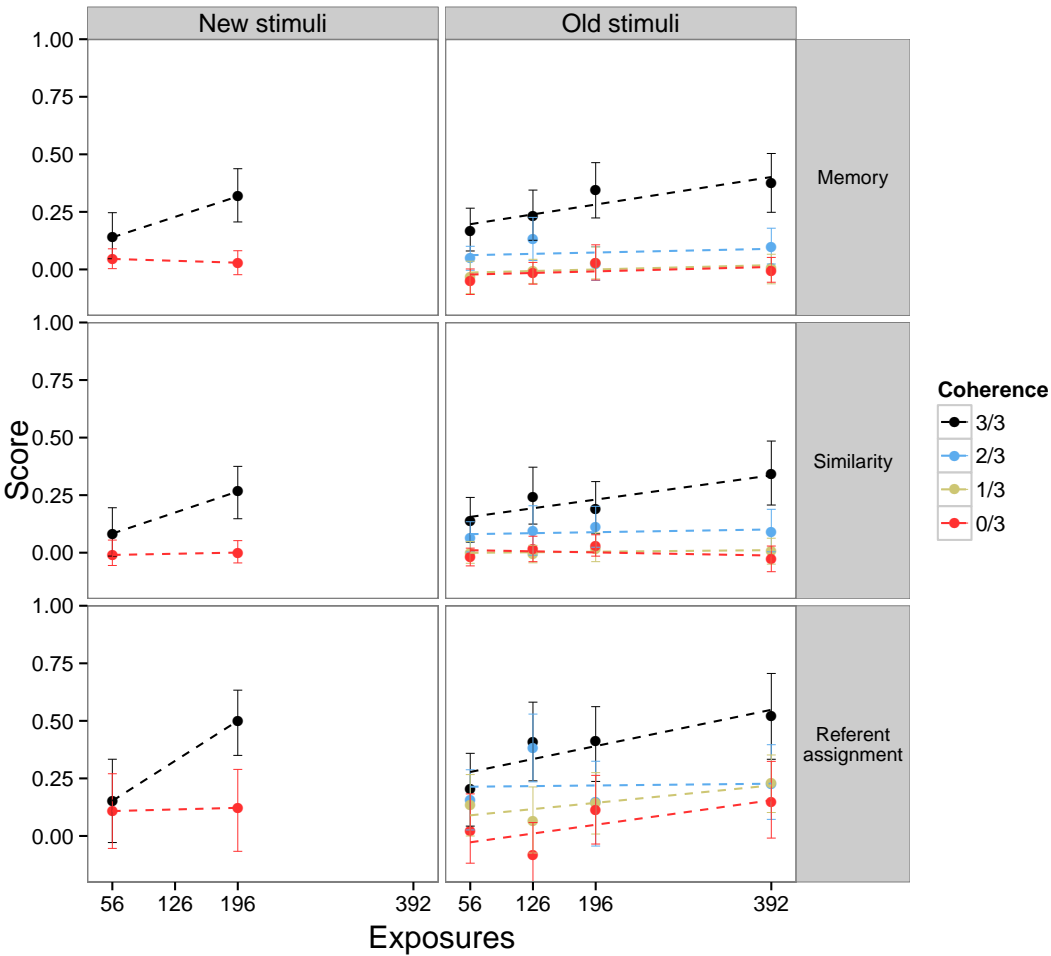
stimuli:

Fig. A1. Old versus new stimuli

Appendix B

t-tests for familiarity orderings in the memory task

Table B1

p values (adjusted using the Holm procedure) for paired-sample t-tests of familiarity ratings for (F)amiliar, (W)ithheld, (C)o-occurrence violation, and (P)osition violation sentence types in Experiments 1a, 1b, 2, and 3.

0/3 coherence: df = 145

	F	W	C
W	*		
C	*	0.58	
P	***	0.25	0.14

1/3 coherence: df = 148

	F	W	C
W	***		
C	***	0.98	
P	***	0.07	0.07

2/3 coherence: df = 159

	F	W	C
W	***		
C	***	***	
P	***	***	***

3/3 coherence: df = 150

	F	W	C
W	***		
C	***	***	
P	***	***	***

Nonce connective: df = 147

	F	W	C
W	0.56		
C	***	***	
P	***	***	***

Onset: df = 148

	F	W	C
W	0.055		
C	***	0.055	
P	***	***	***

Rime: df = 184

	F	W	C
W	0.16		
C	**	0.17	
P	***	***	***

Syllable count: df = 140

	F	W	C
W	**		
C	***	0.4	
P	***	**	**

Semantic baseline: df = 144

	F	W	C
W	***		
C	***	0.83	
P	***	***	***

*: $p < 0.05$ **: $p < 0.01$ ***: $p < 0.001$