

Semantic Coherence Facilitates Distributional Learning of Word Meanings

Long Ouyang, Lera Boroditsky, and Michael C. Frank

{louyang, lera, mcfrank}@stanford.edu

Department of Psychology

Stanford University

Abstract

Computational work has suggested that one could, in principle, learn aspects of word meaning simply from patterns of co-occurrence between words. The extent to which humans can do this distributional learning is an open question – artificial language learning experiments have yielded mixed results, prompting suggestions that distributional cues must be correlated with other cues, such as phonological regularities, for successful learning. We conducted a large-scale experiment comparing how distributional learning is afforded by two different types of cues – phonological regularities and semantic coherence. We found that semantic coherence more strongly facilitates distributional learning than onset-consonant phonological regularities.

Keywords: word learning; distributional learning; semantic coherence

Introduction

How do people learn what words mean? According to the distributional hypothesis (Harris, 1951; Firth, 1957), patterns of word co-occurrence are a powerful source of information about word meaning. It may be possible to acquire some facets of word meaning by simply keeping track of which linguistic contexts words appear in, even without access to any physical or social cues. Computational models have lent quantitative support for the distributional hypothesis. For example, the Latent Semantic Analysis model of Landauer & Dumais (1997) clustered words according to their patterns of occurrence across documents and was able to closely match human performance on synonym tests. The success of early models like LSA led to a proliferation of models that use distributional information to learn word meaning (see Riordan & Jones, 2010 for an overview) as well as other linguistic properties such as syntactic category (e.g., Redington, Chater, & Finch, 1998).

We know that distributional learning is computationally possible, but the evidence of human capacity for distributional learning is mixed. Some research has found that people learn co-occurrence statistics and use them to inform categorization (Mintz, 2002; Gerken, Wilson, & Lewis, 2005; Reeder, Newport, & Aslin, 2010), but other work has reached the opposite conclusion (Braine, 1966; Smith, 1966; Frank & Gibson, 2009). A holistic reading of the literature suggests that learning is more likely to occur when additional cues are correlated with distributional properties. Below, we review two representative findings using the *MNPQ* artificial language learning paradigm (Braine, 1966; Smith, 1966).

The *MNPQ* grammar contains four categories of words – *M*, *N*, *P*, and *Q* – and sentences take one of two forms: *MN* and *PQ*. Early investigations (Braine, 1966; Smith, 1966)

found that subjects recall “grammatical” *MN* and *PQ* sentences but also tend to recall 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*). *MNPQ* learning has been an important paradigm because it provides an empirical means to consider purely distributional learning of word categories.

Braine (1987) investigated the effect of correlating co-occurrence regularities with natural gender. Subjects acquired an artificial language by learning to name pictures of referents. In the experimental condition, all pictures of men were labeled by *M* words and all pictures of women were labeled by *P* words. Learning of the co-occurrence regularities was significantly higher in the experimental condition than in a control condition where natural gender was not predictive of *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 learned *MNPQ* co-occurrence regularities when they were aligned with the number of syllables in a word – that is, when *N* words were disyllabic and *Q* words were monosyllabic, but *not* when the number of syllables was not predictive of *N/Q* membership.

The defining feature of the artificial language learning paradigm is that, at the outset of the experiment, subjects do not know the meanings of any of the words. Yet, this condition generally does not hold true for real language learners, who typically know the meanings of some words but not others. It is plausible that the presence of known words facilitates distributional learning. If this is true, how effective are such semantic cues, and how do they compare to (e.g.) phonological cues? Here, we report the initial results of the first large-scale study comparing the effectiveness of various correlated cues on distributional learning. We presented subjects with an *MNPQ* language where sentences took the form “*M* and *N*” or “*P* and *Q*”. We hypothesized that distributional learning would be afforded given a certain level of *semantic coherence*, where all *M* and *P* words were familiar and adhered to some semantic organization (e.g., *M* = animal words, *P* = vehicle words). For instance, hearing the four sentences “cat and dax”, “cat and ziv”, “car and wug”, and “car and pif” might allow learners to infer that daxes and zivs belong to the

same category, as both words co-occur with “cat”, 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 Experiments 2 and 3, we compared semantic coherence to phonological coherence and semantic incoherence (the presence of known words that *do not* adhere to some semantic organization).

Experiment 1: Semantic Coherence

We parametrically varied two independent variables: (1) the amount of exposure to the language and (2) semantic coherence – the fraction of known words that adhered to a taxonomic organization (M = animal words, P = vehicle words). We then applied three measures of $MNPQ$ learning – sentence memory, similarity rating, and a referent assignment task.

Method

Participants 654 Amazon Mechanical Turk (MTurk) workers participated in the study. Using MTurk’s worker qualifications we limited participation to workers located in the US and with a previous HIT approval rate $\geq 90\%$. We chose MTurk workers because the number of experimental conditions required a large number of subjects.

Materials Sentences took the form “ M and N ” or “ P and Q ” (see Figure 1). 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 {*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*} (animal words) and familiar P ’s were drawn from {*car*, *bus*, *truck*} (vehicle words). 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. 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 number of familiar M and P words. 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

Figure 1: The $MNPQ$ language. Underlined sentences were withheld from exposure in Experiments 1-3.

<u>$m_1 n_1$</u>	$m_1 n_2$	$m_1 n_3$	<u>$p_1 q_1$</u>	$p_1 q_2$	$p_1 q_3$
$m_2 n_1$	<u>$m_2 n_2$</u>	$m_2 n_3$	$p_2 q_1$	<u>$p_2 q_2$</u>	$p_2 q_3$
$m_3 n_1$	$m_3 n_2$	<u>$m_3 n_3$</u>	$p_3 q_1$	$p_3 q_2$	<u>$p_3 q_3$</u>

to turn on their speakers and click a button, which played a spoken English word. Subjects were required to correctly type the word to continue. The experiment had four phases – exposure, similarity, memory, and referent assign. Below, we detail these phases (for purposes of 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 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 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 , and *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, instead of a sentence from the $MNPQ$ language, we played 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 respecting the co-occurrence relationships as more familiar than the cross-category sentences violating the co-occurrence relationships.

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** and **thite**?”); to facilitate mapping between visual and spoken word forms, the speaker button next to each word played the spoken word when clicked. In two catch trials, 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 rated as more similar than cross-category pairs.

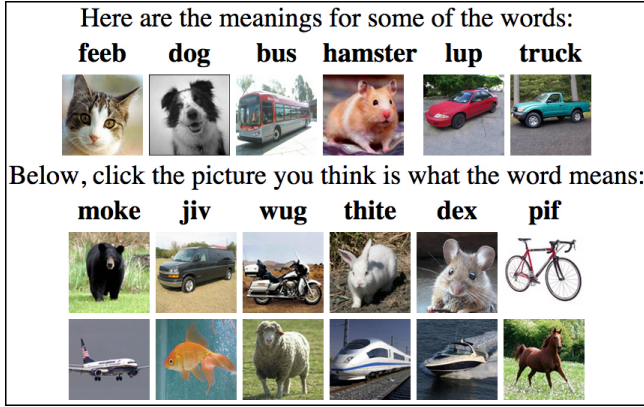


Figure 2: The referent assignment task.

Referent assignment. Subjects made 2AFC referent assignments for the N and Q words (see Figure 2). At the top of the screen, we displayed the M and P words in random order. Under each word, we showed an image of an associated referent. The referents corresponded to the familiar pools for M and P words: CAT, DOG, HAMSTER, CAR, BUS, and TRUCK. Familiar words were always associated with the obvious referents (e.g., “dog” was always paired with an image of a dog). Below the “seeded” word meanings, we displayed a row containing the N and Q words. Under each word, we displayed a 2AFC referent choice between an animal (the “correct” choice for N words) and vehicle words (the “correct” choice for Q words); 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).

Results and Discussion

We excluded the 47 subjects who did not correctly answer all of the catch trials. Results are shown in Figure 3. Next, 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. We analyzed two aspects of the data. First, we were in-

terested in main effects of coherence on score. Second, as one hallmark of statistical learning is sensitivity to the amount of evidence observed, we were interested in the relationship between amount of exposure and score. Accordingly, we looked for exposure \times coherence interactions. A significant interaction 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 an interactive regression model: score \sim exposures \times condition. To examine the differences between the different coherence levels, we used Helmert contrasts analyzing (i) the difference between the 1/3 and 0/3 conditions, (ii) the difference between the 2/3 condition and the 0/3 and 1/3 conditions combined, and (iii) the difference between the 3/3 condition and the 0/3, 1/3, and 2/3 conditions combined. Results of these analyses are shown in Table 1.

Table 1: Regression models

Regressor	β	t	p
Memory			
Exposures	< 0.001	2.67	< 0.01*
Condition: 1/3 – (0/3)	-0.003	-0.03	0.97
Condition: 2/3 – (0/3,1/3)	0.101	1.78	0.074
Condition: 3/3 – (0/3,1/3,2/3)	0.154	3.62	< 0.01*
E \times C: 1/3 – (0/3)	< 0.001	0.17	0.86
E \times C: 2/3 – (0/3,1/3)	> -0.001	-0.20	0.83
E \times C: 3/3 – (0/3,1/3,2/3)	< 0.001	2.76	< 0.01*
Similarity			
Exposures	< 0.001	1.82	0.06
Condition: 1/3 – (0/3)	-0.039	-0.36	0.71
Condition: 2/3 – (0/3,1/3)	0.075	1.33	0.18
Condition: 3/3 – (0/3,1/3,2/3)	0.097	2.30	0.02*
E \times C: 1/3 – (0/3)	< 0.001	0.53	0.59
E \times C: 2/3 – (0/3,1/3)	< 0.001	0.42	0.67
E \times C: 3/3 – (0/3,1/3,2/3)	< 0.001	2.66	< 0.02*
Referent assignment			
Exposures	0.001	3.16	< 0.01*
Condition: 1/3 – (0/3)	0.153	1.01	0.31
Condition: 2/3 – (0/3,1/3)	0.183	2.26	0.02*
Condition: 3/3 – (0/3,1/3,2/3)	0.121	2.00	0.04*
E \times C: 1/3 – (0/3)	> -0.001	-0.29	0.77
E \times C: 2/3 – (0/3,1/3)	> -0.001	-0.89	0.37
E \times C: 3/3 – (0/3,1/3,2/3)	< 0.001	1.29	0.19

Memory There were significant main effects of exposure and condition, with scores in the 3/3 condition being 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, suggesting greater efficiency of statistical learning in 3/3. Thus, more 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, with scores in 3/3 being 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 and 3/3 scores were significantly higher than the rest of the scores combined. 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. We also computed this model using coherence as a continuous variable; this continuous regressor significantly predicted increases in score, $\beta = 0.29$, $t(650) = 3.06$, $p < 0.005$, indicating that *parametrically* increasing coherence resulted in *parametric* increases in referent assignment score.

To summarize, in Experiment 1, we found that higher coherence (1) increased ability to distinguish novel grammatical sentences from sentences violating co-occurrence regularities, (2) sharpened sensitivity to lexical category boundaries related to the co-occurrence regularities, and (3) increased inductive bias in associating words with referents. How does coherence bring about these effects? Frank & Gibson (2009) have shown that *MNPQ* learning can be bolstered by easing working memory demands. Furthermore, 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 conceivable that the presence of semantically coherent known words reduced memory demands and thus improved *MNPQ* learning. 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 0.28 to 0.12, Sobel $z = 7.74$, $p < 0.05$; this reduced value was significantly greater than zero, $t(657) = 3.60$, $p < 0.0005$, indicating partial mediation. After controlling for memory, the regression coefficient relating coherence and referent assignment score decreased significantly from 0.31 to 0.19, Sobel $z = 5.19$, $p < 0.05$; this reduced value was significantly greater than zero, $t(651) = 3.67$, $p < 0.0005$, 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.

Given this result, it is natural to ask whether known words *per se* can sufficiently ease memory demands so as to facilitate *MNPQ* learning, or whether they must have semantic coherence; we consider this question in Experiment 3. A different, though not mutually exclusive, possibility is that any relatively salient type of coherence, such as phonological co-

herence, is sufficient to facilitate distributional learning. We consider this possibility in Experiment 2.

Experiment 2: Phonological Coherence

As noted previously, Lany and Saffran (2010) found evidence of successful *MNPQ* learning when co-occurrence regularities were perfectly correlated with a phonological property – the number of syllables in an *N/Q* word. We sought to compare phonological coherence with semantic coherence. Thus, in Experiment 2, we measured the result of a phonological manipulation in which all *M* words began with “r” and all *P* words began with “z”.

Method

Participants 157 MTurk workers participated in the study.

Design The method was similar to that of Experiment 1. *M*’s were {*rull*, *rudge*, *ruck*} and *P*’s were {*zof*, *zerm*, *zabe*}.

Results and Discussion

Results are shown in Figure 4. We compared the phonological condition results with the 0/3 and 3/3 conditions of Experiment 1 using a regression model with Helmert contrasts analyzing (i) the difference between the 0/3 and phonological conditions and (ii) the difference between the 3/3 condition and the 0/3 and phonological conditions combined.

Memory Phonological scores were not significantly different from 0/3 scores, $t(466) = 0.96$, $p > 0.05$ and both combined were significantly lower than 3/3 scores, $\beta = 0.2$, $t(466) = 3.44$, $p < 0.001$. Phonological efficiency was not significantly different from 0/3 efficiency, $t(466) = -0.17$, $p > 0.05$ and both combined were significantly lower than 3/3 efficiency, $\beta = 0.0007$, $t(466) = 2.76$, $p < 0.01$.

Similarity Phonological scores were not significantly different from 0/3 scores, $t(466) = -0.16$, $p > 0.05$, and both combined were significantly lower than 3/3 scores, $\beta = 0.14$, $t(466) = 2.54$, $p < 0.05$. Phonological efficiency was not significantly different from 0/3 efficiency, $t(466) = -0.29$, $p > 0.05$, and both combined were significantly lower than 3/3 efficiency, $\beta = 0.0008$, $t(466) = 3.2$, $p < 0.005$.

Referent assignment Phonological scores were not significantly different from 0/3 scores, $t(466) = 1.47$, $p > 0.05$, and both combined were significantly less than 3/3 scores, $\beta = 0.2$, $t(466) = 2.35$, $p < 0.05$. There were no differences in efficiency (recall that this was also the case in Experiment 1). In terms of facilitating acquisition of the *MN* and *PQ* co-occurrence regularities, the phonological manipulation was indistinguishable from the 0/3 condition, and hence was markedly less effective than semantic coherence at the 3/3 level. It must be noted that the phonological regularity we introduced was the onset consonant (*r*- words versus *z*- words) applied to *M/P* words, whereas Lany and Saffran (2010) used syllable length (monosyllabic versus disyllabic words) applied to *N/Q* words, making direct comparison difficult. Presently, we have established that a particular kind

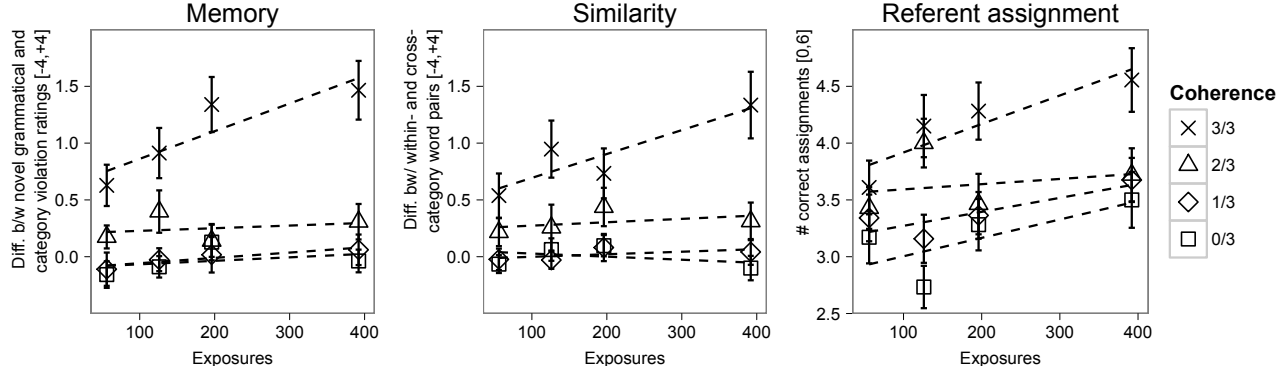


Figure 3: Experiment 1 results. Each plot shows data for one measure (memory, similarity, meaning) in Experiment 1. Dots show condition means, while dashed lines show the best-fitting linear trend.

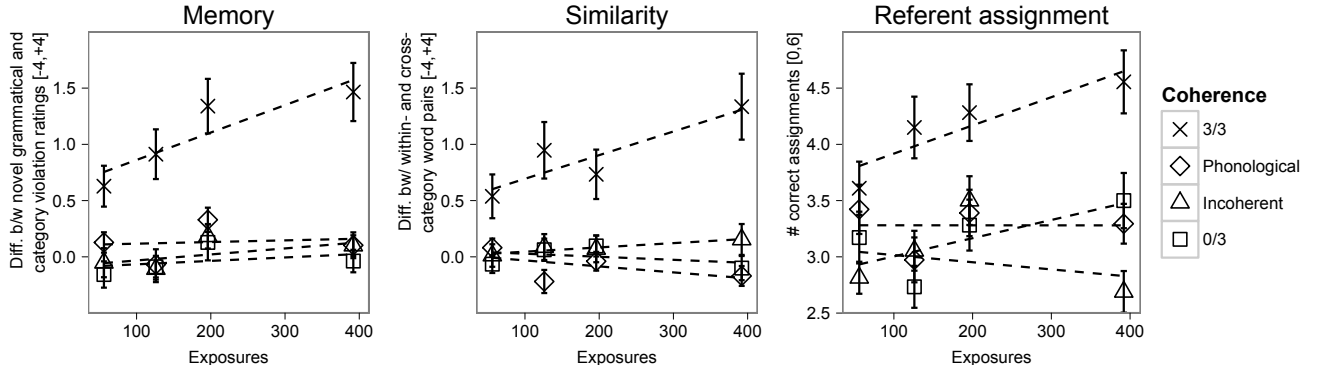


Figure 4: Experiment 2 and 3 results. Each plot shows data for one measure (memory, similarity, meaning) in Experiments 2 and 3. Dots show condition means, while dashed lines show the best-fitting linear trend.

of phonological regularity (onset consonant of a co-occurring word) is a far weaker correlated cue than semantic coherence.

Experiment 3: Semantic Incoherence

The *M* and *P* words in Experiment 1 were all known and had semantic coherence. In Experiment 3, we explored whether coherence is necessary 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.

Method

Participants 151 MTurk workers participated in the study.

Design and Procedure The method was similar to that of Experiment 2. The *M* and *P* words were known but did not adhere to any clear semantic organization. The specific *M* and *P* words were drawn 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 card.

Results and Discussion

Memory Incoherent scores were not significantly different from 0/3 scores, $t(460) = 0.056$, $p > 0.05$ and both combined were significantly lower than 3/3 scores, $\beta = 0.23$, $t(460) = 3.95$, $p < 0.0001$. Incoherent efficiency was not significantly different from 0/3 efficiency, $t(460) = 0.25$, $p > 0.05$ and both combined were significantly lower than 3/3 efficiency, $\beta = 0.0006$, $t(460) = 2.55$, $p < 0.05$.

Similarity Incoherent scores were not significantly different from 0/3 scores, $t(460) = -0.22$, $p > 0.05$, and both combined were significantly lower than 3/3 scores, $\beta = 0.15$, $t(460) = 2.47$, $p < 0.05$. Incoherent efficiency was not significantly different from 0/3 efficiency, $t(460) = 0.723$, $p > 0.05$, and both combined were significantly lower than 3/3 efficiency, $\beta = 0.0006$, $t(460) = 2.54$, $p < 0.05$.

Referent assignment Incoherent scores were not significantly different from 0/3 scores, $t(460) = 0.82$, $p > 0.05$, and both combined were significantly less than 3/3 scores, $\beta = 0.23$, $t(460) = 2.83$, $p < 0.01$. There were no differences in efficiency (recall that this was also the case in Experiments 1 and 2). 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 distributional learning.

General discussion

We have conducted the first large-scale systematic investigation of the effects of exposure and various correlated cues (semantic and phonological coherence) on distributional learning of word meanings. We have shown that semantic coherence aids distributional learning in the *MNPQ* regime far more than phonological (onset consonant) coherence. Additionally, our experiments indicate that *coherence* is necessary – semantically incoherent linguistic input provided virtually no benefit. Additionally, we showed that coherence works in part by alleviating memory limitations, though our data suggests that there may be aspects of distributional word learning not bottlenecked by memory resources.

We conjecture that word learners may be using semantic coherence to infer the topic of discourse and that this gist topic meaning influences the representations of co-occurring novel words (cf. the topic-learning models of Griffiths, Steyvers, & Tenenbaum, 2007). It may be through this process that people know *tort* to be a legal term and *transducer* to be an engineering term, despite not knowing the precise meanings of these words. Under this account, learning would be easier for words that occur in contexts high in semantic information and coherence. Thus, we would expect learning to have a “contiguous” character (faster learning for words occurring in familiar contexts than for words occurring in less familiar contexts), a possibility we plan to test in future work.

Our experiments highlight a limitation of artificial language learning paradigms. 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.

Our work on distributional learning of semantic properties is in agreement with the extant literature on distributional learning of syntactic properties (viz. grammatical gender). Results from these studies (e.g., Brooks et al., 1993; Frigo & McDonald, 1998) indicate that children and adults fail to learn *MNPQ* categories without correlated cues, but they can learn given (e.g.,) correlated phonological markers. We believe that correlated cues – be they semantic, phonological, or otherwise – serve a common purpose: to reduce the space of possible categories.

Acknowledgments

We thank Paul Thibodeau and Jay McClelland for helpful discussions.

References

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

- Braine, M. D. S. (1987). What is learned in acquiring word classes: A step toward an acquisition theory. In *Mechanisms of language acquisition* (pp. 65–87). Hillsdale, NJ: Lawrence Erlbaum.
- Brooks, P. J., Braine, M. D. S., Catalano, L., Brody, R. E., & Sudhalter, V. (1993, February). 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.
- Firth, J. R. (1957). A Synopsis of Linguistic Theory 1930–1955. In *Studies in linguistic analysis* (pp. 1–32). Oxford: Blackwell.
- Frank, M., & Gibson, E. (2011). Overcoming memory limitations in rule learning. *Language Learning and Development*, 7(2), 130–148.
- Frigo, L., & McDonald, J. L. (1998, August). Properties of Phonological Markers That Affect the Acquisition of Gender-Like Subclasses. *Journal of Memory and Language*, 39(2), 218–245.
- Gerken, L., Wilson, R., & Lewis, W. (2005). Infants can use distributional cues to form syntactic categories. *Journal of Child Language*.
- Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological review*, 114(2), 211–244.
- Harris, Z. S. (1951). *Methods in structural linguistics*. Chicago: University of Chicago Press.
- 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), 211–240.
- Lany, J., & Saffran, J. (2010). From Statistics to Meaning: Infants’ Acquisition of Lexical Categories. *Psychological Science*, 21(2), 284.
- Mintz, T. (2002). Category induction from distributional cues in an artificial language. *Memory and Cognition*, 30(5), 678.
- Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, 22(4), 425–469.
- Reeder, P., Newport, E., & Aslin, R. (2010). Novel Words in Novel Contexts: The Role of Distributional Information in Form-class Category Learning. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*.
- Riordan, B., & Jones, M. N. (2010). Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*, 3(2), 303–345.
- 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.
- Treiman, R., & Danis, C. (1988). Short-term memory errors for spoken syllables are affected by the linguistic structure of the syllables. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(1), 145–152.